the Availability Digest

www.availabilitydigest.com

Lightning Downs Amazon – Not!

September 2011

A power failure in the evening of Sunday, August 7, 2011, took down an Availability Zone in Amazon's Dublin data center, which houses Amazon's European region for its Elastic Compute Cloud. Thousands of users in dozens of European countries found that they had no access to their applications nor to their data. It was days before service was restored. The power utility reported that the power loss was caused by a lightning strike that caused a massive transformer in an electrical substation outside of Dublin, Ireland, to explode.

Why should a power failure cause days of havoc? Where were the backup generators? As it turns out, there were several factors that led to a failure chain totally unanticipated by Amazon. Factors included hardware faults, software bugs, and human errors. Lightning was not one of them. As is Amazon's practice, it was very forthcoming with updates on the status of the outage via its Service Health Dashboard. However, the complexity of the failure chain was evident in some of the confusion exhibited by Amazon as it tried to give a running commentary on the situation.

Amazon's Availability Zones

Before delving further into this disaster, let us briefly review Amazon's Availability Zones, since they played a role in some of the problems as well as the resolution of the problems.[1]

Regions

Amazon's Elastic Compute Cloud (EC2) is arguably the leading service today for deploying custom applications in a cloud environment. Amazon has gone to great lengths to ensure the availability of its cloud services. It has broken its cloud infrastructure into five geographically dispersed regional data centers - US East (Virginia), US West (Northern California), EU (Ireland), Asia Pacific (Singapore), and Asia Pacific (Japan).

Availability Zones

Within each region, Amazon provides independent Availability Zones. An Availability Zone (AZ) is a data center that is independent of all other data centers in the region, though the AZ data centers within a region are collocated in the regional data center. The U.S.-East Region data center, for instance, has four Availability Zones. Should an AZ fail, the others continue uninterrupted. A customer can run a critical application in multiple Availability Zones within a region to ensure availability. If desired, an application can also have a redundant instance in other regions.

[1] Amazon's Cloud Downed by Fat Finger, *Availability Digest*; May 2011.
http://www.availabilitydigest.com/public_articles/0605/amazon_ebs.pdf

Elastic Block Store (EBS)

Amazon offers two storage services – S3 (Simple Storage Service) and EBS (Elastic Block Store). EBS offers persistent block storage that can be attached to an EC2 instance. The storage then can be used like any block storage device, such as a SAN. The contents of each EBS volume are replicated to multiple backup EBS volumes. An EBS volume can be accessed by only one EC2 instance, and they both must be in the same Availability Zone.

To provide high availability, a customer can run several copies of an EC2 instance in different Availability Zones within a region. Amazon will replicate data between the EBS volumes in different AZs. Remote EC2 instances can also be run in other regions to provide disaster recovery. However, data replication between EBS volumes is the responsibility of the customer in these cases.

Simple Storage Service (S3)

S3, on the other hand, is independent of EC2. It provides Storage as a Service features through web interfaces such as SOAP and REST. It can be used to store and retrieve data from anywhere on the web. It is hugely scalable and stores data redundantly for high availability.

Remirroring

A factor that contributed to Amazon's problems was a process called remirroring. When an EBS volume loses connectivity to one of its replica volumes, it searches for a new volume to which it can replicate its data. This is called *remirroring*. To do this, the primary volume searches its EBS cluster for another volume that has enough storage capacity to act as a replica. The transfer of volume data is then initiated to the new replica.

To ensure consistency, access to volumes that are being remirrored are blocked until the remirroring has been completed and a primary replica is identified. While this is happening, EC2 instances attempting to access this data are blocked – in EC2 terms, they are *stuck*.

The EU Region Outage

At 10:41 AM PDT (6:41 PM Dublin time) on Sunday, August 7th, an Availability Zone in Amazon's European data center in Dublin lost power. The local power utility blamed the power loss on a lightning strike that caused an explosion of a 100 kilovolt, 10 megawatt transformer in a power substation outside of Dublin. Even worse, the surge from the lightning strike appeared to have damaged the controllers for the backup generators for the AZ, which consequently failed to come up. The Availability Zone went dark.

An early post by Amazon said:

> "The transient electric deviation caused by the explosion was large enough that it propagated to a portion of the phase control system that synchronizes the backup generator plant, disabling some of them. Power sources must be phase-synchronized before they can be brought online to load. Bringing these generators online required manual synchronization."

Availability zones in a region are generally powered by independent sources, so other AZs in the European region were unaffected.

By 1:49 PM PDT, some power was restored and enough network capacity was available to allow access to the AZ. Amazon then focused on bringing the failed EC2 instances and their EBS volumes back on line, but progress was slower than anticipated.
.
Then to Amazon's horror, they discovered that an error in Amazon Web Services cleanup software resulted in some customers having data deleted from their backup data snapshots, preventing data

recovery in those instances in which data was lost. In order to restore the snapshots, Amazon had to copy all the data from each inconsistent EBS node to S3, where they could rebuild the volume and convert it to snapshot format. This was a massive task – each EBS volume can hold up to a terabyte of data. Amazon had to truck in additional servers to aid in the recovery process, an effort that was compounded because it was nighttime in Dublin. Early Tuesday morning, Amazon announced that half of the volumes that had been in an inconsistent state had been recovered. Affected EC2 customers had now been down for almost two days.

Finally, in the evening of Wednesday, August 10th, 98% of the European EC2 services were restored. Some customers had been down for over three days.

What Really Happened?

It seems that ESB Networks, the Irish electricity provider, had a later version of events. It said that the problem arose following a fault in one of its CityWest substations, noting that an alternate power source should have kicked-in automatically in less than a second.

The utility added that there was no report of an explosion or fire. And to compound matters, it said there was no record of a lightning strike in the Dublin area at the time. It said that its original assessment that a lightning strike was to blame was wrong:

> "ESB Networks can confirm that at 18:16 on Sunday August 7th, a number of customers in CityWest lost electricity supply. … In this case, the problem was the failure of a 110kV transformer in the CityWest 110kV substation. The cause of this failure is still being investigated at this time but our initial assessment of lightning as the cause has now been ruled out. This initial supply disruption lasted for approximately 1 hour as ESB Networks worked to restore supply. There was an ongoing partial outage in the area until 11pm. The interruption affected about 100 customers in the Citywest area, including Amazon and a number of other data centres. Another Amazon data centre served by ESB in South County Dublin was not directly affected by the outage, though it did experience a voltage dip which lasted for less than one second."

As is usually the case, once all problems had been resolved, Amazon published a detailed account of the outage and the efforts to recover from it.[2] It is quite an extensive account and is summarized as follows.

> 10:41 PM PDT: The AZ lost power. "Normally, when utility power fails, electrical load is seamlessly picked up by backup generators. Programmable Logic Controllers (PLCs) assure that the electrical phase is synchronized between generators before their power is brought online. In this case, one of the PLCs did not complete the connection of a portion of the generators to bring them online … With no utility power, and backup generators for a large portion of this Availability Zone disabled, there was insufficient power for all of the servers in the Availability Zone to continue operating. Uninterruptable Power Supplies (UPSs) that provide a short period of battery power quickly drained[3] and we lost power to almost all of the EC2 instances and 58% of the EBS volumes in that Availability Zone. We also lost power to the EC2 networking gear that connects this Availability Zone to the Internet and connects this Availability Zone to the other Availability Zones in the Region."
>
> 11:05 AM PDT: "We were seeing launch delays and API errors in all EU West Availability Zones. There were two primary factors that contributed to this. First, our EC2 management service … has servers in each Availability Zone. The management servers which receive requests continued to route requests to management servers in the affected Availability Zone. Because the management servers in the affected Availability Zone were inaccessible, requests routed to those servers failed. Second, the EC2 management servers receiving requests were continuing to accept RunInstances

[2] Summary of the Amazon EC2, Amazon EBS, and Amazon RDS Service Event in the EU West Region, *Amazon message.* http://aws.amazon.com/message/65648/

[3] The 25-minute difference between the power companies report of power loss and Amazon's report of power loss was perhaps the time it took for the backup batteries to drain.

requests targeted at the impacted Availability Zone. Rather than failing these requests immediately, they were queued and our management servers attempted to process them. Fairly quickly, a large number of these requests began to queue up and we overloaded the management servers receiving requests …"

11:54 AM PDT: "We had been able to bring some of the backup generators online by manually phase-synchronizing the power sources. This restored power to many of the EC2 instances and EBS volumes, but a majority of the networking gear was still without power, so these restored instances were still inaccessible.

1:49 PM PDT: "Power had been restored to enough of our network devices that we were able to re-establish connectivity to the Availability Zone. Many of the instances and volumes in the Availability Zone became accessible at this time."

EBS Node Recovery: "EBS volumes in the affected Availability Zone entered one of three states: (1) online – none of the nodes holding a volume's data lost power, (2) re-mirroring – a subset of the nodes storing the volume were offline due to power loss and the remaining nodes were re-replicating their data, and (3) offline – all nodes lost power.

"In the first case, EBS volumes continued to function normally.

"In the second case, the majority of nodes were able to leverage the significant amount of spare capacity in the affected Availability Zone to successfully re-mirror, and enable the volume to recover. But, because we had such an unusually large number of EBS volumes lose power, the spare capacity we had on hand to support re-mirroring wasn't enough. We ran out of spare capacity before all of the volumes were able to successfully re-mirror. As a result, a number of customers' volumes became "stuck" as they attempted to write to their volume, but their volume had not yet found a new node to receive a replica. In order to get the "stuck" volumes back online, we had to add more capacity. We brought in additional labor to get more onsite capacity online and trucked in servers from another Availability Zone in the Region. There were delays as this was nighttime in Dublin and the logistics of trucking required mobilizing transportation some distance from the datacenter. Once the additional capacity was available, we were able to recover the remaining volumes waiting for space to complete a successful re-mirror.

"In the third case, when an EC2 instance and all nodes containing EBS volume replicas concurrently lose power, we cannot verify that all of the writes to all of the nodes are completely consistent. … Bringing a volume back in an inconsistent state without the customer being aware could cause undetectable, latent data corruption issues which could trigger a serious impact later. For the volumes we assumed were inconsistent, we produced a recovery snapshot to enable customers to create a new volume and check its consistency before trying to use it. The process of producing recovery snapshots was time-consuming because we had to first copy all of the data from each node to Amazon Simple Storage Service (Amazon S3), process that data to turn it into the snapshot storage format, and re-copy the data to make it accessible from a customer's account. Many of the volumes contained a lot of data (EBS volumes can hold as much as 1 TB per volume). … By 8:25 PM PDT on August 10th, we were 98% complete, with the remaining few snapshots requiring manual attention."

Multi-AZ Failover Faults: "… a portion of Multi-AZ database instances experienced prolonged failover times. … Multi-AZ database instances consist of a "primary" database instance and a synchronously replicated "secondary" database instance in another Availability Zone. When the system detects that a primary database instance might be failing, upon verification via a health check that the primary is no longer accepting traffic, the secondary is promoted to primary. This verification is important to avoid a "split brain" situation, one where both the primary and the secondary database instances are accepting writes and some writes exist on one database while some exist on another. During the event, there were failures of Multi-AZ primary database instances in the affected Availability Zone.

“For a portion of these Multi-AZ primary-secondary pairs, a DNS connectivity issue related to the power loss prevented the health check from finding the IP address it needed to contact and kept the secondary from immediately assuming the role of the primary. … The DNS connectivity issues triggered a software bug that caused failover times to the secondary database instance to extend significantly for a small subset of Multi-AZ deployments.”

Snapshot Failures: “Separately, and independent from issues emanating from the power disruption, we discovered an error in the EBS software that cleans up unused storage for snapshots after customers have deleted an EBS snapshot. … At least one week passes from the time the snapshot cleanup identification process runs before any blocks it has flagged for deletion are allowed to be removed. Each day, it updates the lists of blocks to delete, … and if any block eligible for deletion the day before now shows up in the most recent list of blocks referenced by active EBS volumes or snapshots, the process flags those blocks for [manual] analysis. Actual deletion is executed by an engineer who first, before running the actual deletion process, evaluates the blocks flagged for analysis and verifies that there are no blocks in the list scheduled to be deleted that have been flagged for analysis. The engineer must present his verification step to another engineer who approves the deletion.

“In one of the days leading up to the Friday, August 5th deletion run, there was a hardware failure that the snapshot cleanup identification software did not correctly detect and handle. The result was that the list of snapshot references used as input to the cleanup process was incomplete. Because the list of snapshot references was incomplete, the snapshot cleanup identification process incorrectly believed a number of blocks were no longer referenced and had flagged those blocks for deletion … On August 5th, the engineer running the snapshot deletion process checked the blocks flagged for analysis before running the actual deletion process in the EU West Region. The human checks in this process failed to detect the error and the deletion process was executed. On Friday evening, an error accessing one of the affected snapshots triggered us to investigate.

“By Sunday morning, August 7th, we had completed the work to fully understand root cause, prevent the problem from recurring, and build a tool that could create recovery snapshots for affected snapshots. We then started to do the work necessary to map these affected snapshots to customers and build the recovery snapshots, with the aim to communicate this information to customers by Sunday night. However, before we got very far in this endeavor, the power event began. We had to temporarily stop work on the snapshot issue to respond to the power event. Once we’d been able to restore the majority of the EBS volumes affected by the power event, we returned to working on the snapshot issue in parallel with restoring the remainder of the EBS volumes that were recovering from the power event. …”

Lessons Learned

As a result of this chain of faults, Amazon is taking several corrective actions:

- It will add redundancy and more isolation for its backup generator control PLCs so they are insulated from other failures.

- It will address the resource saturation that affected API calls at the beginning of the disruption. It will implement better load balancing to quickly take failed API management service hosts out of production.

- It will continue to create additional capabilities that make it easier to develop and deploy applications in multiple Availability Zones.

- It drastically reduced the long recovery time required to recover stuck or inconsistent EBS volumes. It will create the capability to recover volumes directly on the EBS servers upon restoration of power without having to move the data off of those servers to S3.

- It has corrected the software bug that inappropriately deleted snapshot data.
- It will improve the handling of health check failures.
- Though its communication of the status of the situation was greatly improved over past instances, it will strive to improve crisis communications even further.

Amazon provided a ten-day usage credit for all customers who had an EBS volume in the compromised AZ, whether or not they were affected by the outage. They provided a thirty-day credit for all customers whose snapshot blocks were inadvertently deleted.

Acknowledgements

In addition to the references previously noted, the following resources were used to provide material for this article:

Lightning Strike in Dublin Downs Amazon, Microsoft Clouds, *CIO.com*; August 8, 2011.
Amazon Cloud Outage: What Can Be Learned?, *Information Week*; August 9, 2011.
Amazon Cloud Outage Cleanup Hits Software Error, *Information Week*; August 10, 2011.
Amazon Outage And Cloud Common Sense, *Information Week*; August 10, 2011.
Questions raised around Amazon's "lightning claims" at Dublin data center, *TNW Insider*; August 10, 2011.
Amazon Admits Multiple Problems at Dublin Datacenter, *Wall Street Journal*; August 15, 2011.