Summary of the AWS Service Event in the US East Region

July 2, 2012

We'd like to share more about the service disruption which occurred last Friday night, June 29th, in one of our Availability Zones in the US East-1 Region. The event was triggered during a large scale electrical storm which swept through the Northern Virginia area. We regret the problems experienced by customers affected by the disruption and, in addition to giving more detail, also wanted to provide information on actions we'll be taking to mitigate these issues in the future.

Our US East-1 Region consists of more than 10 datacenters structured into multiple Availability Zones. These Availability Zones are in distinct physical locations and are engineered to isolate failure from each other. Last Friday, due to weather warnings of the approaching storm, all change activity in the US East-1 Region had been cancelled and extra personnel had been called into the datacenters for the evening.

On Friday night, as the storm progressed, several US East-1 datacenters in Availability Zones which would remain unaffected by events that evening saw utility power fluctuations. Backup systems in those datacenters responded as designed, resulting in no loss of power or customer impact. At 7:24pm PDT, a large voltage spike was experienced by the electrical switching equipment in two of the US East-1 datacenters supporting a single Availability Zone. All utility electrical switches in both datacenters initiated transfer to generator power. In one of the datacenters, the transfer completed without incident. In the other, the generators started successfully, but each generator independently failed to provide stable voltage as they were brought into service. As a result, the generators did not pick up the load and servers operated without interruption during this period on the Uninterruptable Power Supply ("UPS") units. Shortly thereafter, utility power was restored and our datacenter personnel transferred the datacenter back to utility power. The utility power in the Region failed a second time at 7:57pm PDT. Again, all rooms of this one facility failed to successfully transfer to generator power while all of our other datacenters in the Region continued to operate without customer impact.

In the single datacenter that did not successfully transfer to the generator backup, all servers continued to operate normally on Uninterruptable Power Supply ("UPS") power. As onsite personnel worked to stabilize the primary and backup power generators, the UPS systems were depleting and servers began losing power at 8:04pm PDT. Ten minutes later, the backup generator power was stabilized, the UPSs were restarted, and power started to be restored by 8:14pm PDT. At 8:24pm PDT, the full facility had power to all racks.

The generators and electrical switching equipment in the datacenter that experienced the failure were all the same brand and all installed in late 2010 and early 2011. Prior to installation in this facility, the generators were rigorously tested by the manufacturer. At datacenter commissioning time, they again passed all load tests (approximately 8 hours of testing) without issue. On May 12th of this year, we conducted a full load test where the entire datacenter switched to and ran successfully on these same generators, and all systems operated correctly. The generators and electrical equipment in this datacenter are less than two years old, maintained by manufacturer representatives to manufacturer standards, and tested weekly. In addition, these generators operated flawlessly, once brought online Friday night, for just over 30 hours until utility power was restored to this datacenter. The equipment will be repaired, recertified by the manufacturer, and retested at full load onsite or it will be replaced entirely. In the interim, because the generators ran successfully for 30 hours after being manually brought online, we are confident they will perform properly if the load is transferred to them. Therefore, prior to completing the engineering work mentioned above, we will lengthen the amount of time the electrical switching equipment gives the generators to reach stable power before the switch board assesses whether the generators are ready to accept the full power load. Additionally, we will expand the power quality tolerances allowed when evaluating whether to switch the load to generator power. We will expand the size of the onsite 24x7 engineering staff to ensure that if there is a repeat event, the switch to generator will be completed manually (if necessary) before UPSs discharge and there is any customer impact.

Though the resources in this datacenter, including Elastic Compute Cloud (EC2) instances, Elastic Block Store (EBS) storage volumes, Relational Database Service (RDS) instances, and Elastic Load Balancer (ELB) instances, represent a single-digit percentage of the total resources in the US East-1 Region, there was significant impact to many customers. The impact manifested in two forms. The first was the unavailability of instances and volumes running in the affected datacenter. This kind of impact was limited to the affected Availability Zone. Other Availability Zones in the US East-1 Region continued functioning normally. The second form of impact was degradation of service "control planes" which allow customers to take action and create, remove, or change resources across the Region. While control planes aren't required for the ongoing use of resources, they are particularly useful in outages where customers are trying to react to the loss of resources in one Availability Zone by moving to another.

EC2 and EBS

Approximately 7% of the EC2 instances in the US-EAST-1 Region were in the impacted Availability Zone and impacted by the power loss. These instances were offline until power was restored and systems restarted. EC2 instances operating in other Availability Zones within the US East-1 Region continued to function as they did prior to the event. Internet connectivity into the Region was unaffected. The vast majority of these instances came back online between 11:15pm PDT and just after midnight. Time for the completion of this recovery was

extended by a bottleneck in our server booting process. Removing this bottleneck is one of the actions we'll take to improve recovery times in the face of power failure. EBS had a comparable percentage (relative to EC2) of its volumes in the Region impacted by this event. The majority of EBS servers had been brought up by 12:25am PDT on Saturday. However, for EBS data volumes that had in-flight writes at the time of the power loss, those volumes had the potential to be in an inconsistent state. Rather than return those volumes in a potentially inconsistent state, once the EBS servers are back up and available, EBS brings customer volumes back online in an impaired state where all I/O on the volume is paused. Customers can then verify the volume is consistent and resume using it. Though the time to recover these EBS volumes has been reduced dramatically over the last 6 months, the number of volumes requiring processing was large enough that it still took several hours to complete the backlog. By 2:45am PDT, 90% of outstanding volumes had been turned over to customers. We have identified several areas in the recovery process that we will further optimize to improve the speed of processing recovered volumes.

The control planes for EC2 and EBS were significantly impacted by the power failure, and calls to create new resources or change existing resources failed. From 8:04pm PDT to 9:10pm PDT, customers were not able to launch new EC2 instances, create EBS volumes, or attach volumes in any Availability Zone in the US-East-1 Region. At 9:10pm PDT, control plane functionality was restored for the Region. Customers trying to attach or detach impacted EBS volumes would have continued to experienced errors until their impacted EBS volumes were recovered. The duration of the recovery time for the EC2 and EBS control planes was the result of our inability to rapidly fail over to a new primary datastore. The EC2 and EBS APIs are implemented on multi-Availability Zone replicated datastores. These datastores are used to store metadata for resources such as instances, volumes, and snapshots. To protect against datastore corruption, currently when the primary copy loses power, the system automatically flips to a read-only mode in the other Availability Zones until power is restored to the affected Availability Zone or until we determine it is safe to promote another copy to primary. We are addressing the sources of blockage which forced manual assessment and required hand-managed failover for the control plane, and have work already underway to have this flip happen automatically.

Elastic Load Balancing

Elastic Load Balancers (ELBs) allow web traffic directed at a single IP address to be spread across many EC2 instances. They are a tool for high availability as traffic to a single end-point can be handled by many redundant servers. ELBs live in individual Availability Zones and front EC2 instances in those same zones or in other Availability Zones.

For single-Availability Zone ELBs, the ELB service maintains one ELB in the specified Availability Zone. If that ELB fails, the ELB control plane assigns its configuration and IP address to another ELB server in that Availability Zone. This normally requires a very short period of time. If there is a large scale issue in the Availability Zone, there may be insufficient capacity to immediately provide a new ELB and replacement will wait for capacity to be made available.

ELBs can also be deployed in multiple Availability Zones. In this configuration, each Availability Zone's end-point will have a separate IP address. A single Domain Name will point to all of the end-points' IP addresses. When a client, such as a web browser, queries DNS with a Domain Name, it receives the IP address ("A") records of all of the ELBs in random order. While some clients only process a single IP address, many (such as newer versions of web-browsers) will retry the subsequent IP addresses if they fail to connect to the first. A large number of non-browser clients only operate with a single IP address.

For multi-Availability Zone ELBs, the ELB service maintains ELBs redundantly in the Availability Zones a customer requests them to be in so that failure of a single machine or datacenter won't take down the end-point. The ELB service avoids impact (even for clients which can only process a single IP address) by detecting failure and eliminating the problematic ELB instance's IP address from the list returned by DNS. The ELB control plane processes all management events for ELBs including traffic shifts due to failure, size scaling for ELB due to traffic growth, and addition and removal of EC2 instances from association with a given ELB.

During the disruption this past Friday night, the control plane (which encompasses calls to add a new ELB, scale an ELB, add EC2 instances to an ELB, and remove traffic from ELBs) began performing traffic shifts to account for the loss of load balancers in the affected Availability Zone. As the power and systems returned, a large number of ELBs came up in a state which triggered a bug we hadn't seen before. The bug caused the ELB control plane to attempt to scale these ELBs to larger ELB instance sizes. This resulted in a sudden flood of requests which began to backlog the control plane. At the same time, customers began launching new EC2 instances to replace capacity lost in the impacted Availability Zone, requesting the instances be added to existing load balancers in the other zones. These requests further increased the ELB control plane backlog. Because the ELB control plane currently manages requests for the US East-1 Region through a shared queue, it fell increasingly behind in processing these requests; and pretty soon, these requests started taking a very long time to complete.

While direct impact was limited to those ELBs which had failed in the power-affected datacenter and hadn't yet had their traffic shifted, the ELB service's inability to quickly process new requests delayed recovery for many customers who were replacing lost EC2 capacity by launching new instances in other Availability Zones. For multi-Availability Zone ELBs, if a client attempted to connect to an ELB in a healthy Availability Zone, it succeeded. If a client attempted to connect to an ELB in the impacted Availability Zone and didn't retry using one of the alternate IP addresses returned, it would fail to connect until the backlogged traffic shift occurred and it issued a new DNS query. As mentioned, many modern web browsers perform multiple attempts when given multiple IP addresses; but many clients, especially game consoles and other consumer electronics, only use one IP address returned from the DNS query.

As a result of these impacts and our learning from them, we are breaking ELB processing into multiple queues to improve overall throughput and to allow more rapid processing of time-sensitive actions such as traffic shifts. We are also going to immediately develop a backup DNS re-weighting that can very quickly shift all ELB traffic away from an impacted Availability Zone without contacting the control plane.

Relational Database Service (RDS)

RDS provides two modes of operation: Single Availability Zone (Single-AZ), where a single database instance operates in one Availability Zone; and Multi Availability Zone (Multi-AZ), where two database instances are synchronously operated in two different Availability Zones. For Multi-AZ RDS, one of the two database instances is the "primary" and the other is a "standby." The primary handles all database requests and replicates to the standby. In the case where a primary fails, the standby is promoted to be the new primary.

Single-AZ RDS Instances, by default, have backups turned on. When a Single-AZ RDS instance fails, there are two kinds of recovery that are possible. If EBS volumes do not require recovery, the database instance can simply be restarted. If recovery is required, the backups are used to restore the database. In some cases, where backups have been turned off by customers, there can be no recovery and the instance is lost unless manual backups have been taken.

Multi-AZ RDS Instances detect failure in the primary or standby and immediately take action. If the primary fails, the DNS CNAME record is updated to point to the standby. If the standby fails, a new instance is launched and instantiated from the primary as the new standby. Once failure is confirmed, failover can take place in less than a minute.

When servers lost power in the impacted datacenter, many Single-AZ RDS instances in that Availability Zone became unavailable. There was no way to recover these instances until servers were powered up, booted, and brought online. By 10pm PDT, a large number of the affected Single-AZ RDS instances had been brought online. There were many remaining instances which required EBS to recover storage volumes. These followed the timeline described above for EBS impact. Once volumes were recovered, customers could apply backups and restore their Single-AZ RDS instances. In addition to the actions noted above with EBS, RDS will be working to improve the speed at which volumes available for recovery can be processed.

At the point of power loss, most Multi-AZ instances almost instantly promoted their standby in a healthy AZ to "primary" as expected. However, a small number of Multi-AZ RDS instances did not complete failover, due to a software bug. The bug was introduced in April when we made changes to the way we handle storage failure. It is only manifested when a certain sequence of communication failure is experienced, situations we saw during this event as a variety of server shutdown sequences occurred. This triggered a failsafe which required manual intervention to complete the failover. In most cases, the manual work could be completed without EBS recovery taking place. The majority of remaining Multi-AZ failovers were completed by 11:00pm PDT. The remaining Multi-AZ instances were processed when EBS volume recovery completed for their storage volumes.

To address the issues we had with some Multi-AZ RDS Instances failovers, we have a mitigation for the bug in test and will be rolling it out in production in the coming weeks.

Final Thoughts

We apologize for the inconvenience and trouble this caused for affected customers. We know how critical our services are to our customers' businesses. If you've followed the history of AWS, the customer focus we have, and the pace with which we iterate, we think you know that we will do everything we can to learn from this event and use it to drive improvement across our services. We will spend many hours over the coming days and weeks improving our understanding of the details of the various parts of this event and determining how to make further changes to improve our services and processes.

Sincerely,
The AWS Team

AWS Cloud Security
What's New
Blogs
Press Releases

Resources for AWS

Getting Started
Training and Certification
AWS Solutions Portfolio
Architecture Center
Product and Technical FAQs
Analyst Reports
AWS Partner Network

Developers on AWS

Developer Center
SDKs & Tools
.NET on AWS
Python on AWS
Java on AWS
PHP on AWS
Javascript on AWS

Help

Contact Us
AWS Careers
File a Support Ticket
Knowledge Center
AWS Support Overview
Legal

Amazon is an Equal Opportunity Employer: *Minority / Women / Disability / Veteran / Gender Identity / Sexual Orientation / Age.*

Language عربي | Bahasa Indonesia | Deutsch | English | Español | Français | Italiano | Português | Tiếng Việt | Türkçe | Русский | ■■■ | □□□ | □□□ | □□ (□□) | □□ (□□)