Summary of the AWS Service Event in the Northern Virginia (US-EAST-1) Region

December 10th, 2021

We want to provide you with some additional information about the service disruption that occurred in the Northern Virginia (US-EAST-1) Region on December 7th, 2021.

Issue Summary

To explain this event, we need to share a little about the internals of the AWS network. While the majority of AWS services and all customer applications run within the main AWS network, AWS makes use of an internal network to host foundational services including monitoring, internal DNS, authorization services, and parts of the EC2 control plane. Because of the importance of these services in this internal network, we connect this network with multiple geographically isolated networking devices and scale the capacity of this network significantly to ensure high availability of this network connection. These networking devices provide additional routing and network address translation that allow AWS services to communicate between the internal network and the main AWS network. At 7:30 AM PST, an automated activity to scale capacity of one of the AWS services hosted in the main AWS network triggered an unexpected behavior from a large number of clients inside the internal network. This resulted in a large surge of connection activity that overwhelmed the networking devices between the internal network and the main AWS network, resulting in delays for communication between these networks. These delays increased latency and errors for services communicating between these networks, resulting in even more connection attempts and retries. This led to persistent congestion and performance issues on the devices connecting the two networks.

This congestion immediately impacted the availability of real-time monitoring data for our internal operations teams, which impaired their ability to find the source of congestion and resolve it. Operators instead relied on logs to understand what was happening and initially identified elevated internal DNS errors. Because internal DNS is foundational for all services and this traffic was believed to be contributing to the congestion, the teams focused on moving the internal DNS traffic away from the congested network paths. At 9:28 AM PST, the team completed this work and DNS resolution errors fully recovered. This change improved the availability of several impacted services by reducing load on the impacted networking devices, but did not fully resolve the AWS service impact or eliminate the congestion. Importantly, monitoring data was still not visible to our operations team so they had to continue resolving the issue with reduced system visibility. Operators continued working on a set of remediation actions to reduce congestion on the internal network including identifying the top sources of traffic to isolate to dedicated network devices, disabling some heavy network traffic services, and bringing additional networking capacity online. This progressed slowly for several reasons. First, the impact on internal monitoring limited our ability to understand the problem. Second, our internal deployment systems, which run in our internal network, were impacted, which further slowed our remediation efforts. Finally, because many AWS services on the main AWS network and AWS customer applications were still operating normally, we wanted to be extremely deliberate while making changes to avoid impacting functioning workloads. As the operations teams continued applying the remediation actions described above, congestion significantly improved by 1:34 PM PST, and all network devices fully recovered by 2:22 PM PST.

We have taken several actions to prevent a recurrence of this event. We immediately disabled the scaling activities that triggered this event and will not resume them until we have deployed all remediations. Our systems are scaled adequately so that we do not need to resume these activities in the near-term. Our networking clients have well tested request back-off behaviors that are designed to allow our systems to recover from these sorts of congestion events, but, a latent issue prevented these clients from adequately backing off during this event. This code path has been in production for many years but the automated scaling activity triggered a previously unobserved behavior. We are developing a fix for this issue and expect to deploy this change over the next two weeks. We have also deployed additional network configuration that protects potentially impacted networking devices even in the face of a similar congestion event. These remediations give us confidence that we will not see a recurrence of this issue.

AWS Service Impact

While AWS customer workloads were not directly impacted from the internal networking issues described above, the networking issues caused impact to a number of AWS Services which in turn impacted customers using these service capabilities. Because the main AWS network was not affected, some customer applications which did not rely on these capabilities only experienced minimal impact from this event.

Several AWS services experienced impact to the control planes that are used for creating and managing AWS resources. These control planes use services hosted in the internal network. For example, while running EC2 instances were unaffected by this event, the EC2 APIs that customers use to launch new instances or to describe their current instances experienced increased error rates and latencies starting at 7:33 AM PST. By 1:15 PM PST, as congestion was improving, EC2 API error rates and latencies began to improve, except for launches of new EC2 instances, which recovered by 2:40 PM PST. Customers of AWS services like Amazon RDS, EMR, Workspaces would not have been able to create new resources because of the inability to launch new EC2 instances during the event. Similarly, existing Elastic Load Balancers remained healthy during the event, but the elevated API error rates and latencies for the ELB APIs resulted in increased provisioning times for new load balancers and delayed instance registration times for adding new instances to existing load balancers. Additionally, Route 53 APIs were impaired from 7:30 AM PST until 2:30 PM PST preventing customers from making changes to their DNS entries, but existing DNS entries and answers to DNS queries were not impacted during this event. Customers also experienced login failures to the AWS Console in the impacted region during the event. Console access was fully restored by 2:22 PM PST. Amazon Secure Token Service (STS) experienced elevated latencies when providing credentials for third party identity providers via OpenID Connect (OIDC). This resulted in login failures for other AWS services that utilize STS for authentication, such as Redshift. While latencies improved at 2:22 PM PST when the issue affecting network devices was addressed, full recovery for STS occurred at 4:28 PM PST.

Customers were also impacted by CloudWatch monitoring delays throughout this event and, as a result, found it difficult to understand impact to their applications. A small amount of CloudWatch monitoring data was not captured during this event and may be missing from some metrics for parts of the event.

Customers accessing Amazon S3 and DynamoDB were not impacted by this event. However, access to Amazon S3 buckets and DynamoDB tables via VPC Endpoints was impaired during this event.

AWS Lambda APIs and invocation of Lambda functions operated normally throughout the event. However, API Gateway, which is often used to invoke Lambda functions as well as an API management service for customer applications, experienced increased error rates. API Gateway servers were impacted by their inability to communicate with the internal network during the early part of this event. As a result of these errors, many API Gateway servers eventually got into a state where they needed to be replaced in order to serve requests successfully. This normally happens through an automated recycling process, but this was not possible until the EC2 APIs began recovering. While API Gateways began seeing recovery at 1:35 PM PST, errors and latencies remained elevated as API Gateway capacity was recycled by the automated process working through the backlog of affected servers. The service largely recovered by 4:37 PM PST, but API Gateway customers may have continued to experience low levels of errors and throttling for several hours as API Gateways fully stabilized. The API Gateway team is working on a set of mitigations to ensure that API Gateway servers remain healthy even when the internal network is unavailable and making improvements to the recycling process to speed recovery efforts in the event of a similar issue in the future. EventBridge, which is also often used in conjunction with Lambda, experienced elevated errors during the initial phases of the event but saw some improvement at 9:28 AM PST when the internal DNS issue was resolved. However, during mitigation efforts to reduce the load on the affected network devices, operators disabled event delivery for EventBridge at 12:35 PM. Event delivery was re-enabled at 2:35 PM PST, however the service experienced elevated event delivery latency until 6:40 PM PST as it processed the backlog of events.

The AWS container services, including Fargate, ECS and EKS, experienced increased API error rates and latencies during the event. While existing container instances (tasks or pods) continued to operate normally during the event, if a container instance was terminated or experienced a failure, it could not be restarted because of the impact to the EC2 control plane APIs described above. At 1:35 PM PST, most of the container-related API error rates returned to normal, but Fargate experienced increased request load due to the backlog of container instances that needed to be started, which led to continued elevated error rates and Insufficient Capacity Errors as container capacity pools were being replenished. At 5:00 PM PST, Fargate API error rates began to return to normal levels. Some customers saw elevated Insufficient Capacity Errors for “4 vCPU” task sizes for several hours following recovery.

Amazon Connect experienced elevated failure rates for handling phone calls, chat sessions, and task contacts during the event. Issues with API Gateways used by Connect for the execution of Lambda functions resulted in elevated failure rates for inbound phone calls, chat sessions or task contacts. At 4:41 PM PST, when the affected API Gateway fully recovered, Amazon Connect resumed normal operations.

Event Communication

We understand that events like this are more impactful and frustrating when information about what’s happening isn’t readily available. The impairment to our monitoring systems delayed our understanding of this event, and the networking congestion impaired our Service Health Dashboard tooling from appropriately failing over to our standby region. By 8:22 AM PST, we were successfully updating the Service Health Dashboard. As the impact to services during this event all stemmed from a single root cause, we opted to provide updates via a global banner on the Service Health Dashboard, which we have since learned makes it difficult for some customers to find information about this issue. Our Support Contact Center also relies on the internal AWS network, so the ability to create support cases was impacted from 7:33 AM until 2:25 PM PST. We have been working on several enhancements to our Support Services to ensure we can more reliably and quickly communicate with customers during operational issues. We expect to release a new version of our Service Health Dashboard early next year that will make it easier to understand service impact and a new support system architecture that actively runs across multiple AWS regions to ensure we do not have delays in communicating with customers.

In closing

Finally, we want to apologize for the impact this event caused for our customers. While we are proud of our track record of availability, we know how critical our services are to our customers, their applications and end users, and their businesses. We know this event impacted many customers in significant ways. We will do everything we can to learn from this event and use it to improve our availability even further.