Azure status history

This page contains all root cause analyses (RCAs) for incidents that occurred on November 20, 2019 or later. Each RCA will be retained on this page for 5 years. RCAs before November 20, 2019 aren't available.

Product: All

Region: All

Date: All

February 2021

2/12 RCA - Azure Cosmos DB connectivity issues affecting downstream services in West US region (Tracking ID CVTV-RB0)

Summary of Impact: Between February 11, 23:23 UTC and February 12, 04:30 UTC, a subset of customers using Azure Cosmos DB in West US may have experienced issues connecting to resources. Additionally, other Azure services that leverage Azure Cosmos DB may have also seen downstream impact during this time. The Cosmos DB outage affected user application requests to West US. A small subset of customers using Cosmos DB in other regions saw an impact on their replication traffic into West US. Customer impact for Azure Cosmos DB accounts was dependent on the Geo-Replication configurations in place:

- Accounts with no Geo-Replication: Read and write requests failed for West US
- Accounts with Geo-Replicated Single-Write + Multiple-Read regions: Read and write requests failed for West US. The Cosmos DB client SDK automatically redirected read requests to a healthy region – an increased latency may have been observed due to longer geographic distances
- Accounts with Geo-Replicated Multiple Write + Read regions: Read and write requests may have failed in West US. The Cosmos DB client SDK automatically redirected read and write requests to a healthy region – an increased latency may have been observed due to longer geographic distances.

Root Cause: On February 11, 10:04 UTC (approximately thirteen hours before the incident started), a Cosmos DB deployment was completed in West US using safe deployment practices; unfortunately, it introduced a code regression that triggered at 23:11 UTC, resulting in the customer impact described above.

A rare failure condition in the configuration store for one of the West US clusters was encountered. The front-end service (which is responsible for request routing of customer traffic) should handle this. Due to the code regression, the cluster's front-end service failed to address the condition and crashed.

Front-end services for other clusters in the region also make calls to the impacted cluster's front-end service to obtain configuration. These calls were timed out because of unavailability, triggering the same unhandled failure condition and resulting crash. This cascading effect impacted most West US Cosmos DB front-end services. Cosmos DB customers in the region would have observed this front-end service outage as a loss of availability.

Mitigation: Cosmos DB internal monitoring detected the failures and triggered high severity alerts. The appropriate teams responded to these alerts immediately and began investigating. During the triage process, Engineers noted that the configuration store's failure condition (which led to the unhandled error) was uncommon and not triggered in any other clusters worldwide.

The team applied a configuration change to disable the offending code causing the process crashes. Automated service recovery then restored all cluster operations.

Next Steps: We apologize for the impact on affected customers. We are continuously taking steps to improve the Microsoft Azure Platform and our processes to help ensure such incidents do not occur in the future. In this case, this includes (but is not limited to):

- Expediting roll out of a hotfix for the Cosmos DB Gateway application to isolate failures for internal metadata requests to reduce the regional and inter-regional impact
- Improving Cosmos DB monitoring to detect unhandled failures
- Improving the Cosmos DB front-end service to remove dependencies on current configuration store in steady-state
- Improving publicly available documentation, with the intent of providing more straightforward guidance on the actions customers can take with each account configuration type in the event of partial, regional, or availability zone outages
- Improving Cosmos DB automated failover logic to accelerate failover progress due to partial regional outages

Provide Feedback: Please help us improve the Azure customer communications experience by taking our survey: https://aka.ms/AzurePIRSurvey

January 2021

1/15 Azure Network Infrastructure service availability issues for customers located in Argentina - Mitigated (Tracking ID DM75-VCB)

Summary of Impact: Between 17:30 and 20:15 UTC on 15 Jan 2021, customers located in Argentina attempting to access the Azure Portal and/or Azure Resources may have experienced degraded performance, network drops, or timeouts. Customers may also have experienced downstream impact to dependent Azure services due to an underlying networking event.

Preliminary Root Cause: We determined that a network device, affecting network traffic in Argentina, experienced a hardware fault and that network traffic was not automatically rerouted.

Mitigation: We took the faulty network device out of rotation and rerouted network traffic to mitigate the issue.

Next Steps: We will continue to investigate to establish the full root cause and prevent future occurrences. Stay informed about Azure service issues by creating custom service health alerts: https://aka.ms/azbi-videos for video tutorials and https://aka.ms/ash-alerts for how-to documentation.

December 2020

12/14 RCA - Azure Active Directory - Authentication errors (Tracking ID PSDT-7P0)

Summary of impact: Between 08:01 and 09:20 UTC on 14 Dec 2020, a subset of users in Europe might have encountered errors while authenticating to Microsoft services and third-party applications. Impacted users would have seen the error message: "AADSTS90033: A transient error has occurred. Please try again". The impact was isolated to users who were served through one specific back end scale unit in Europe. Availability for Azure Active Directory (AD) authentication in Europe dropped to a 95.85% success rate during the incident. Availability in regions outside of Europe region remained within Service Level Agreement (SLA).

Root Cause: The Azure AD back end is a geo-distributed and partitioned cloud directory store. The back end is partitioned into many scale units with each scale unit having multiple storage units distributed across multiple regions. Request processing for one of the back end scale units experienced high latency and timeouts due to high thread contention. The thread contention happened on the scale unit due to a particular combination of requests and a recent change in service topology for the scale unit rolled out previously.

Mitigation: To mitigate the problem, engineers updated the backend request routing to spread the requests to additional storage units. Engineers also rolled back the service topology change that triggered high thread contention.

Next Steps: We apologize for the impact to affected customers. We are continuously taking steps to improve the Microsoft Azure Platform and our processes to help ensure such incidents do not occur in the future. In this case, this includes (but is not limited to):

- Augment existing load testing to validate the combination of call patterns that caused the problem.
- Further root cause the reason for thread contention and make necessary fixes before re-enabling the service topology change.

Provide Feedback: Please help us improve the Azure customer communications experience by taking our survey: https://aka.ms/AzurePIRSurvey

October 2020

10/27 RCA - Azure Active Directory B2C - North Europe / West Europe (Tracking ID 8SHB-PD0)

Summary of Impact: Between 08:40 UTC and 11:10 UTC on 27 Oct 2020, a subset of customers using Azure Active Directory B2C (AAD B2C) in North Europe/West Europe may have experienced errors when connecting to the service. Customers may have received an HTTP status code 502 (Bad Gateway) or HTTP status code 504 (Gateway Timeout).

Root Cause: In the North Europe/West Europe regions a configuration change was compounded by a surge in traffic which exceeded the regions' operational thresholds and required the Azure AD B2C Service to be augmented.

Mitigation: We performed a change to the service configuration, routing all traffic for the affected regions to an alternate production environment. This production environment, which was located in the same regions, had the necessary operational thresholds and measures in place.

Next Steps: We apologize for the impact to affected customers. We are continuously taking steps to improve the Microsoft Azure Platform and our processes to help ensure such incidents do not occur in the future. In this case, this includes (but is not limited to):

- Ensuring that the affected regions' operational thresholds are set appropriately for the service.
- Thorough testing of the new environment to ensure that it operates and scales as expected.
- Reviewing our monitoring/alerts and making adjustments to ensure that proximity to operational thresholds is detected much earlier, enabling us to take proactive action to prevent such issues.
- Ensuring that failover systems are in place to allow for more rapid routing of traffic between environments.

Provide Feedback: Please help us improve the Azure customer communications experience by taking our survey: https://aka.ms/AzurePIRSurvey

10/19 RCA - Azure Resource Manager - Issues accessing Azure resources via ARM (Tracking ID 2LXD-HT8)

Summary of Impact: Between 19:07 UTC and 22:20 UTC on 19 Oct 2020, a subset of customers using resources that leverage Azure Resource Manager (ARM) may have received intermittent errors while accessing or performing service management operations - such as create, update, delete - for multiple resources from the Azure portal or when using CLI.

Root Cause: The issue was caused by a misconfiguration in the broad phase of a deployment for ARM services, which resulted in unanticipated utilization of a single partition of Cosmos DB. The impact period was due to the normal organic increase in requests exceeding limits for that single Cosmos DB partition, which triggered throttling on those requests, and as a result, the failures or errors were received for those ARM requests. We were alerted to impact based on internal telemetry at 19:07 UTC and commenced investigation. By 20:30 UTC the impact had become more widespread.

During integration testing and in the early phases of the rollout, in-line with safe deployment practices, the deployment did not show any problems or regression.

Mitigation: A recent deployment was identified as the likely root cause. In parallel, teams worked to disable the calls to Cosmos DB, which were introduced by the deployment while also scaling up the Cosmos DB instance, which collectively mitigated the impact. By 21:15 UTC telemetry showed the expected decrease in errors and by 22:20 UTC impact had subsided.

Next Steps: We apologize for the impact to affected customers. We are continuously taking steps to improve the Microsoft Azure Platform and our processes to help ensure such incidents do not occur in the future. In this case, this includes (but is not limited to):

- Investigate auto-scaling and other resiliency techniques for Cosmos DB and other dependencies.
- Review and ensure proactive monitoring procedures include expected thresholds for Cosmos DB and dependent services in test, Pilot and Early phases of deployment.
- Review procedures, and create additional automated rules to catch this class of misconfiguration in the code during testing phase.

Provide Feedback: Please help us improve the Azure customer communications experience by taking our survey: https://aka.ms/AzurePIRSurvey

10/7 RCA - Issues accessing Microsoft and Azure services (Tracking ID 8TY8-HT0)

Summary of Impact: Between 18:20 UTC and 18:42 UTC on 07 Oct 2020, a subset of customers may have encountered increased latency, packet loss, failed connections and authentication failures across multiple Azure services. Retries may have succeeded during this time and users who had authenticated prior to the impact start time were less likely to experience authentication issues.

Network resources were restored at 18:42 UTC. Azure services began auto-mitigation. While other services had to undergo manual intervention to recover this could have led to varying times of recovery for Azure services. By 21:30 UTC it was confirmed that all Azure services had recovered.

Root Cause: The incident was caused by a code defect in a version update of a component that controls network traffic routing between Azure regions. Because the main parameters of the new code were invoked only at production scale and scope levels, the pre-production validation process did not flag an issue. Following the deployment into production, the code defect prevented anomaly detection from occurring, which normally would catch an abnormal, sudden increase in the number of unhealthy devices and force a health validation of those devices before removing routes from the network. In this instance, due to the prevention of the anomaly detection process, the Wide Area Network Software Defined Network (WAN SDN) controller removed the corresponding routes to these devices from the network. This code defect was triggered 1 hour after rollout of the service update at 18:20 UTC and caused traffic to see abnormal routes, in turn causing network congestion and packet loss.

Mitigation: The WAN SDN controller automatically recovered after a transient issue with health signals improved. The controller validated full health of network devices and then added the routes back on the devices, mitigating the network issue by 18:42 UTC. Affected Azure services began to auto-mitigate shortly thereafter, including Azure AD which recovered by 18:45 UTC. To prevent recurrence, we rolled back the software change to use the previous version of the traffic routing system.

Next Steps: We apologize for the impact to affected customers. We are continuously taking steps to improve the Microsoft Azure Platform and our processes to help ensure such incidents do not occur in the future. In this case, this includes (but is not limited to) the following:

- Improving resiliency of fixed scores of network state. Preventing bad data from propagating through the SDN controller pipeline through additional anomaly detection.
- Increase the length of time new versions of service run in pre-production before global deployment.
- Increase test coverage in the virtualized environment (Open Network Emulator) that emulates production network and improve the SDN controller resiliency to transients that occur naturally in the virtualized environment and to new injected faults.

Provide Feedback: Please help us improve the Azure customer communications experience by taking our survey: https://aka.ms/AzurePIRSurvey

10/6 Azure Front Door - Mitigated (Tracking ID BKND-JP8)

Summary of Impact: Between 17:00 and 21:19 UTC on 06 Oct 2020, a subset of customers may have experienced traffic routing to unhealthy backends.

Preliminary Root Cause: A configuration change was deployed, causing the incorrect routing of traffic to unhealthy backends.

Mitigation: We reverted the recent change to a previous healthy configuration.

Next Steps: We will continue to investigate to establish the full root cause and prevent future occurrences.

September 2020

9/28 RCA - Authentication errors across multiple Microsoft services and Azure Active Directory integrated applications (Tracking ID SM79-F8B)

Summary of Impact: Between approximately 21:25 UTC on September 28, 2020 and 00:23 UTC on September 29, 2020, customers may have encountered errors performing authentication operations for all Microsoft and third-party applications and services that depend on Azure Active Directory (Azure AD) for authentication. Applications using Azure AD B2C for authentication were also impacted.

Users who were not already authenticated to cloud services using Azure AD were more likely to experience issues and may have seen multiple authentication request failures corresponding to the average availability numbers shown below. These have been aggregated across different customers and workloads.

- Europe: 81% success rate for the duration of the incident.
- Americas: 17% success rate for the duration of the incident, improving to 37% just before mitigation.
- Asia: 72% success rate in the first 120 minutes of the incident. As business-hours peak traffic started, availability dropped to 32% at its lowest.
- Australia: 37% success rate for the duration of the incident.

Service was restored to normal operational availability for the majority of customers by 00:23 UTC on September 29, 2020; however, we observed infrequent authentication request failures which may have impacted customers until 02:25 UTC.

Users who had authenticated prior to the impact start time were less likely to experience issues depending on the applications or services they were accessing.

Resilience measures in place protected Managed Identities services for Virtual Machines, Virtual Machine Scale Sets, and Azure Kubernetes Services with an average availability of 99.8% throughout the duration of the incident.

Root Cause: On September 28 at 21:25 UTC, a service update targeting an internal validation ring was deployed, causing a crash upon startup in the Azure AD backend services. A latent code defect in the Azure AD backend service Safe Deployment Process (SDP) system caused this to deploy directly into our production environment, bypassing our normal validation process.

Azure AD is designed to be a geo-distributed service deployed in an active-active configuration with multiple partitions across multiple data centers around the world, built with isolation boundaries. Normally, changes initially target a validation ring containing non-customer data, followed by an inner ring that contains Microsoft only users, and lastly our production environment. These changes are deployed in phases across five rings over several days.

In this case, the SDP system failed to correctly target the validation test ring due to a latent defect that impacted the system's ability to interpret deployment metadata. Consequently, all rings were targeted concurrently. The incorrect deployment caused service availability to degrade.

Within minutes of impact, we took steps to revert the change using automated rollback systems which would normally have limited the duration and severity of impact. However, the latent defect in our SDP system had corrupted the deployment metadata, and we had to resort to manual rollback processes. This significantly extended the time to mitigate the issue.

Mitigation: Our monitoring detected the service degradation within minutes of initial impact, and we engaged immediately to initiate troubleshooting. The following mitigation activities were undertaken:

- The impact started at 21:25 UTC, and within 5 minutes our monitoring detected an unhealthy condition and engineering was immediately engaged.
- Over the next 30 minutes, in concurrency with troubleshooting the issue, a series of steps were undertaken to attempt to minimize customer impact and expedite mitigation. This included proactively scaling out some of the Azure AD services to handle anticipated load once a mitigation would have been applied and failing over certain workloads to a backup Azure AD Authentication system.
- At 22:02 UTC, we established the root cause, began remediation, and initiated our automated rollback mechanisms.
- Automated rollback failed due to the corruption of the SDP metadata. At 22:47 UTC we initiated the process to manually update the service configuration which bypasses the SDP system, and the entire operation completed by 23:59 UTC.
- By 00:23 UTC enough backend service instances returned to a healthy state to reach normal service operational parameters.
- All service instances with residual impact were recovered by 02:25 UTC.

Next Steps: We sincerely apologize for the impact to affected customers. We are continuously taking steps to improve the Microsoft Azure Platform and our processes to help ensure such incidents do not occur in the future. In this case, this includes (but is not limited to) the following actions:

We have already completed

- Fixed the latent code defect in the Azure AD backend SDP system.
- Fixed the existing rollback system to allow restoring the last known-good metadata to protect against corruption.
- Expand the scope and frequency of rollback operation drills.

The remaining steps include

- Apply additional protections to the Azure AD service backend SDP system to prevent the class of issues identified here.
- Expedite the rollout of Azure AD backup authentication system to all key services as a top priority to significantly reduce the impact of a similar type of issue in the future.
- Onboard Azure AD scenarios to the automated communications pipeline which posts initial communication to affected customers within 15 minutes of impact.

Provide Feedback: Please help us improve the Azure customer communications experience by taking our survey: https://aka.ms/AzurePIRSurvey

9/18 RCA - Azure Storage Premium File Shares - East US (Tracking ID SMSC-F50)

Summary of Impact: Between 11:30 UTC and 19:51 UTC on 18 Sep 2020, a subset of customers using Azure Storage Premium File Shares in East US may have experienced issues accessing services. Other downstream services may have seen impact or experienced service degradation.

Root Cause: On a single storage scale unit in East US, a feature was applied to optimize the performance of IO operations. The feature contained a code bug in an infrequent error path, which when hit would cause a storage front end process to become unhealthy. The incident started when a small number of clients entered an invalid state, triggered by a combination of a routine network maintenance operation which happened on the storage scale unit at the same time and a code bug on the client side. This caused the faulty error path to be hit more frequently. The series of events led to multiple front ends becoming unhealthy, which resulted in failed requests and increased latencies for the duration of the incident.

Mitigation: We mitigated the incident by applying a configuration change to disable the performance optimization feature that introduced the bug. Once the front end processes became healthy again, we applied another configuration change to balance the load across the front ends in order to speed up the recovery.

Next Steps: We sincerely apologize for the impact to affected customers. We are continuously taking steps to improve the Microsoft Azure Platform and our processes to help ensure such incidents do not occur in the future. In this case, this includes (but is not limited to) the following actions:

- The performance optimization feature has been temporarily disabled in other storage scale units in order to prevent similar incidents until the code fix is deployed.
- A code fix has been developed and will be validated and deployed before re-enabling the performance optimization feature.
- Improving testing and validation to help catch similar issues before they roll out to production.
- Investigating the reason why the monitoring system did not trigger an early warning alert when the front end processes started failing.

Provide Feedback: Please help us improve the Azure customer communications experience by taking our survey: https://aka.ms/AzurePIRSurvey

9/14 RCA - Connectivity Issues - UK South (Tracking ID CSDC-3Z8)

Summary of Impact: Between 13:30 UTC on 14 Sep and 00:41 UTC on 15 Sep 2020, a subset of customers in the UK South may have encountered issues connecting to Azure services hosted in this region. Customers leveraging Availability Zones and configured for zone redundancy would not have experienced a loss in service availability. In some instances, the ability to perform service management would have been impacted. Zone Redundant Storage (ZRS) remained available throughout the incident.

Root Cause and Mitigation: On 14th September 2020, a customer impacting event occurred in a single datacenter in UK South due to a cooling plant issue. The issue occurred when a maintenance activity that was being performed at our facility had the site shut down the water tower makeup pumps via their Building Automation System (BAS). This was shut down in error and was noticed at approximately13:30 UTC when our teams began to inspect the plant.

By this time, the issue had begun to impact downstream mechanical systems resulting in the electrical infrastructure that supports the mechanical systems shutting down. Microsoft operates its datacenters with 2N design meaning that we operate with a fully redundant, mirrored system. The 2N design is meant to protect against interruptions which could cause potential downtime; however, in this case, the cascading failures impacted both sides of the electrical infrastructure that supports mechanical systems. When the thermal event was detected by our internal systems, automation began to power down various resources of the Network, Storage, and Compute infrastructure to protect hardware and data durability. There were portions of our infrastructure that could not be powered down automatically (for example due to connectivity issues); some of these were shut down via manual intervention.

It took approximately 120 minutes for the team to diagnose the root cause and begin to remediate the mechanical plant issues, with cooling being restored at 15:45 UTC. By 16:30 UTC temperatures across the affected parts of the data center had returned to normal operational ranges.

Networking recovery began at approximately 16:30 UTC by beginning power-cycling network switches to recover them from the self-preservation state they entered when overheated. The recovery order was prioritized to first bring Azure management infrastructure, Storage clusters, and then Compute clusters online. When network switches providing connectivity to a set of resources were power-cycled and started to show health, engineers began recovering the other classes of resources. Network recovery was completed at 23:32 UTC. Shortly after this, any impacted Storage and Compute clusters regained connectivity, and engineers took further steps to bring any remaining unhealthy servers back online.

Next Steps: We apologize for the impact to affected customers. We are continuously taking steps to improve the Microsoft Azure Platform and our processes to help ensure such incidents do not occur in the future. In this case, this includes (but is not limited to):

- Review the logs and alarms from all affected mechanical and electrical gear to help ensure there was no damage or failed components. This is complete.
- Review and update Operational Procedure and Change Management to help ensure that the correct checks are in place and system changes via commands across systems are validated visually prior to commencement of work or return to a normal state.
- Validate and update the discrimination study for the Mechanical and Electrical systems.

Provide Feedback: Please help us improve the Azure customer communications experience by taking our survey: https://aka.ms/AzurePIRSurvey