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

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 ZLXD-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 use sub-optimal 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 recent 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 feed sources 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 8KND-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-F88)

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 test 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 that contains no 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:

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-FS0)

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 operations which happened on the storage scale unit at the 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

9/3 RCA - Network Latency Issue – West Europe (Tracking ID 8KLC-1T8)

Summary of Impact: Between 09:21 and 17:32 UTC on 03 Sep 2020, a subset of customers may have experienced intermittent latency or issues connecting to resources hosted in West Europe. Retries may have worked during this timeframe.

Root Cause: Two separate events occurred in close succession prior to the start of impact from this incident:

- Approximately 4 hours before the impact start, some local activity (likely construction) in the vicinity of the data centre cause an increase in the number of packets corrupted during transmission over fiber optic cables between data centers in the West Europe region. These errored packets were detected and dropped, and our networking automation systems took the links out of service and opened tickets with the local site to have them repaired. This is a standard process, and our automated safety checks validated that there was no impact related to this.
- Separately, between 09:21 and 09:26 UTC a significant fiber-cut occurred approximately 5 kilometres from the data centre on one one of the other paths between the data centers. This cut impacted 50% of the capacity for that route, but again, this event on its own would have no impact on traffic overall in the West Europe region.

Each of the events in isolation would have had no perceptible impact on the networking operations for West Europe, but when combined, they resulted in 9 links between data centres receiving an unequal share of traffic, becoming congested, and dropping packets (the impact was to less than 2% of the total capacity on the impacted links). Connections that travelled over these congested links would have experienced increased packet loss and latency. As connections are spread over the available links, services that retried requests by opening new connections were likely to have been unaffected and successful.

The time to mitigate was extended by the need for on-call engineers to identify that there were multiple causes for down links and identify the best way to reduce congestion and rebalance traffic. During the initial response, the large number of concurrent alerts resulted in on-call engineers taking actions that moved the congestion from one link to another, but did not resolve it.

Mitigation: Mitigation was achieved by engineers manually determining which of the links that had experienced errors could be put back into service and rebalancing traffic across the links in service. Full mitigation was declared at 17:32 UTC, but most customers would have see improvement in advance of this time. Full restoration was achieved by September 4 02:00 UTC when the significant fiber cut was repaired.

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):

- Accelerate the readiness of additional fiber paths between these data centres to reduce the impact of future fiber cuts.
- Improve the tooling used by on-call engineers when responding to complex incidents with multiple causes of downed links, so that they can reduce congestion faster and achieve mitigation more quickly.

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

August 2020

8/21 Content Delivery Network (CDN) - Service Degradation - Mitigated (Tracking ID DLYY-ND8)

Summary of Impact: Between 18:05 and 19:55 UTC on 21 Aug 2020, a subset of customers using Azure CDN from Verizon may have experienced service degradation.

Preliminary Root Cause: We determined that a recent deployment task impacted connectivity to origins, causing dynamic or cache miss requests to fail.

Mitigation: The CDN provider rolled out an update that fixed 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/ash-videos for video tutorials and https://aka.ms/ash-alerts for how-to documentation.

8/14 RCA - Degraded connectivity to Microsoft Services within the Southeast region of the United States (Tracking ID 9MDM-TT8)

Summary of Impact: Between approximately 02:20 UTC to 03:30 UTC and 04:07 UTC to 04:52 UTC on 14 Aug 2020, a subset of customers connecting through one of Microsoft's edge-nodes (https://aka.ms/MSGlobalNetwork) in the Southeast United States (US) may have experienced intermittent periods of degraded connectivity when attempting to connect to Azure, Microsoft 365, and Xbox resources.

Root Cause: Microsoft's Global Network consists of edge-nodes that connect to the Internet externally and two or more Backbone sites internally via diverse optical fiber paths for redundancy during failure scenarios.

On 14 Aug 2020 at 02:20 UTC, we experienced a dual fiber path failure isolating one of our edge-nodes in the Southeastern US. The initial fiber path incident occurred on 13 Aug 2020 at 18:34 UTC due to a fiber cut causing the path to be removed from Microsoft's Global Network. Traffic was then routed to our secondary fiber path per design. Meanwhile, our fiber provider had dispatched a technician to work on resolving the initial fiber incident. While working on that incident, the technician inadvertently disconnected our secondary fiber path at 02:20 UTC, which resulted in the secondary path to be removed from Microsoft's Global Network isolating this edge-node site.

Our network is designed to withstand site isolation and all traffic should have rerouted to the next closest edge-node in the region. However, we identified a router in this edge-node site that continued to advertise a few local prefixes to the Internet, which resulted in the blackholing of all Internet traffic destined to those prefixes in the edge-node site. The route advertisement of the local prefixes should have been withdrawn by the router when the site was isolated from Microsoft Global Network during the secondary fiber path incident but that did not occur due to a missing configuration at this site to detect site isolation and resulted in an outage. In addition, customer notification of the event was delayed due to correlation of the event and the impact.

Mitigation: The outage was mitigated when the fiber provider technician completely restored the fiber connectivity at 04:52 UTC on 14 Aug 2020.

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):

- Taking steps to prevent dual failures from occurring, reduce the degree of impact, and shorten time-to-mitigate by implementing improved failover operations to backup sites.
- Modifying our router configurations globally, to implement conditional prefix advertisement and withdrawal to ensure routers disconnect as expected during isolation events.
- Improving our alert correlation to notify fiber technicians in a timely manner, and to improve the overall notification experience.

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