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:	Region:	Date:
All	All	All

April 2021

4/20 Intermittent 503 errors accessing Azure Portal – Mitigated (Tracking ID HNS6-1SZ)

Summary of Impact: Between approximately 10:30 and 12:11 UTC, and again between 13:49 and 14:09 UTC on 20 Apr 2021, a subset of customers may have experienced intermittent HTTP 503 errors when trying to access the Azure Portal.

Preliminary Root Cause: We observed a subset of Azure Portal instances in UK West became unhealthy, causing intermittent issues accessing the Azure Portal.

Mitigation: We've removed the region from the global Azure Portal rotation to restore functionality. In tandem we also scaled-out resources in other regions to ensure there was no impact related to the load rebalancing. As part of the engineering repair work, the UK instances were briefly brought online at 13:49 UTC, but this caused a recurrence of the issue, so these have now been taken fully offline pending a full Root Cause Analysis.

Next Steps: We apologize for the impact to affected customers and will continue to investigate and establish the full root cause.

4/1 RCA - DNS issue impacting multiple Microsoft services (Tracking ID GVY5-TZZ)

Summary of Impact: Between 21:21 UTC and 22:00 UTC on 1 Apr 2021, Azure DNS experienced a service availability issue. This resulted in customers being unable to resolve domain names for services they use, which resulted in intermittent failures accessing or managing Azure and Microsoft services. Due to the nature of DNS, the impact of the issue was observed across multiple regions. Recovery time varied by service, but the majority of services recovered by 22:30 UTC.

Root Cause: Azure DNS servers experienced an anomalous surge in DNS queries from across the globe targeting a set of domains hosted on Azure. Normally, Azure's layers of caches and traffic shaping would mitigate this surge. In this incident, one specific sequence of events exposed a code defect in our DNS service that reduced the efficiency of our DNS Edge caches. As our DNS service became overloaded, DNS clients began frequent retries of their requests which added workload to the DNS service. Since client retries are considered legitimate DNS traffic, this traffic was not dropped by our volumetric spike mitigation systems. This increase in traffic led to decreased availability of our DNS service.

Mitigation: The decrease in service availability triggered our monitoring systems and engaged our engineers. Our DNS services automatically recovered themselves by 22:00 UTC. This recovery time exceeded our design goal, and our engineers prepared additional serving capacity and the ability to answer DNS queries from the volumetric spike mitigation system in case further mitigation steps were needed. The majority of services were fully recovered by 22:30 UTC. Immediately after the incident, we updated the logic on the volumetric spike mitigation system to protect the DNS service from excessive retries.

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):

- Repair the code defect so that all requests can be efficiently handled in cache.
- Improve the automatic detection and mitigation of anomalous traffic patterns.

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

March 2021

3/18 RCA - Azure Key Vault - Intermittent failures (Tracking ID 5LJ1-3CZ)

Summary of impact: Between 23:00 UTC on 18 Mar 2021 and 02:15 UTC on 19 Mar 2021, a subset of customers experienced issues and/or encountered error message "InternalServerErrors" when accessing their vaults in West Europe and North Europe regions. These errors were directly impacting customers performing operations on the Control Plane or Data Plane for Azure Key Vault or for supported scenarios that used Customer Managed Keys for encryption at rest for Azure resource providers, in which case those resources were unavailable.

Timeline:

- 3/18/2021 23:00 UTC - First Impact Observed in West Europe.
- 3/18/2021 23:10 UTC - West Europe Key Vault service fails over to North Europe.
- 3/19/2021 00:00 UTC - North Europe Vaults impacted by same issue.
- 3/19/2021 01:50 UTC - Mitigations completed by deploying new VMs. North Europe fully recovered.
- 3/19/2021 02:15 UTC - West Europe fully recovered.

Root Cause: Azure Key Vault's microservice that handles storage transactions in the West Europe region was impacted by a high CPU usage event in one of the processes that runs on the VMs. The resource utilization of the process was not constrained and it impacted underlying machines supporting Azure Key Vault service at 23:00 UTC.

This increased CPU utilization led to connection drops to other backend systems. The monitoring system detected the failures and triggered an automatic failover of the West Europe services to North Europe at 3/18/2021 23:10UTC, and Vaults in West Europe region were then limited to read-only operations. West Europe VMs for the microservice continued to be unhealthy and North Europe served the traffic for both regions.

At 3/19/2021 00:00 UTC the same background process experienced high CPU usage in North Europe. At this point the failures across both regions caused an outage for both read and write availability of Vaults in both regions. While working on the mitigation, the automated system triggered a series of failovers and failbacks which led to the service switching traffic between West Europe and North Europe as each region alternately became healthy and then unhealthy with the shifting traffic.

Mitigation: As a first measure to remediate the situation underlying VMs supporting Azure Key Vault were rebooted. However, the CPU usage continued to be high in the VMs. Engineers then deployed new VMs with higher capacity to handle the increased CPU usage and redirected traffic to them. Once this was completed both regions recovered. Also as a preliminary measure to prevent recurrence in other regions, the capacity was increased globally.

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 first precautionary measure taken was to increase the number of VMs and capacity for the microservice globally to prevent outages from a similar issue.
- New monitors have been added to watch for increased resource usage from processes.
- Failover and failback is being constrained to prevent "ping-pong" between paired regions, which extended recovery times.
- The team is working on modifying the failover pattern for the service so that the paired region is not affected by failover traffic from another region.
- We are looking into resource usage by processes on the underlying VMs that support Azure Key Vault and working to build a capacity model and restrict high resource usage by any single process.

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

3/15 RCA - Authentication errors across multiple Microsoft services (Tracking ID LN01-P8Z)

Summary of Impact: Between 19:00 UTC on March 15, 2021 and 09:37 UTC on March 16, 2021, customers may have encountered errors performing authentication operations for any Microsoft services and third-party applications that depend on Azure Active Directory (Azure AD) for authentication. Mitigation for the Azure AD service was finalized at 21:05 UTC on 15 March 2021. A growing percentage of traffic for services then recovered. Below is a list of the major services with their extended recovery times:

22:39 UTC 15 March 2021 Azure Resource Manager.
01:00 UTC 16 March 2021 Azure Key Vault (for most regions).
01:18 UTC 16 March 2021 Azure Storage configuration update was applied to first production tenant as part of safe deployment process.
01:50 UTC 16 March 2021 Azure Portal functionality was fully restored.
04:04 UTC 16 March 2021 Azure Storage configuration change applied to most regions.
04:30 UTC 16 March 2021 the remaining Azure Key Vault regions (West US, Central US, and East US 2).
09:25 UTC 16 March 2021 Azure Storage completed their recovery and we declared the incident fully mitigated.

Root Cause and Mitigation: Azure AD utilizes keys to support the use of OpenID and other Identity standard protocols for cryptographic signing operations. As part of standard security hygiene, an automated system, on a time-based schedule, removes keys that are no longer in use. Over the last few weeks, a particular key was marked as "retain" for longer than normal to support a complex cross-cloud migration. This exposed a bug where the automation incorrectly ignored that "retain" state, leading it to remove that particular key.

Metadata about the signing keys is published by Azure AD to a global location in line with Internet Identity standard protocols. Once the public metadata was changed at 19:00 UTC on 15 March 2021, applications using these protocols with Azure AD began to pick up the new metadata and stopped trusting tokens/assertions signed with the key that was removed. At that point, end users were no longer able to access those applications.

Service telemetry identified the problem, and the engineering team was automatically engaged. At 19:35 UTC on 15 March 2021, we reverted deployment of the last backend infrastructure change that was in progress. Once the key removal operation was identified as the root cause, the key metadata was rolled back to its prior state at 21:05 UTC.

Applications then needed to pick up the rolled back metadata and refresh their caches with the correct metadata. The time to mitigate for individual applications varies due to a variety of server implementations that handle caching differently. A subset of Storage resources experienced residual impact due to cached metadata. We deployed an update to invalidate these entries and force a refresh. This process completed and mitigation for the residually impacted customers was declared at 09:37 UTC on 16 March 2021.

Azure AD is in a multi-phase effort to apply additional protections to the backend Safe Deployment Process (SDP) system to prevent a class of risks including this problem. The first phase does provide protections for adding a new key, but the remove key component is in the second phase which is scheduled to be finished by mid-year. A previous Azure AD incident occurred on September 28th, 2020 and both incidents are in the class of risks that will be prevented once the multi-phase SDP effort is completed.

Next Steps: We understand how incredibly impactful and unacceptable this incident is and apologize deeply. 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 the September incident, we indicated our plans to "apply additional protections to the Azure AD service backend SDP system to prevent the class of issues identified here."

- The first phase of those SDP changes is finished, and the second phase is in a very carefully staged deployment that will finish mid-year. The initial analysis does indicate that once that is fully deployed, it will prevent the type of outage that happened today, as well as the related incident in September 2020. In the meantime, additional safeguards have been added to our key removal process which will remain until the second phase of the SDP deployment is completed.
- In that September incident we also referred to our rollout of Azure AD backup authentication. That effort is progressing well. Unfortunately, it did not help in this case as it provided coverage for token issuance but did not provide coverage for token validation as that was dependent on the impacted metadata endpoint.
- During the recent outage we did communicate via Service Health for customers using Azure Active Directory, but we did not successfully communicate for all the impacted downstream services. We have assessed that we have tooling deficiencies that will be addressed to enable us to do this in the future.
- We should have kept customers more up to date with our investigations and progress. We identified some differences in detail and timing across Azure, Microsoft 365 and Dynamics 365 which caused confusion for customers using multiple Microsoft services. We have a repair item to provide greater consistency and transparency across our services.

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

3/9 Argentina and Uruguay - Issue Accessing Azure Resources (Tracking ID 8NVQ-HD8)

Summary of Impact: Between 17:21 and 17:37 UTC on 09 Mar 2021, a network infrastructure issue occurred impacting traffic into and out of Argentina and Uruguay. During this time, customers in these areas may have experienced intermittent issues connecting to Azure resources.

Root Cause: A regional networking fiber cut resulted in a brief loss of connectivity to Microsoft resources.

Mitigation: An automated failover of network traffic to an alternative fiber route mitigated the issue.

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.

February 2021

2/26 RCA - Azure Storage and dependent services - Japan East (Tracking ID PLWV-BT0)

Summary of Impact: Between 03:26 UTC and 10:02 UTC on 26 Feb 2021, a subset of customers in Japan East may have experienced service degradation and increased latency for resources utilizing Azure Storage, including failure of virtual machine disks. Some Azure services utilizing Storage may have also experienced downstream impact.

Summary Root Cause: During this incident, the impacted storage scale unit was under heavier than normal utilization. This was due to:

- Incorrect limits set on the scale unit which allowed more load than desirable to be placed on it. This reduced the headroom that is usually available for unexpected events such as sudden spikes in growth which allows time to take load-balancing actions.
- Additionally, the load balancing automation was not sufficiently spreading the load to other scale units within the region.

This high utilization triggered heavy throttling of storage operations to protect the scale unit from catastrophic failures. This throttling resulted in failures or increased latencies for storage operations on the scale unit.

Note: The original RCA mistakenly identified a deployment as a triggering event for the increased load. This is because during an upgrade, the nodes to be upgraded are removed from rotation, temporarily increasing load on remaining nodes. An upgrade was in queue on the scale unit but had not yet started. Our apologies for the initial mistake.

Background: An internal automated load balancing system actively monitors resource utilization of storage scale units to optimize load across scale units within an Azure region. For example, resources such as disk space, CPU, memory and network bandwidth are targeted for balancing. During this load balancing, storage data is migrated to a new scale unit, validated for data integrity at the destination and finally the data is cleaned up on the source to return free resources. This automated load-balancing happens continuously and in real-time to ensure workloads are properly optimized across available resources.

Detailed Root Cause: Prior to the start of impact, our automated load-balancing system had detected high utilization on the scale-unit and was performing data migrations to balance the load. Some of these load-balancing migrations did not make sufficient progress, creating a situation where the resource utilization on the scale unit reached levels that were above the safe thresholds that we try to maintain for sustained production operation. This kick-started automated throttling on incoming storage write requests to protect the scale unit from catastrophic failures. When our engineers were engaged, they also detected that the utilization limits that were set on the scale unit to control how much data and traffic should be directed to the scale unit was higher than expected. This did not give us sufficient headroom to complete load-balancing actions to prevent customer facing impact.

Mitigation: To mitigate customer impact as fast as possible, we took the following actions:

- Engineers took steps to aggressively balance resource load out of the storage scale unit. The load-balancing migrations that were previously unable to finish were manually unblocked and completed, allowing a sizeable quantity of resources to be freed up for use. Additionally, load-balancing operations were tuned to improve its throughput to more effectively distribute load.
- We prioritized recovery of nodes with hardware failures that had been taken out of rotation to bring additional resources online.

These actions brought the resource utilization on the scale unit to a safe level which was well below throttling thresholds. Once Storage services were recovered around 06:56 UTC, dependent services started recovering. We declared full mitigation at 10:02 UTC.

Next steps: We sincerely apologize for the impact this event had on our customers. Next steps include but are not limited to:

- Optimize the maximum allowed resource utilization levels on this scale unit to provide increased headroom in the face of multiple unexpected events.
- Improve existing detection and alerting for cases when load-balancing is not keeping up, so corrective action can be triggered early to help avoid customer impact.
- Improve load-balancing automation to handle certain edge-cases under resource pressure where manual intervention is currently required to help prevent impactful events.
- Improve emergency-levers to allow for faster mitigation of impactful resource utilization related events.

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

2/16 Azure Frontdoor - Europe - Timeouts connecting to resources (Tracking ID ZN8_-VT8)

Summary of Impact: Between approximately 12:00 UTC and 13:30 UTC a subset of customers using Azure Frontdoor in Europe may have experience timeouts and/or issues connecting to resources.

Root Cause: Engineers determined that a backend network device became unhealthy, and traffic was not automatically rerouted. This resulted in Azure Front Door requests to fail.

Mitigation: We manually removed the faulty backend network device and rerouted network traffic. This mitigated the issue.

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.

2/12 RCA - Azure Cosmos DB connectivity issues affecting downstream services in West US region (Tracking ID CVTV-R80)

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 impact), 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 DM7S-VC8)

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/ash-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 PS0T-790)

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 had 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

1 2 3 4 5 6 ▸