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

April 2022

4/8

Service Management Operation Errors Across Azure Services in East US 2 (Tracking ID Y__S-9C0)

Summary of Impact: Between 12:25 UTC on 08 Apr 2022 and 14:40 UTC on 09 Apr 2022, customers running services in the East US 2 region may have experienced service management errors, delays, and/or timeouts. Customers may have experienced issues that caused GET and PUT errors impacting the Azure portal itself, as well as services including Azure Virtual Machines (VMs), Virtual Machine Scale Sets (VMSS), Azure Data Factory (ADF), Azure Databricks, Azure Synapse, Azure Backup, Azure Site Recovery (ASR), and Azure Virtual Desktop (AVD) Customers may have seen errors including "The network connectivity issue encountered for Microsoft.Compute cannot fulfill the request. For some downstream services that have auto-scale enabled, this service management issue may have caused data plane impact.

Root Cause: We determined that the Compute Resource Provider (CRP) Gateway service experienced an issue which severely reduced its throughput. The underlying issue was a retry storm triggered by the zonal failure of a related Allocator service. While we were able to recover the Allocator service by restarting the instances in the failed zone, the backlog of work exposed a potential issue with .Net CLR and Garbage Collector. This resulted in a large percentage of incoming calls to the CRP Gateway to fail. The retries triggered by the upstream services only made the load situation worse. Under normal circumstances, the Gateway instances are overprovisioned for such retry storms but the combination of the reduced throughput across all instances and continuous retries (some services which normally make 25K calls in 1 minutes were making 150K calls in the same period due to retries) resulted in a prolonged impact.

Deeper investigation into process profile data exposed that the processes were experiencing a high rate of timeout exceptions for ongoing operations and .Net Garbage Collector was overworked due to high heap churn under above mentioned spike in load. A high rate of exceptions and the simultaneous pressure on the .Net GC exposed an unfavorable interaction in the .Net runtime's process wide lock.

Mitigation: To mitigate the situation, the below steps were taken:

- Two large services were temporarily throttled more aggressively to ensure they do not continue to overload gateway.
- Once the underlying issue of the throughput reduction was partially understood, the gateway services were restarted multiple times until they got out of the wedged state.

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):

- As a long-term fix, we initiated a CRP gateway hotfix that will prevent the gateway from entering into the wedged state. The hotfix roll-out is progressing as per our Safe Deployment Practices.
- We are flighting a configuration change to make the .Net GC work less hard and avoid interaction with the process wide lock which is surfaced with exception handling.
- Repair items have been identified to optimize areas of code which were causing heap churn

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

March 2022

3/16

RCA - Azure AD B2C – Authentication Failures and Error Notifications (Tracking ID TTCR-NTZ)

Summary of Impact: Between 09:13 and 10:22 UTC on March 16, 2022, end-users of customers using Azure Active Directory B2C may have experienced errors and timeouts when attempting to sign in or sign up. Retry attempts were likely to succeed during this incident

Root Cause: The service experienced a significant increase in workload in the affected region during a planned maintenance operation. As a result, a subset of sign-in requests was queued up by the system, which resulted in timeouts and, in some cases sign in attempts by end-users may have timed out.

Mitigation: The service automatically scaled up compute resources in response to the increase in workload, which provided partial relief. In addition, we rerouted subsets of the workload to alternate capacity to achieve complete mitigation.

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):

- Improve planned maintenance Standard Operating Procedures (SOP) by pre-provisioning of additional capacity to affected regions to help handle unanticipated workload.
- Improve planned maintenance SOP to include pro-active assessment of similar pre-provisioning in other regions beyond the affected region for the service.

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

3/1

RCA - Azure Resource Manager - Service Management Operation Failures (Tracking ID ZNRZ-HDG)

Summary of Impact: Between 11:49 EST on 01 March 2022 and 03:08 EST on 03 Mar 2022, a subset of customers experienced errors when using Azure Resource Manager to perform service management operations in the Azure Government cloud.

Root Cause: A synchronization issue occurred between backend components used to permit certain ARM requests. A configuration change was applied to these backend components, which resulted in some instances of the ARM service becoming unreachable, causing errors for a subset of operation requests.

Mitigation: We rolled out a hotfix to affected components, restoring the ARM service, which allowed operation requests to complete as expected.

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):

- Update ARM component configuration methods to help prevent synchronization issues when similar updates are required.

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

February 2022

2/16

RCA - SQL Database and App Service - West Europe (Tracking ID 9TDP-N8G)

Summary of Impact: Between 07:31 UTC and 15:31 UTC on 16 Feb 2022, a subset of customers using SQL Database instances in West Europe may have experienced database connectivity errors in this region including when attempting to create new connections. Retries may have been successful.

Additionally, customers utilizing Azure App Service in West Europe may have experienced issues while performing service management operations such as site create, delete, and move resources on App Service (Web, Mobile and API Apps) applications. Autoscaling and loading site metrics may have also been impacted.

Root Cause: Due to a memory hardware failure in a network router, a control plane/data plane synchronization process in that router failed during a regular automated maintenance operation. This router was one of 8 redundant routers in that tier of the network, and the failure of the synchronization process led to the router dropping packets to a subset of the IP addresses it handled. The result was the failure of up to 12% of network flows to endpoints below the router. In particular, up to 12% of new connections to a subset of Virtual IP addresses (VIPs) would have failed. Retries would have likely succeeded and connections worked properly once established. As a downstream effect of the availability impact to some SQL services, App Service resources in the region with a dependency on SQL may have experienced issues.

The mitigation took longer than expected as the impacted device was going through automated maintenance, during which alerts were suppressed. Alerts were triggered as expected by the automated alerting system once the device was brought back into rotation.

This network device was being upgraded to a new firmware version which has, among other capabilities, the ability to automatically recover the control plane/data plane synchronization process in case of failure. This update helps prevent such failure scenarios in the future.

Mitigation:

- Engineers isolated the impacted network switch to mitigate the incident.
- Engineers verified that no other network switches were impacted due to the same issue.

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):

- Expedite the upgrade of network switches to the firmware that contains auto recovery of the control plane/data plane synchronization process.
- Enhancements to automated maintenance and alerting platforms to improve alerting for devices undergoing maintenance.

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

2/12

RCA - Azure SQL DB and Cosmos DB Unavailable (Tracking ID SL1P-TSZ)

Summary of Impact: Between 11:45 UTC on 12 Feb 2022 and 11:43 UTC on 15 Feb 2022, a limited subset of customers using SQL Databases or Cosmos DB experienced database unavailability and may have seen errors when connecting to their database instances. This issue affected a specific generation of hardware hosting SQL and Cosmos DB resources in six regions.

Root Cause: Our telemetry has shown that a subset of nodes running our newest hardware generation hosting SQL DB and Cosmos DB experienced a loss of connectivity to the network control plane starting on 12 Feb 2022. Our newest hardware generation uses a new and optimized network control plane and data plane designed to improve performance and reduce irregularities and latency, with a new control plane secured by TLS authentication. The TLS certificates that secure this channel are rotated regularly and expire within a short timespan for security purposes. A race condition in the underlying Remote Procedure Call (RPC) mechanism caused the network control plane channel in some cases to not pick up the rotated certificate, leading to connectivity failures once the certificate expired.

Routine maintenance in the general Azure compute fleet had fixed this issue with a code update in January. However, based on telemetry, we did not believe this was a significant risk to SQL and Cosmos DB environments, which were scheduled to receive the code update later in February. On 12 Feb 2022, a significant number of previously rotated certificates expired simultaneously, causing impact. We first mitigated the impacted nodes and DBs, and then pushed the code update using an emergency process through the SQL and Cosmos DB fleets to make sure the impact would not recur.

Mitigation: Once impact was identified, we recovered nodes with network control plane connectivity loss and brought DBs back to a healthy state over the course of 12 Feb - 15 Feb. While most DBs recovered earlier in the incident, running a full update to get the code fix to all nodes had to proceed slowly to ensure uptime and data integrity, so it took until 14 Feb for Cosmos DB and 15 Feb for SQL DB to complete the rollout.

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):

- Improved procedures to help determine if code updates for Azure servers are critical, so we can patch critical data services like SQL and Cosmos DB sooner, instead of taking them at a normally scheduled pace.
- Faster automated recovery and deployment for critical fixes to SQL and Cosmos DB, to help reduce impact time for critical node fixes.
- Improved telemetry to help detect certificate expiration risk.

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

2/12

Virtual Machines - West US - Resolved (Tracking ID 2S1T-LCG)

Impact Statement: Between 04:38 UTC and 6:30 UTC on 12 Feb 2022, you were identified as a customer using Virtual Machines, Azure SQL and Storage in West US who may have experienced connection failures when trying to access some resources hosted in the region. Additional downstream services may have also been impacted.

Preliminary Root Cause: We determined that a subset of storage resources experienced a drop in network connectivity.

Mitigation: We restored the network connectivity to storage resources to mitigate the issue.

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

2/2

RCA - Azure AD - Service Management Failures (Tracking ID SMWW-BDZ)

Summary of Impact: Between 19:50 UTC and 22:06 UTC on Feb 2, 2022, customers using Azure Active Directory (Azure AD) may have experienced failures when performing any service management operations.

During this incident, the Azure AD REST API service experienced some availability impact globally. However, the overall availability was at 99.996% through the duration of the incident. This means that retries had a high probability of success and customer impact would have been minimal or unnoticed.

After investigation, it was determined that there was no impact to Azure AD B2C, as previously reported.

Even though customer impact may have been minimal or unnoticed, the global nature of the incident led us to attend to this issue at a high severity and overcommunicate to ensure the potentially broad impact was notified.

Root Cause: As part of planned maintenance, a change to a dependency was rolled out which affected the availability of the Azure AD REST API service. Resiliency measures initially delayed the impact by relying on caches. When said measures were exhausted, the first request every 10 seconds resulted in a call to the dependency, causing a timeout and the request to fail.

Mitigation: Engineers rolled back the change to the dependency, mitigating the incident.

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):

- Improve monitoring of the dependency to help ensure that its availability is more accurately observed regardless of underlying resiliency measures, which would help prevent customer impact much earlier.
- The dependency on the service that caused the impact is being re-evaluated to determine if it can be removed from the Azure AD REST API.

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

January 2022

1/13

RCA - Azure Resource Manager - Issues with management and resource operations (Tracking ID 8V39-P9Z)

Summary of Impact: Between 07:00 UTC on 13 Jan 2022 and 20:00 UTC on 14 Jan 2022, a subset of customers using Azure Resource Manager (ARM) to deploy, modify, or remove Azure resources experienced delays, timeouts, and failures which were visible for long running operations executed on the platform. Impact was most severe for a period of 5 hours starting at 15:30 UTC on Jan 13 and another period of 8 hours starting at 00:00 UTC on Jan 14, and in regions including but not limited to West US, West US 2, South Central US, North Europe, West Europe, East Asia and Southeast Asia.

Impact to customers will have been broad, as numerous Azure services rely on service management operations orchestrated by the ARM platform. Most customers will have experienced delays and timeouts, but many customers will have seen deployment or resource management failures.

Root Cause: A code modification which started rolling out on 6 Jan 2022 exposed a latent defect in the infrastructure used to process long running operations (informally, "jobs"). The code modification resulted in an exception for a tiny fraction of job executions, each one of them disabled a small part of the job execution infrastructure. Over the course of hours, the job executions shifted entirely away from the regions that received the new code to their backup paired regions. For a period of 16 hours, there was no customer impact as the backup paired regions executed the jobs as intended. The impact spread to backup paired regions as the new code was deployed, resulting in job queue up, latency delays, and timeouts. In some cases, the jobs executed with such prolonged delays that they were unable to succeed, and customers will have seen failures in these cases.

As a result of the way that the job execution infrastructure was implemented, the compounding failures were not visible in our telemetry - leading to engineer's mis-identifying the cause initially and attempting rollbacks which did not improve the underlying health of the service. The consequence of this was a second period of impact starting at 00:00 UTC on 14 Jan 2022 and extending for approximately 8 hours.

Mitigation: Identifying the source of the problems in this case took time, as some parts of the job infrastructure remained healthy and processing jobs, while other key parts were being disabled. At the time we were unable to clearly identify the newly released code as correlating with the impact we were seeing. When the nature of the problem became clear we immediately started to roll back to a previous build. This change rolled out progressively completed at 20:00 UTC on 14 Jan 2022.

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):

- Reviewing and improving our monitoring and alerting strategy for our job execution infrastructure to improve our ability to detect problems like this one before they become customer-impacting.
- Fixing the underlying problem which allows a single rare exception to disable parts of the job execution infrastructure.
- Providing better visibility for operators when a paired region has assumed responsibility for job execution, in order to indicate a reduced-redundancy state and signal the need to pause or roll back a deployment.

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

1/13

Azure Data Factory V2 - West Europe - Mitigated (Tracking ID PKJ8-TTZ)

Summary of Impact: Between 10:10 UTC on 13 Jan 2022 and 20:00 UTC on 14 Jan 2022, a subset of customers may have experienced issues, timeouts, or failures for some service management operations for services leveraging Azure Resource Manager (ARM). This could have also included issues with operations attempted to manage resources or resource groups. This could have resulted in a downstream impact on other Azure services that rely on Azure Resource Manager.

Preliminary Root Cause: We have identified a change to backend role instances leveraged by Azure Resource Manager causing the timeouts and is root cause of the failure.

Mitigation: We mitigated background job execution systems causing failures and performed a roll back of recent change following our safe deployment practices (SDP) to return ARM to a previous healthy state, impacted the issue. The roll back took several hours to complete globally following our SDP process.

Next Steps: We will also continue to investigate to establish the full root cause and prevent future occurrences. You can stay informed about Azure service issues, maintenance events, or advisories by creating custom service health alerts (<https://aka.ms/ash-videos" class="wa-link-status">https://aka.ms/ash-videos" rel="noopener noreferrer" target="_blank">https://aka.ms/ash-videos" class="wa-link-status">https://aka.ms/ash-videos for video tutorials and https://aka.ms/ash-alerts" class="wa-link-status">https://aka.ms/ash-alerts" rel="noopener noreferrer" target="_blank">https://aka.ms/ash-alerts" class="wa-link-status">https://aka.ms/ash-alerts for how-to documentation) and you will be notified via your preferred communication channel(s).

1/4

Azure Cosmos DB - East US (Tracking ID 9VT8-HPG)

Summary of Impact: Between 12:30 UTC on 04 Jan 2022 and 7:41 UTC on 5 Jan 2022, customers with Azure Cosmos DB accounts in East US may have experienced connectivity and service availability errors while accessing their Cosmos DB databases. One Cosmos DB cluster in the East US region was unavailable during this time, so both new and existing connections to databases in this subscription in this region may have resulted in errors or timeouts.

Root Cause: Cosmos DB uses Azure Service Fabric as the underlying platform for providing fault tolerance in the cluster. Service Fabric uses the ring topology, and each node establishes a lease relationship with nodes in its proximity (i.e. neighborhood) to detect failure. It has a set of nodes that are responsible for determining cluster memberships of other nodes, known as Arbitrators. A node that fails to refresh lease within a timeout period will be reported by its neighbors, and the arbitrators need to determine whether the node should leave the cluster. This check is done in a timer callback.

During this incident, the timer callback on one of the nodes was fired multiple times at a frequency higher than intended. This resulted in the node's neighbors getting incorrectly reported as unavailable. By design, the Arbitrators trusted this information as they did not receive any healthy uptime notification within the stipulated time frame. This continued until the quorum of nodes was lost, and the cluster went down eventually. The cluster came back up once the culprit node was manually rebooted as part of the mitigation efforts.

Mitigation: After the initial investigation, the cluster was marked as offline at 14:08 UTC on 04 Jan 2202 which triggered regional failover for accounts that had multiple regions and automatic failover enabled. Customers that did not have automatic failover enabled continued to be impacted until the cluster was recovered.

The cluster was recovered by rebooting the Service Fabric Infrastructure nodes after removing the culprit node. However, recovery of the cluster was delayed due to overload of the configuration store as the service was restarting. Cosmos DB Engineers initially tried to reduce the load on the configuration store by delaying the startup of about 20% of the nodes. This approach did not fully resolve the problems. Engineers then manually applied configuration changes to increase the timeout on the requests used to fetch data from the configuration store. This change allowed the recovery to continually make progress. Availability to the cluster was incrementally restored as service back end processes started running. Recovery was completed at 07:41 UTC on 05 Dec 2022.

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):

Service Fabric team to develop a fix to improve resilience in case of misfired timer(s) reporting incorrect node health status within Azure Service Fabric.

Azure Cosmos DB to improve monitoring to better identify culprit nodes early on if this failure pattern reoccurs.

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