Azure status history

Product: All

Region: All

Date: Most recent

May 2019

5/22 RCA - Service Management Operations - West Europe

Summary of Impact: Between 15:10 and 21:00 UTC on 22 May 2019, a subset of customers in West Europe experienced intermittent service management delays or failures for resources hosted in this region. Impacted services included Azure Databricks, Azure Backup, Cloud Shell, HDInsight, and Virtual Machines.

Between 23:20 and 23:50 UTC on 22 May 2019, during the deployment of the permanent fix, a small subset of customers using Virtual Machines and Azure Databricks experienced increased latency or timeout failures when attempting service management operations in West Europe.

Root Cause: The issue was attributed to performance degradation in the Regional Network Manager (RNM) component of Azure software stack. The RNM component, called the partition manager, is a stateful service and has multiple replicas. This component saw an increase in latency due to a build up of replicas being created for the service. Prolonged operational delays triggered a RNM bug which caused the primary replica to re-build on two occasions. This caused service management operations to fail while one of the other replicas was taking on the primary role.

Mitigation: Engineers identified the RNM bug and applied a hotfix to the region which helped resolve network operation failures. During the outage, an security validation process on RNM nodes was performing scans which slowed the replica buildout for impacted nodes. Engineers terminated the scans to improve performance. The network operation job queues began to drain and latency returned to normal.

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):

- Develop and roll out dedicated partition for large customers (in progress)
- Implement automatic throttling to balance load (in progress)
- Root cause and fix the cause for high commit latency (in progress)

Provide feedback: Please help us improve the Azure customer communications experience by taking our survey https://aka.ms/XKB-5FZ

5/13 RCA - Network Connectivity - Increased Latency

Summary of impact: Between 09:05 and 15:33 UTC on 13 May 2019, a subset of customers in North America and Europe may have experienced intermittent connectivity issues when accessing some Azure services.

Root cause and mitigation: The impact was the result of inconsistent data replication in a networking infrastructure service. This resulted in unexpected throttling of network traffic to our name resolution servers. Once the issue was detected, engineers mitigated it by updating the configuration of the affected network infrastructure service to override the effect of this data inconsistency. Simultaneously, engineers performed operations to repair the data inconsistency.

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):

- Improve our monitoring to detect data inconsistencies similar to the one that caused this issue.
- Improvements in the system to help ensure such inconsistencies do not occur in the future.

Provide feedback: Please help us improve the Azure customer communications experience by taking our survey: https://aka.ms/7CC-HXG

5/7 SQL Services - West Europe

Summary of impact: Between 10:57 and 12:48 UTC on 07 May 2019, a subset of customers using SQL Database, SQL Data Warehouse, Azure Database for PostgreSQL, Azure Database for MySQL, Azure Database for MariaDB, in West Europe may have experienced issues performing service management operations – such as create, update, rename and delete- for resources hosted in this region.
In addition, customers may have been unable to see their list of databases using SMSS. However as this was a Service Management issue, these databases would not have been impacted (despite not being visible from SMSS).

Preliminary root cause: Engineers identified a back-end database service responsible for processing service management requests in the region became unhealthy preventing the requests from completing.

Mitigation: Engineers performed a manual restart of the impacting back-end service, which restored its capacity to process requests, mitigating the issue.

Next steps: Engineers 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

5/2 RCA - Network Connectivity - DNS Resolution

Summary of impact: Between 19:29 and 22:35 UTC on 02 May 2019, customers may have experienced connectivity issues with Microsoft cloud services including Azure, Microsoft 365, Dynamics 365 and Azure DevOps. Most services were recovered by 21:40 UTC with the remaining recovered by 22:35 UTC.

Root cause: As part of planned maintenance activity, Microsoft engineers executed a configuration change to update one of the name servers for DNS zones used to reach several Microsoft services, including Azure Storage and Azure SQL Database. A failure in the change process resulted in one of the four name servers' records for these zones to point to a DNS server having blank zone data and returning negative responses. The result was that approximately 25% of the queries for domains used by these services (such as database.windows.net) produced incorrect results, and reachability to these services was degraded. Consequently, multiple other Azure and Microsoft services that depend upon these core services were also impacted to varying degrees.

More details: This incident resulted from the coincidence of two separate errors. Either error by itself would have been non-impacting:

1) Microsoft engineers executed a name server delegation change to update one name server for several Microsoft zones including Azure Storage and Azure SQL Database. Each of these zones has four name servers for redundancy, and the update was made to only one name server during this maintenance. A misconfiguration in the parameters of the automation being used to make the change resulted in an incorrect delegation for the name server under maintenance.
2) As an artifact of automation from prior maintenance, empty zone files existed on servers that were not the intended target of the assigned delegation. This by itself was not a problem as these name servers were not serving the zones in question.

Due to the configuration error in change automation in this instance, the name server delegation made during the maintenance targeted a name server that had an empty copy of the zones. As a result, this name server replied with negative (nxdomain) answers to all queries in the zones. Since only one out of the four name server's records for the zones was incorrect, approximately one in four queries for the impacted zones would have received an incorrect negative response.

DNS resolvers may cache negative responses for some period of time (negative caching), so even though erroneous configuration was promptly fixed, customers continued to be impacted by this change for varying lengths of time.

Mitigation: To mitigate the issue, Microsoft engineers corrected the delegation issue by reverting the name server value to the previous setting. Engineers verified that all responses were then correct, and the DNS resolvers began returning correct results within 5 minutes. Some applications and services that accessed the incorrect values and cached the results may have experienced longer restoration times until the expiration of the incorrect cached information.

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):

1) Additional checks in the code that performs nameserver updates to prevent unintended changes [in progress].
2) Pre-execution modeling to accurately predict the outcome of the change and detect potential problems before execution [in progress].
3) Improve per-zone, per-nameserver monitors to immediately detect changes that cause one nameserver's drift from the others [in progress].
4) Improve DNS namespace design to better allow staged rollouts of changes with lower incremental impact [in progress].

Provide feedback: Please help us improve the Azure customer communications experience by taking our survey https://aka.ms/R50C-5RZ

5/2 RCA - Azure Map

Summary of impact: Between 04:35 and 11:00 UTC on 02 May 2019, a subset of customers using Azure Maps may have experienced 500 errors when attempting to make calls to Azure Maps Rest APIs.

Root cause and mitigation: Engineers were notified by internal monitoring that connectivity between internal components/services required by Azure Maps were disrupted. This lead to the failure to fulfill incoming customer requests. Upon investigation, engineers found that the authentication application was inadvertently removed in error during regular maintenance operations. Restoration of authentication application was performed manually, which led to the mitigation of this incident.

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):

- Moving the authentication applications to a more protected repository, and setting up fail safes to prevent future incidents.
- Augmenting monitoring for authentication issues and improving troubleshooting guides for faster response.
- Longer term, engineers are planning to move to a different authentication platform which will provide better insights earlier.

5/1 Issue signing in to https://shell.azure.com

Summary of impact: Between 18:00 UTC on 30 Apr 2019 and 23:20 UTC on 01 May 2019, customers may have experienced issues signing in to https://shell.azure.com
During this time, customers were able to access Cloud Shell through the Azure portal at https://portal.azure.com

Preliminary root cause: Engineers identified a mis-match between a configuration file which had been recently updated and its corresponding code in shell.azure.com

Mitigation: The Cloud Shell team developed, tested, and rolled out a new build which addressed and corrected the issue.

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

April 2019

4/19 RCA - Availability degradation for Azure DevOps

Summary of impact: Between 03:30 and 15:20 UTC, and then again between 17:00 and 17:32 UTC on the 19 Apr 2019, a subset of customers experienced issues connecting to Azure DevOps. These issues primarily affected customers physically located on the East Coast and those whose organizations are located on the East Coast.

Root cause: During a planned maintenance event for Azure Front Door (AFD), a configuration change caused network traffic to be incorrectly advertised. The AFD ring impacted by this maintenance hosted Azure DevOps and other Microsoft internal tenants. This may have resulted in timeouts and 500 errors for customers of Azure DevOps.
The maintenance event started at 3:30 UTC, which started dropping around 5-10% of requests. When the environment severely degraded at 14:44 UTC, engineering observed the major impact start. The maintenance event was on a ToR (Top of Rack) switch. The standard operating procedure is to take the environment offline by removing edge machines. By design, the MUX stopped advertising BGP (Border Gateway Protocol) routes and traffic is not routed through these MUX. Within this environment one of the MUX Load Balancers was in an unhealthy state but the BGP session between the load balancer and the TOR was still active. Consequently, the MUX was still active in the environment and the TOR was advertising traffic incorrectly.

Mitigation: The first impact window was mitigated by withdrawing the invalid route so that traffic would be routed correctly. The recurrence was caused by the maintenance process resetting the configuration back to the previous state, publishing an invalid route. The 2nd mitigation was re-applying the change again.

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):

- reviewing and implementing more stringent measures for when we take environments offline for maintenance events.

Provide feedback: Please help us improve the Azure customer communications experience by taking our survey https://aka.ms/WCMY-JQG

4/16 RCA - Networking Degradation - Australia Southeast / Australia East

Summary of impact: Between 07:12 and 08:02 UTC on 16 Apr 2019, a subset of customers with resources in Australia Southeast / Australia East may have experienced difficulties connecting to Azure endpoints, which in-turn may have caused errors when accessing Microsoft Services in the impacted regions.

Root cause: Microsoft received automated notification alerts that the Australia East and Australia Southeast regions were experiencing degraded network availability from a select number of Internet Service Providers (ISPs). During this time, a subset of network prefix paths changed for the select number of ISPs, this manifested in traffic not reaching the destinations within the Australia East and Australia Southeast regions. The issue stemmed from a routing anomaly due to an erroneous advertisement of prefixes received via an ExpressRoute circuit to an Internet Exchange (IX).

Mitigation: Microsoft disabled the incorrect ExpressRoute peering. The IX also identified a high amount of traffic and automatically mitigated by bringing down the peering with the IX. Once the peerings were brought down by Microsoft and the IX, availability was restored to Australia East and Australia Southeast regions.

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):

- Engage Internet Service Providers to add additional policies/protections to Internet facing routing infrastructure to block future routing anomalies [Complete]
- Add additional automated route mitigation steps within the Azure platform to reduce mitigation time [In Progress]
- Investigate further route optimizations in the Azure/Microsoft ecosystem to inherently block future routing anomalies [In Progress]

Provide feedback: Please help us improve the Azure customer communications experience by taking our survey https://aka.ms/QFVY-1FG

4/12 RCA - Cognitive Services

Summary of impact: Between 01:50 and 11:30 UTC on 12 Apr 2019 a subset of customers using Cognitive Services including Computer Vision, Face and Text Analytics in West Europe and/or West Central US may have experienced 500-level response codes, high latency and/or timeouts when connecting to resources hosted in this region.

Root cause: Engineers determined a recent deployment introduced a software regression, manifesting in increased latency across two regions.

Mitigation: The issue was not detected in pre-deployment testing, however, once manually detected, engineers proceeded to roll-back the recent deployment task to mitigate the issue.

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):

- Improve pre-deployment tests to catch this kind of issue in the future [In Progress]
- Improve monitoring to more closely represent production traffic patterns [In Progress]

Provide feedback: Please help us improve the Azure customer communications experience by taking our survey https://aka.ms/Q4JN-7FG

4/9 Virtual Machines - North Central US

Summary of impact: Between 21:39 on 9 Apr 2019 and 6:20 UTC on 10 Apr 2019, a subset of customers using Virtual Machines in North Central US may have experienced connection failures when trying to access some Virtual Machines hosted in the region. These Virtual Machines may have also restarted unexpectedly. Some residual impact was detected, impacting a small subset of recovered Virtual Machine connectivity with the underlying disk storage.

Root cause: Azure Storage team made a configuration change on 9 April 2019 at 21:30 UTC to our back-end infrastructure in North Central US to improve performance and latency consistency for Azure Disks running inside Azure Virtual Machines. This change was designed to be transparent to customers. It was enabled following our normal deployment process, first to our test environment, and lower impact scale units before being rolled out to the North Central US region. However, this region hit bugs which impacted customer VM availability. Due to a bug, VM hosts were able to establish session with the storage scale unit but hit issues when trying to receive/send data from/to storage scale unit. This situation was designed to be handled with fallback to our existing data path, but an additional bug led to failure in the fallback path and resulted in in VM reboots.

Mitigation: The system automatically recovered. Some of the customer VMs which didn't auto recover, needed an additional recovery step.

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):

- We have paused further deployment of this configuration change until the underlying bugs are fixed [complete].
- Fix bugs that caused the background operation to have customer-facing impact [in progress].
- Additional validation rigor to cover the scenario that caused the bugs to be missed in test environment [in progress].

March 2019

3/29 RCA - SQL Database

Summary of impact: Between 16:45 and 22:05 UTC on 29 Mar 2019, a subset of customers may have experienced the following:

- Difficulties connecting to SQL Database resources in the East US, UK South, and West US 2 regions
- Difficulties connecting to Service Bus and Event Hubs resources in the East US and UK South regions
- Failures when attempting service management operations for App Service resources in the UK South and East US regions
- Failures when attempting service management operations for Azure IoT Hub resources

Root cause: Azure SQL DB supports VNET service endpoints for connecting specific databases to specific VNETs. A component used in this functionality, called the virtual network plugin, runs on each VM used by Azure SQL DB, and is invoked at VM restart or reboot. A deployment of the virtual network plugin was rolling out worldwide. Deployments in Azure follow the Safe Deployment Practice (SDP), which aims to ensure deployment related incidents do not occur in many regions at the same time. SDP achieves this in part by limiting the rate of deployment for any one change. Prior to the start of the incident this particular deployment had already successfully occurred across multiple regions and for multiple days such that the deployment had reached the later stages of SDP, where changes are deployed to several regions at once. This deployment was using a VM restart capability, which occurs without impact to running workloads on those VMs.

On 5 capacity units across 3 regions, an error in the plugin load process caused the VM to fail to restart. The virtual network plugin is configured as 'required to start', as absence of it prevents key VNET service endpoint functionality from being used on that VM. The error led to repeated restart attempts causing the VMs to continuously cycle. This occurred on enough VMs across those 5 capacity units that there were not enough resources available to provide placement for all databases in those units causing those databases became unavailable. The plugin error was specific to the hardware types and configurations on the impacted capacity units.

The 5 capacity units affected included some of the databases used by Service Bus, Event Hub and App Services in those regions which led to the impact to those services. An impacted database in East US was the global service management state for Azure IoT Hub, hence the broad impact to that service.

Mitigation: Impacted databases using the Azure SQL DB AutoDR capability were failed over to resources in other regions. Some impacted databases were moved to healthy capacity within the region. Full recovery occurred when sufficient affected VMs were manually rebooted on the impacted capacity units. This brought enough healthy capacity online for all databases to become available.

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):

- Fix the error in deployment, which led to continuous recycling on the specific hardware types and configurations [in progress].
- Repair deployment block system - it stopped the deployment in each capacity unit before the entire unit became unhealthy, but not soon enough [in progress].
- Improve detection mechanism - it detected correlated impact at region level, but would have detected faster if each capacity unit was treated separately [in progress].
- Improve service resiliency for IoT Hub [in progress].

Provide feedback: Please help us improve the Azure customer communications experience by taking our survey https://aka.ms/F4SN-7VZ

3/28 RCA - Data Lake Storage / Data Lake Analytics

Summary of impact: Between 22:10 on 28 Mar 2019 and 03:23 UTC on 29 Mar 2019, a subset of customers using Data Lake Storage and/or Data Lake Analytics may have experienced impact in three regions:

- East US 2 experienced impact from 23:40 UTC on 28 Mar to 03:23 UTC on 29 Mar 2019.
- West Europe and Japan East experienced impact from 22:10 to 23:50 UTC on 28 Mar 2019.

Impact symptoms would have been the same for all regions:

- Customers using Azure Data Lake Storage may have experienced difficulties accessing Data Lake Storage accounts hosted in the region. In addition, data ingress or egress operations may have timed out or failed.
- Customers using Azure Data Lake Analytics may have seen U-SQL job failures.

Root cause:

Background: ADLS Gen1 uses a microservice to manage the metadata related to placement of data. This is a partitioned microservice where each partition serves a subset of the metadata. Each partition is served by a fault tolerant group of servers. Load across various partitions is managed by an XML config file called partition.config – this a master file which has information about all instances of the microservice; a per region file is generated by a tool. (This tool is applied to all config files, not just partition.config.) Load balancing actions are done in response to the overall load in the region and load on specific partitions. Frequency of load balancing actions is dependent on the overall load in the region. Currently, these load-balancing actions are not automated.

All (code and config) microservice deployments are staged and controlled such that deployment goes to a few machines in a region then to all the machines in a region before moving to the next region. A software component called watchdogs is responsible for testing the service continually and raising errors, which will stop a deployment after the first scale unit or two and revert the bad deployment. The watchdogs can also raise alerts that result in engineers being paged. Moving to next region requires success of deployment in the current region AND approval of the engineer.

What happened: Some of the microservice instances across different regions needed balancing of load to continue to provide best experience and availability. An engineer made changes to the global partition.config file for the identified regions and triggered deployment using the process described above. After observing success in a canary region, the engineer approved deployment in all remaining regions. After deployment completed successfully, the engineer received alerts in two regions: East Japan and West Europe.

Investigation revealed a syntax error in the partition.config. The tool which generates this per region config file, deleted the previous version of the region specific partition.config file and failed to generate a new region specific partition.config file. This did not cause any problem for the metadata service and the deployments succeeded. But later, when for unrelated reasons a new metadata service Front End (FE) process would start, the missing partition.config would cause FE to crash. The deployment in the canary region and other regions succeeded because there were no FE starts so the errors were not seen.

Mitigation: The engineer reverted the bad syntax error in the partition.config file. This new version of partition.config fixed the syntax error, mitigating those two regions as FEs stopped crashing. But this revealed a logic error specific to US East2 region in the partition.config which now caused failures in that region until the engineer fixed that error as well restoring the service availability.

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):

- Mandatory test run automatically at submit time, that sanity-checks partition.config. This test would catch both the syntax error and the logic error.
- Hardening the config deployment mechanism, so that it has built-in delay between regions instead of manual approvals.
- Enhance the watchdogs so that they catch more errors and cause deployments to fail automatically and revert.
- Enhance microservice logic to deal more gracefully with errors in partition.config.
- Fix the tool that generates per region config file for the issue that caused it to delete the output file; instead have it raise an error to fail the deployment.
- Move partition.config to a data folder with separate file for each region, so that an error in one region doesn't affect other regions.

Provide feedback: Please help us improve the Azure customer communications experience by taking our survey https://aka.ms/WFTJ-3XG

3/27 RCA - Service Management Failures - West Europe

Summary of impact: Between approximately 15:20 UTC on 27 Mar 2019 and 17:30 UTC on 28 Mar 2019, a subset of customers may have received failure notifications when performing service management operations such as create, update, deploy, scale, and delete for resources hosted in the West Europe region.

Root cause and mitigation:

Root Cause: Regional Network Manager (RNM) is a core component of the network control plane in Azure. RNM is an infrastructure service that works with another component called the Network Service provider (NRP) to orchestrate the network control plane and drive the networking goal state on host machines. Days leading up to the incident, peak load in RNM's partition manager sub-component had been increasing steadily due to organic growth and load spikes. In anticipation of this, the engineering team had prepared a code improvement to the lock acquisition logic to enhance the efficiency of queue draining and improve performance. On the day of the incident, before the change could be deployed, the load increased sharply, concentrating on a few subscriptions. This pushed RNM to a tipping point. The load caused operations to time out, resulting in failures. Most of the load was concentrated on a few subscriptions, leading to lock contentions where one thread was waiting on the other, causing a slow drain of operations. The gateway component in RNM started to aggressively add the failures back in to the queue as retries, leading to a rapid snowball effect. Higher layers in the stack such as ARM and Compute Resource Provider (CRP) further aggravated load with retries.

Mitigation: To mitigate the situation and restore RNM to its standard operating levels, the reties had to be stopped. A hotfix to stop the gateway component in RNM from adding retry jobs to the queue was successfully applied. In addition, the few subscriptions that were generating peak load were blocked from sending control plane requests to West Europe. Timeout value to obtain locks was extended to help operations succeed. As a result, RNM recovered steadily and the load returned to operating levels. Finally the originally planned code change was rolled out to all replicas of the RNM, bringing RNM back to its standard operating levels and providing it the ability to take higher loads and improving its performance.

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. This includes (but is not limited to):

- Improve lock contention in RNM
- Improve RNM performance and scale out capacity with enough headroom
- Mechanisms to throttle workload before RNM service hits tipping point

Provide feedback: Please help us improve the Azure customer communications experience by taking our survey at https://aka.ms/Q6DN-3BG