Post Incident Review Report
For
Microsoft 365

Report Date: February 4, 2019

Report By: ICC

Microsoft 365 Customer Ready Post Incident Review

Incident Information

Incident ID	EX172491/EX172564
Incident Title	Users unable to access email
Service(s) Impacted	Exchange Online

User Impact

There were two types of user impact for this incident:

1. EX172491 - Users hosted in the wider capacity unit, but NOT served by the affected account forest experienced latency, slow mail flow and in some cases connectivity issues.
2. EX172564 - Users hosted in a specific capacity unit AND served by the affected account forest were unable to connect to the Exchange Online service using any protocol.

Scope of Impact

Approximately 10% of customers served from our European data centers may have seen intermitted mail delays and, on rare occasions, connectivity issues (EX172491). For approximately 1% of customers served from our European data centers, users would have been unable to connect to the Exchange Online service for an extended period of time (EX172564).

The following graph shows relative impact across all users on the affected infrastructure between January 24th and January 25th. Customers affected by EX172564 would have been impacted during the full extent of the dips in availability and were the primary driver in the relative drop in availability for the affected infrastructure.

The graph above shows an aggregation of account forest availability. This does not reflect the true availability of the affected account forest we detailed in EX172564, but a representation of the whole

region's account forests. The first dip above indicates the start of impact and for customers impacted by EX172564 would be when Exchange Online availability was impacted. For those not homed in the affected account forest, availability would have been low during this time, and once access granted, the user would experience delays and latency to most functions over any protocol. The second dip above (25th January), indicates that the start of impact as a tipping point was reached with more details on this below under Root Cause.

We also found that, while the graph above showed us aggregated availability across all account forests, it didn't provide the exact detail we needed to find a solution more effectively. We're making improvements to our telemetry to allow us greater visibility at forest level.

Incident Start Date and Time

Thursday, January 24, 2019, at 9:00 AM UTC

Incident End Date and Time

Friday, January 25, 2019, at 4:00 PM UTC

Root Cause

There were two events that caused impact between January 24th and January 25th:

January 24th

A Windows Server component that handles User Datagram Protocol (UDP) transactions caused a kernel lock to be held for an extended period, creating a kernel error and system crash in some Domain Controllers (DC). This then increased load to the remaining DCs, and caused them to also experience a system crash, reducing the number of DCs available to handle traffic to a level that could not effectively handle the service volume.

January 25th

We experienced a second, previously unknown issue where under high load the DC caches grew larger than physical RAM. This caused a dramatic reduction in efficiency for data that should be sourced from RAM, or from the solid-state drives (SSD) versus the Hard Disk Drives (HDD). As a result, the DCs were not processing requests as expected.

We believe the load which caused the incident is normal load that grew over time. Earlier in the week, on Monday and Tuesday, the data cache management issues, described in the January 25th entry above, were present, but generating no alerts or causing impact to customers. From this data, we can conclude that this growth caused the environment to reach a tipping point.

Since the fixes went in over the weekend of January 26th and January 27th, the account forest has been more stable than before the incident. The data cache management issues are no longer present, and Active Directory (AD) has been performing optimally. Containment procedures have been improved, and now AD is able to handle a much higher update rate.

Actions Taken (All times UTC)

Thursday, January 24, 2019

9:00 AM – The incident began.

9:17 AM – Internal automation triggered an alert for investigation.

9:18 AM – Engineers began investigating this issue as high priority. They began reviewing service logs.
10:13 AM – Engineers identified a potential network issue and began investigating network logs.
12:02 AM – Engineers identified potential issues with some unhealthy Domain Controllers and began investigating them.
12:17 AM – Engineers identified that a portion of the Domain Controllers had crashed due to kernel errors. Engineers proceeded to restart the affected Domain Controllers to clear the kernel errors.
1:29 PM – Engineers began investigating Domain Controller logs to identify the source of the issue.
4:14 PM – Engineers identified higher than expected User Datagram Protocol (UDP) transactions and began investigating ways to alleviate the excessive volume of traffic.
5:54 PM – Engineers reconfigured some UDP related firewall rules on an affected machine and noticed an improvement in availability. They began testing to see if this could be replicated to the remaining impacted Domain Controllers.
6:38 PM – Engineers began implementing the new firewall rules on the remaining Domain Controllers.
7:15 PM – Engineers confirmed that the affected Domain Controllers were returning to a healthy state.
8:18 PM – Engineers performed the rule change on all remaining impacted Domain Controllers and began monitoring the backlog of mail queues to ensure that email was being processed and delivered as expected.
8:39 PM – Engineers began rebalancing load to ensure that mail queues would be processed as efficiently as possible.

Friday, January 25, 2019
12:20 AM – Engineers confirmed that the mail queues were successfully processed and began monitoring the environment to ensure that there were no further issues.
9:23 AM – Availability started to drop again. Engineers started gathering logs from the affected Domain Controllers to understand why they became unresponsive again.
1:57 PM – Engineers concluded that the secondary drop in availability was caused by unexpectedly high password authentication traffic exhausting a cache, sending multiple requests, which, due to the Domain Controllers' bad health, was unable to process smoothly. The cause of this event was different from Thursday's event, which was fixed by implementing the new firewall rules.
6:00 PM – Engineers updated the cache, which eased traffic, returning availability to expected levels. With this information, and information gathered from diagnostic data, engineers began to formulate an action plan.

Saturday, January 26, and Sunday, January 27, 2019
12:00 AM – Over the weekend, outside of core business hours for the affected region, Engineers performed the following optimization, code and configuration fixes to prevent both scenarios from happening again:

- Updated the UDP firewall rule generally, allowing more efficient traffic handling.
- Built and deployed new Domain Controllers, doubling existing capacity in the specific capacity unit most affected.
- Optimized configuration within Active Directory (AD) to improve cache performance.
- Reduced load on password services.
- Upscaled performance counters for related databases for more effective monitoring.
- Implemented a code update for the CAFÉ arrays to capture specific errors. This would modify caching to prevent crashes and impact to other capacity units.

- Modified connection settings, allowing for more efficient load balancing across capacity units.

After completing the work above, Engineers decided to carry out a period of extended monitoring during peak business hours to ensure that these actions had restored service.

Monday, January 28, 2109
8:00 AM - Initial data from telemetry indicated that the service was working as expected. There were no new customer reports of impact.
12:00 PM – After the heaviest part of the European business day, there were still no reports of issues and the service remained stable.
4:00 PM – After a day's monitoring, with telemetry showing everything up and running with no traffic spikes, the incident was declared resolved.

Next Steps are in addition to the actions already taken

Findings	Action	Completion Date
Domain controllers became unresponsive due to a critical capacity failure.	We're conducting a full architectural design review to determine additional scalability and resiliency options.	February 2019
Automated recovery features weren't configured to recover service for this very specific scenario.	We're reviewing our code for improved performance and potential automated recovery options to reduce or avoid similar impact in the future.	February 2019
Containment procedures weren't running as expected, resulting in impact to other capacity units.	We're reviewing our service isolation protocols to ensure that any future incidents are more contained to the infrastructure where they occur.	February 2019
While our monitoring systems alerted engineers of an issue, it took engineers' analysis to understand the nature of the incident.	We're analyzing performance data and trends on the affected systems to help prevent this problem from happening again.	February 2019
Additional monitoring data is required.	We're improving our trace data logging systems to better diagnose the source of faults. We're also improving the reach of our optics to provide greater insight at a more granular level.	February 2019