



Reason For Outage Report (RFO)

Incident: Service instability caused by a high availability (HA) traffic system upgrade.

London, May 21, 2025

This is the Reason For Outage Report (RFO) detailing the service disruption that occurred on May 12, 2025.

This report outlines the incident timeline, the actions taken to restore stability, a comprehensive post-mortem with root-cause analysis, and the measures we are implementing to prevent a recurrence.

Incident Timeline and Actions Taken

- **[2025-05-12 at 15:43 UTC]:** Deployment of an upgrade to the high availability (HA) stack began, one cluster at a time, with no observable degradation or issue during our initial tests.
- **[2025-05-12 at 16:35 UTC]:** Monitoring systems detected noticeable performance degradation and intermittent timeouts, triggering system alerts across multiple clusters.
- **[2025-05-12 at 17:54 UTC]:** We began rolling back the upgrade on the affected clusters to restore stability.
- **[2025-05-12 at 19:19 UTC]:** All services returned to normal, and the incident was officially closed.

Incident Background

On May 12, 2025, a critical incident occurred due to a malfunction in a newly upgraded system responsible for sharing and redirecting internet traffic between servers. This failure led to widespread service disruption across backend servers in multiple clusters.

The incident was first detected at 16:35 UTC through system alerts and was fully resolved by 19:19 UTC. The total duration was approximately 2 hours and 44 minutes, though the impact varied across affected systems and was not uniform in length.

The issue caused performance drops, including timeouts and brief unavailability for some uncached requests. Recovery followed after the new traffic handling system was disabled on affected clusters.

Comprehensive Post-Mortem and Root-Cause Analysis

The incident was caused by an upgrade to the system in the High-Availability stack that manages traffic between servers.

Tests and a limited rollout were completed without issues. The system performed correctly under those conditions.

As the rollout expanded, some systems started experiencing service degradation. The problems appeared gradually and in different areas, which made it harder to trace them back to the upgrade.

The root cause was a scale-sensitive issue that only appeared when the system was fully deployed. Certain traffic patterns and loads that did not occur during testing triggered the failures.

Monitoring was in place throughout the upgrade, but the delayed and scattered nature of the issues made it difficult to immediately identify the cause.

In summary, the upgrade introduced problems that only became visible at scale.

Lessons Learned and Future Measures

As part of our post-incident analysis, we've identified key lessons and actionable improvements to enhance the resilience of our systems. The following measures are being implemented to address the gaps observed and to strengthen our ability to detect, respond to, and prevent similar issues in the future:

Incremental Rollout Strategy: This incident showed that even careful rollouts can miss issues that only appear under full-scale load.

- **Action:** Going forward, we will divide upgrades into smaller phases to reduce the risk of undetected issues.

Monitoring and Correlation Improvements: Our monitoring systems gave early warnings, but the delayed and scattered failures made it hard to link them to the upgrade.

- **Action:** We will improve our investigation processes to more quickly correlate unusual behavior with recent changes, even when symptoms are intermittent or spread across systems.

These changes will strengthen our ability to deliver reliable, high-performance services as we continue to evolve our infrastructure.

Thank you for your continued trust and commitment.

Team Pressidium

For assistance or further information, please open a ticket from the [Pressidium Dashboard](#) or contact us at support@pressidium.com.