VTEX

The Enterprise Digital Commerce Platform

# Incident Report

## Aug 24, 2023: Elevated 5xx Errors Rate in Navigation flow

| | |
|---|---|
| Status Page URL | Period 1: N/A<br>Period 2: https://status.vtex.com/incidents/0lx9b68z8k80<br>Period 3: https://status.vtex.com/incidents/xb000zzh5857 |
| Impacted accounts | Stores using Store Framework (a VTEX IO storefront) |
| Duration | Period 1: 2 hours and 52 minutes (03:04 to 05:46 UTC)<br>Period 2: 5 hours and 58 minutes (05:46 to 11:54 UTC)<br>Period 3: 26 minutes (21:03 to 21:30 UTC) |
| Availability | Stores were partially available, facing intermittent 5xx errors. |

## Summary

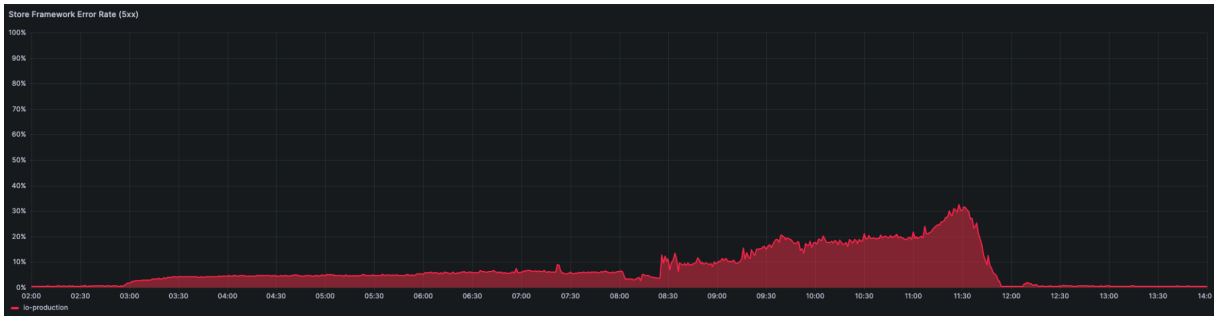
On August 24, 2023, between 03:04 UTC and 11:54 UTC, shoppers experienced two-time windows (see duration above) of increased latency and intermittent errors in the navigation flow in stores using Store Framework. Up to 30% of network requests were affected by these issues, with varying degrees of impact throughout the incident timeline – see **Image 1** for more details.

The issue was triggered by sudden abnormal node[1] interruption[2] rates in our platform, causing cascading failures that led two isolated clusters of several replicated services of our storefront platform to become partially degraded. As an unintended side-effect of one of our mitigation actions, a blue domain parking page was rendered in some navigation sessions.

In addition, on August 24, 2023, between 21:03 and 21:30 UTC, we observed elevated error rates in our platform. This was caused by changes that were performed in the configuration of our cloud infrastructure as follow-up actions of the incident reported earlier.

[1] Nodes are a logical collection of IT resources that runs workloads for one or more containers in a cluster.
[2] Interruptions can be caused by our cloud provider due to capacity, price or other constraints.

**Image 1** – Store Framework Error Rate (5xx)




*Error rates were approx. 7% until 08:00 UTC, but deteriorated with increased platform traffic*

# Symptoms

1. Intermittent errors and slowness in navigation flow

Shoppers navigating in affected stores experienced increased latency and intermittently saw error messages such as "Internal Server Error", "Render server error", "The circuit is now open and is not allowing calls", and "Something exploded". This symptom could be observed in some navigation sessions happening between 03:04 and 11:54 UTC.

2. Blue domain parking page rendered in some navigation sessions

A domain parking page template provided by ParkingCrew, seen on www.exemplo.com, was displayed in some navigation sessions due to a traffic maneuver performed to remove all incoming traffic to unhealthy replicas and reestablish their health. This symptom could be observed in some navigation sessions happening between 08:50 and 12:25 UTC.

# Timeline

## Period 1 (No status, automated actions)

| | |
|---|---|
| **[2023-08-24 03:04 UTC]** | Sudden abnormal node interruption rates from our cloud provider impacted the storefront platform. The issue triggered our automated self-healing mechanisms. |
| **[2023-08-24 03:10 UTC]** | Automated self-healing mechanisms reestablished part of our platform. Two isolated clusters had a partial impact and didn't fully recover, presenting slowly increasing error rates. |

| | |
|---|---|
| | Posterior analysis indicates that our alarms failed to detect issues in the two isolated clusters due to slowly increasing error rates and low traffic from 03:04 UTC to 05:46 UTC. |

## Period 2 (Status 0lx9b68z8k80)

| | |
|---|---|
| **[2023-08-24 05:46 UTC]** | We received the first customer report of intermittent errors in the navigation flow of their stores. Our support teams could not consistently replicate the reported issues. |
| **[2023-08-24 06:53 UTC]** | As new customer reports started coming in, the issue was escalated to the Product Support team. |
| **[2023-08-24 07:37 UTC]** | Our incident response team was alerted and started investigating the issue. |
| **[2023-08-24 08:12 UTC]** | We redistributed part of the traffic to healthy replicas while continuing to repair the degraded replicas. |
| **[2023-08-24 08:50 UTC]** | We started deploying  additional isolated clusters that could receive excess traffic, to fully recover our navigation flow.<br><br>We continued to apply other changes to mitigate the issue, such as manually scaling critical services used by our storefront platform and applying traffic maneuvers to remove all incoming traffic to unhealthy replicas. |
| **[2023-08-24 11:00 UTC]** | We finished deploying isolated clusters that could receive excess traffic, to fully recover our navigation flow. |
| **[2023-08-24 11:30 UTC]** | We found an issue in our traffic maneuvers applied at 08:50 UTC, causing the Blue domain parking page rendered in some navigation sessions symptom. |
| **[2023-08-24 11:54 UTC]** | Platform behavior was reestablished after we finished removing the remaining traffic from the degraded replicas (approx. 10% of requests). Our team continued monitoring. |
| **[2023-08-24 12:25 UTC]** | Side-effects from the 08:50 UTC traffic maneuver ceased. |

| | |
|---|---|
| **[2023-08-24 12:59 UTC]** | We completed a full analysis of the healthy metrics for all replicas to ensure that we were fully recovered. |
| **[2023-08-24 13:35 UTC]** | We declared the incident resolved. |

## Period 3 (Status xb000zzh5857)

| | |
|---|---|
| **[2023-08-24 21:03 UTC]** | Our team applied a change in our cloud infrastructure configuration, as a follow-up action of Period 2.<br><br>This action was applied to reduce the size of the failure domain associated with third-party applications, by isolating them from core VTEX applications. |
| **[2023-08-24 21:09 UTC]** | Our alarms were triggered alerting our incident response team of a rapid increase in error rates in our platform. |
| **[2023-08-24 21:14 UTC]** | Our team identified that one of our clusters had degraded performance and started mitigation actions.<br><br>These actions involve redirecting network traffic from the degraded cluster to healthy clusters. Abrupt changes in traffic volume could negatively impact the healthy clusters and generate a global outage, so this was done carefully. |
| **[2023-08-24 21:30 UTC]** | The mitigation actions were completed and errors ceased. |

# Mitigation strategy

Several response actions were taken to mitigate the impact on our storefront platform and resolve the issue. Here is a summary of the key response actions:

- Manual scaling of critical services used by our storefront platform.
- Removal of all the traffic of unhealthy service replicas to reestablish their health.
- Traffic redistribution to healthy clusters of service replicas.

- Creation of new isolated clusters of replicas to rebalance traffic.

## Follow-up actions: Preventing future failures

We have designed a comprehensive strategy to prevent similar issues in the future and improve our incident response process. All actions described are expected to be completed before Black Friday 2023. The strategy encompasses several key areas:

**Improved Alerting and Detection**

- **[Done]** Briefly after the incident, we adjusted the current alerts by replicas and added an additional alert to detect degradation in low-traffic scenarios.

**Capacity Planning and Service Disruption Budget**

- **[Done]** We reviewed our capacity planning for Store Framework clusters and critical services to reduce the risk of coordinated and sudden abnormal node interruption rates triggered by our cloud provider.
- **[Done]** We reviewed anti-affinity rules among critical services to minimize the impact of eventual interruptions caused by our underlying infrastructure.
- **[In Progress]** We are defining a disruption budget for critical services to maintain node disruption levels that allow our self-healing mechanisms to work as expected.

**Infrastructure Review**

- **[In Progress]** We are reviewing the isolation of critical services of the storefront platform and the control plane responsible for operating the platform. After this review, we expect fewer infrastructure resource interruptions in these components, preventing incidents like this from happening again on the same factors.

**Automation Improvements**

- **[Done]** We created automation to lock minimum resources to critical services of the storefront platform to accelerate our incident response actions.
- **[To Do]** We will continue our already ongoing work on the automation of several manual steps involved in the creation of cluster replicas to accelerate our incident response actions.

**Traffic Management**

- **[In Progress]** We are implementing a gradual rollout based on tenants in the storefront platform to seamlessly redistribute traffic from specific accounts to healthy clusters of replicas. This will allow us to recover specific stores quickly during incidents and modernize our ability to gradually roll out changes in the storefront platform.