Canvas Incident Report

Many users were unable to access or use Canvas for three and a half hours because of a largescale Amazon Web Services networking issue on 28 February, 2017.

Summary

Canvas users on accounts in our US-based Amazon Web Services (AWS) hosting region could not access or use Canvas for three and a half hours between 10:40 AM Mountain Time (12:40 PM Eastern Time / 9:40 AM Pacific Time) and 2:10 PM US Mountain Time (4:10 PM ET / 1:10 PM PT) on 28 February, 2017.

The root cause of this incident was a large-scale AWS event that affected Canvas and many other websites and web applications. We apologize for the impact today’s downtime had on your users. We are working with AWS to understand what happened to their infrastructure today and what we can do to limit our exposure to any future, similar events.

Details

At 10:40 AM US Mountain Time, Canvas users in US-hosted accounts began to experience significant slowness and error messages in Canvas. Automated monitoring alerted our DevOps team immediately, and an influx of Support tickets confirmed that the issue was widespread and impacting users in a meaningful way.

DevOps quickly worked through their protocol for triaging critical system issues. They ruled out everything to do with Canvas itself and noted signs of system stress that pointed to a problem in the AWS hosting environment. They also noticed problems with many non-Canvas websites and web applications that are also hosted on AWS—for instance: Slack (a live chat tool), Box and Citrix ShareFile (file storage / file sharing tools), etc. About this time, AWS confirmed that the problem was on their side with the first of a series of updates on their own status page.

Throughout today’s incident, our engineers worked directly with their peers on the AWS team:

- They brainstormed together about things we could try from our side to restore service or at least reach a state where users could use some Canvas features. None of these ideas worked, in part because more AWS services were impacted as time passed.
AWS kept us informed about their progress toward restoring service. This helped us plan ahead for the influx of load on Canvas as users logged back in and queued requests and jobs began to run.

At about 1:00 PM MT, AWS gave us a 60-90 minute ETA. Canvas began to respond at about 2:00 PM MT, and most users were able to access and use Canvas as usual by 2:10 PM MT (4:10 PM ET / 1:10 PM PT).

Mitigation

We're confident that our partners at AWS will study what happened today, identify the root cause, and take appropriate measures. The AWS infrastructure meets the highest standards of resiliency and redundancy and typically withstands even the most severe stresses. Long incidents like today’s are very rare. We’ve built our six-year 99.9%+ Canvas uptime track record on strong performance by the AWS cloud, and we anticipate a bright future from the partnership for many years to come.

As started our post-mortem process this evening, we thought of two very reasonable questions you may have in mind:

- **Q1**: I understand the root issue today was an AWS problem. But AWS has three hosting regions in the US, and you’ve placed all Canvas accounts in just one of them: US-EAST. Why not use all three regions?
- **A**: At this stage of Canvas growth and development, that is probably wise. We will explore options with AWS. Our working plan is to distribute Canvas database clusters across all three US-based AWS regions. A given region-wide AWS event would then impact only a third or so of Canvas accounts.

- **Q2**: Instructure has a Canvas disaster-recovery plan that includes the option to fail over from Amazon’s US-EAST region to the US-WEST region in some circumstances. Why didn’t you do that during today’s incident?
- **A**: Today’s incident did not meet the criteria we have set for triggering the drastic, crossregional failover measure. AWS downtime events fall into three categories:
  1. Events that affect a single “availability zone” (AZ) within a region. This sort of event is by far the most common, capturing 99%+ of all AWS incidents. Every AWS
region includes multiple AZs. We host Canvas on three separate AZs within the US-EAST region, and we can shift load around amongst the three AZs on the fly. This has protected our users during many AZ-level incidents over the years.

2. Events that affect an entire AWS region for less than 24 hours. These are very rare: less than 1% of all incidents fall into this category. Today’s incident was such an event. In cases like this, our best DR strategy is to “shelter in place” and rely on AWS to recover from the problem as quickly as possible.

3. Events that affect an entire AWS region for more than 24 hours. This sort of event is incredibly rare. In fact, we have never encountered one in our six-plus years in the AWS cloud. But we can imagine how one could happen: for instance, after a very severe natural or man-made disaster. We built our cross-regional failover DR measure to safeguard against this kind of highly-unlikely event.

Conclusion

You rely on us to provide a highly performant learning management system that is available whenever and wherever your teachers and students choose to use it. We feel this responsibility keenly, and we set high standards for ourselves. When we don’t measure up, we learn everything we can and make changes so we can do better in the future. Today’s incident shed light on some changes we can make. It’s a privilege to work with you. Thanks for all you do.