Provisional Slack Root Cause Analysis (RCA) Report

Date: 2021-01-05
What: RCA for *Slack outage*
Date of Incident: 2021-01-04 7:00am PST - 10:40am PST

Issue Summary
Starting around 6:00am PST on 2021-01-04 some customers started experiencing occasional errors and increased latency while using Slack. Around 7:00am PST there was a rapid increase in errors and Slack was not usable for all customers. Around 8:45am PST some customers began to see improvements, but others who were trying to launch their Slack clients were not able to do so. By around 9:15am PST most customers were able to use Slack again normally. We continued to experience elevated errors until 10:40am PST, after which all customers were able to use Slack again normally.

We have been working continuously on ensuring that Slack remains available, fast, and reliable. We are working with our cloud vendor to understand and resolve the underlying issues. We're confident that we have a path forward and that we have proper mitigations in place. Understanding and remediating this issue is our highest priority and we'll send the final report that includes all of our corrective actions to you once it's ready.

Root Cause
Around 6:00am PST we began to experience packet loss between servers caused by a routing problem between network boundaries on the network of our cloud provider. By 6:30am PST the packet loss began to worsen, causing increased error rates from our backend servers. We were paged at 6:46am PST due to the high error rate. In addition, many backend servers were busy servicing high latency requests due to network problems between our backend servers and our database servers.

Subsequently, our load balancer servers marked many backend servers as unhealthy due to the network problems, which significantly reduced the number of healthy servers serving traffic. Around 7:00am PST this meant that there were an insufficient number of backend servers to meet our capacity needs. Customers either could not load their Slack clients, or saw error pages directing them to the Slack status page.

Our backend server fleet then began to automatically scale up to meet traffic demand. Between 7:01 and 7:15am PST our automation attempted to simultaneously add 1,200 servers to the backend fleet, a much higher rate of server provisioning than we normally handle. Our provisioning service, which configures new servers in our cloud environment, was unable to keep up with the multiple tasks of configuring servers, such as setting up DNS configuration, resulting in unhealthy servers in the fleet. Newly provisioned servers were unhealthy: some could not contact the provisioning service or the configuration service, and thus get far enough through the provisioning process to start their services. Others had additional problems we are still investigating. This resulted in partially provisioned servers which could not take traffic. Exacerbated by network instability and the number of servers actively polling it, our provisioning service no longer functioned correctly.

Our observability platform was also not reachable for most of the incident because its connection to its database was subject to the network instability. This complicated our debugging efforts and extended the timeline to recovery. We attempted to reprovision observability platform servers for troubleshooting, but were unable to do so due to the problems with the provision service. We began to do direct queries to our metrics backend.

Around 8:05am PST we diagnosed that our provision service had run out of open file handles, because it kept an open file handle for each server that it attempted to provision and was now partially provisioned in an unhealthy state. We fixed that by increasing the file handle limit at 8:13am PST. We were then able to successfully provision servers which entered service and served traffic, which indicated that our cloud provider's network instability had improved. We later were told by our cloud provider that during this time period they increased their network capacity in an effort to reduce the instability. We were also experiencing a rate limit from our provider that limited how quickly we could launch new servers in their cloud. We worked with our cloud provider and the rate limit was lifted.

We provisioned new backend servers and observed them successfully taking production traffic. We steadily increased our backend server fleet size and began to see successful customer traffic. By 9:15am PST customers were able to use Slack again. We continued to experience elevated errors due to the network instability until 10:40am PST when our cloud provider had finished increasing their capacity. All customers were able to use Slack again normally at this time.

Corrective Action

We are currently completing our detailed investigation, running our post-incident process, and compiling our corrective actions to prevent future incidents of this kind. We will provide a corrective action plan by 2021-01-18.

- We have turned off the down-scaling on our backend server fleet, and it is currently provisioned in excess of our predicted traffic needs.
- Our cloud provider has increased the capacity of their cross-boundary network traffic systems. We are working with them to understand how this happened and have requested a detailed RCA.
- We have a new runbook for how to debug our systems through direct queries to our metrics backend without our observability platform.

- We have prepared methods to configure some services to reduce cross-boundary network transit. If the problem occurs again we can use these methods.