Students experienced issues when a number of them were trying to login to the hub at the same time during start of class.

What Happened

Due to a large number of users starting up at the same time, the concurrent spawn limit of 64 was reached quickly. New nodes had to be brought up by the autoscaler, and since this took rougly 10 mins from start to end, users trying again after 1 minute aren't guaranteed to get things immediately placed.

Resolution

1. Increase the concurrent spawn limit from 64 to 100 https://github.com/2i2c-org/infrastructure/pull/6674
2. Put ucmerced users on larger nodes, so fewer node spinups are needed https://github.com/2i2c-org/infrastructure/pull/6673

Where we got lucky

1. On GCP, we have extensive log persistent capabilities. This allowed us to look back at logs past kubernetes' default retention period, resolving the issue. We lack this on AWS, so we got lucky that this hub was on GCP

What Went Well?

1. Once we could see the 429 in the logs, we could put some mitigations in place easily.

What Didn't Go So Well?

1. We do not have an alert for this, so we had to find out about the issue from users rather than automated alerts
2. JupyterHub's metrics don't seem to expose multiple 429 status codes correctly

Action Items

1. Collect pod logs and control plane logging for AWS too: https://github.com/2i2c-org/infrastructure/issues/6688 https://github.com/2i2c-org/infrastructure/issues/6219
2. Increase the concurrent server limit from 64 https://github.com/2i2c-org/infrastructure/pull/6674 (done)
3. Investigate why 429 status responses weren't showing up in Grafana https://github.com/2i2c-org/infrastructure/issues/6689
4. Reduce the number of new nodes that need to come up to serve ucmerced https://github.com/2i2c-org/infrastructure/pull/6673
5. Investigate an alert for many user server startups being throttled https://github.com/2i2c
issues/6690

Yuvi Panda

IMPACT TIME

Aug 29 at 09:00

DURATION

4d 1h 30m

*All times listed
Pacific Time (US

9:46 AM Due to influx of users, the autoscaler goes from 2 to 7 user nodes. Request for

9:50 AM 63 users are pending their servers starting up. JupyterHub's concurrent pendi
starts responding to users with a '429' status code, asking them to try again ir
since the new nodes are not up yet, trying again after one minute (rougly) gives

9:58 AM 7 user nodes are up, and users are able to login fine when they try to login now

3:33 PM The issue is reported to us via freshdesk: https://2i2c.freshdesk.com/a/tickets

3:52 PM

Triggered by Yuvi Panda through the website.
Description: UCMerced Outage (View Message)

INCIDENT #1317

UCMerced: Too Many Users Starting up at the same time

Sep 2, 2025

9:00 AM

Resolved by Yuvi Panda through the website.

INCIDENT #1317

UCMerced: Too Many Users Starting up at the same time