A view of your load balancer (end of task 2)

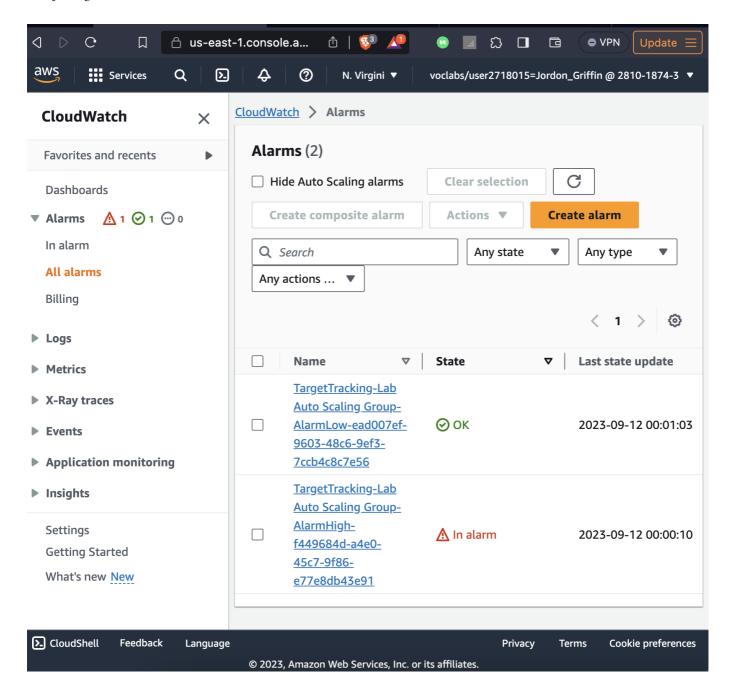# A view of your auto scaling launch template (end of task 3)

Your instances showing that auto scaling is working (end of task 5)

Load balancing is the process of distributing network traffic across multiple servers. This ensures that no single server is overwhelmed with too many requests, which can degrade performance. A load balancer acts as a gatekeeper, directing incoming user requests to the most appropriate server based on various factors, ensuring efficient distribution.

Auto scaling is the process that automatically adjusts the number of computational resources in a server farm, either by increasing or decreasing, based on the actual demand. In simpler terms, it's like having a system that can automatically add or remove servers based on how many users or tasks it's handling at any given moment. It ensures that your cloud resources closely match the demand by monitoring the usage. If the demand increases, like during a sudden increase of website visitors, auto scaling adds more resources to handle the traffic. Conversely, if the demand drops, it removes unnecessary resources. This means you have just the right amount of resources at any time, neither too much (which can be costly) nor too little (which can degrade performance). It's a way to ensure optimal performance while also managing costs.