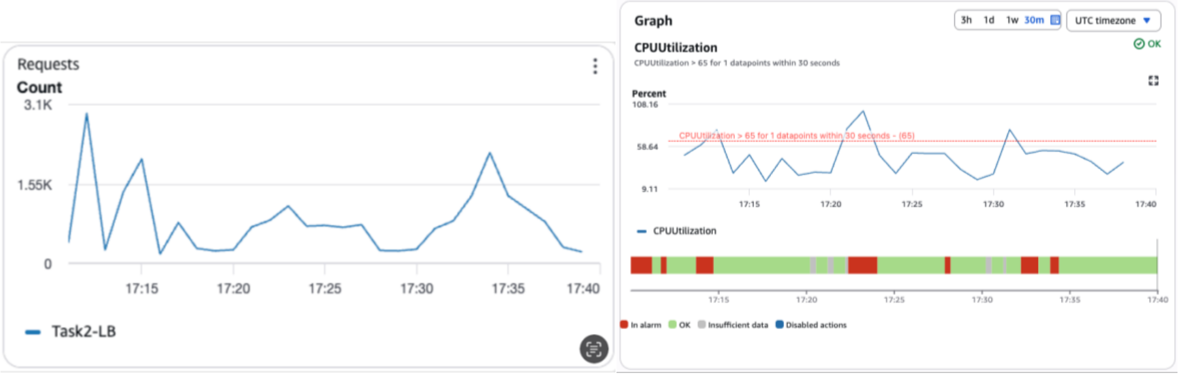
1. What traffic patterns did you see in the load sent to the ELB by the load generator? How did you actually figure out the load pattern? (Please provide appropriate screenshots from the AWS dashboard wherever necessary)

   The traffic sent to the ELB by the load generator shows a highly fluctuating pattern. The load pattern was identified through the requests sent to the load balancer and CPU utilization. The peaks of the "Task2-LB" line in the request count graph represent when the load generator was most active. The CPU utilization also follows a similar trend to the request count. When there are traffic spikes in the request count, CPU utilization also increases, triggering the 65% threshold. When there is less traffic, CPU utilization also decreases. By cross-referencing these two graphs, the traffic patterns can be said to be unpredictable with sudden bursts.



   (The left graph shows the request count, and the right graph shows CPU utilization over time.)

2. Briefly explain the rationale and decision-making process of how you designed and refined the Auto Scaling Group policies. Describe how the insights you gained in the previous question influenced your approach.

   For the CPU lower threshold, I started off with a low value of 30, but from experimentation, I learned that despite high maximum requests per second (RPS), the instance-hours (IH) metric was very high, almost at 400. However, as the CPU lower threshold parameter increases, the average RPS tends to decrease. For the CPU higher threshold, I experimented with a high value of 70, but this led to low average RPS and low max RPS. The higher threshold means instances are less likely to be launched, so this was not a sufficient number to achieve our RPS goals. Then, I experimented with the threshold with lower numbers and learned that lower CPU upper threshold leads to higher IH. As discussed in the previous question, the request pattern is highly fluctuating, and thus, finding the right balance of reactivity was my focus. By adjusting the thresholds along with cool-down periods to maintain the right level of scaling.