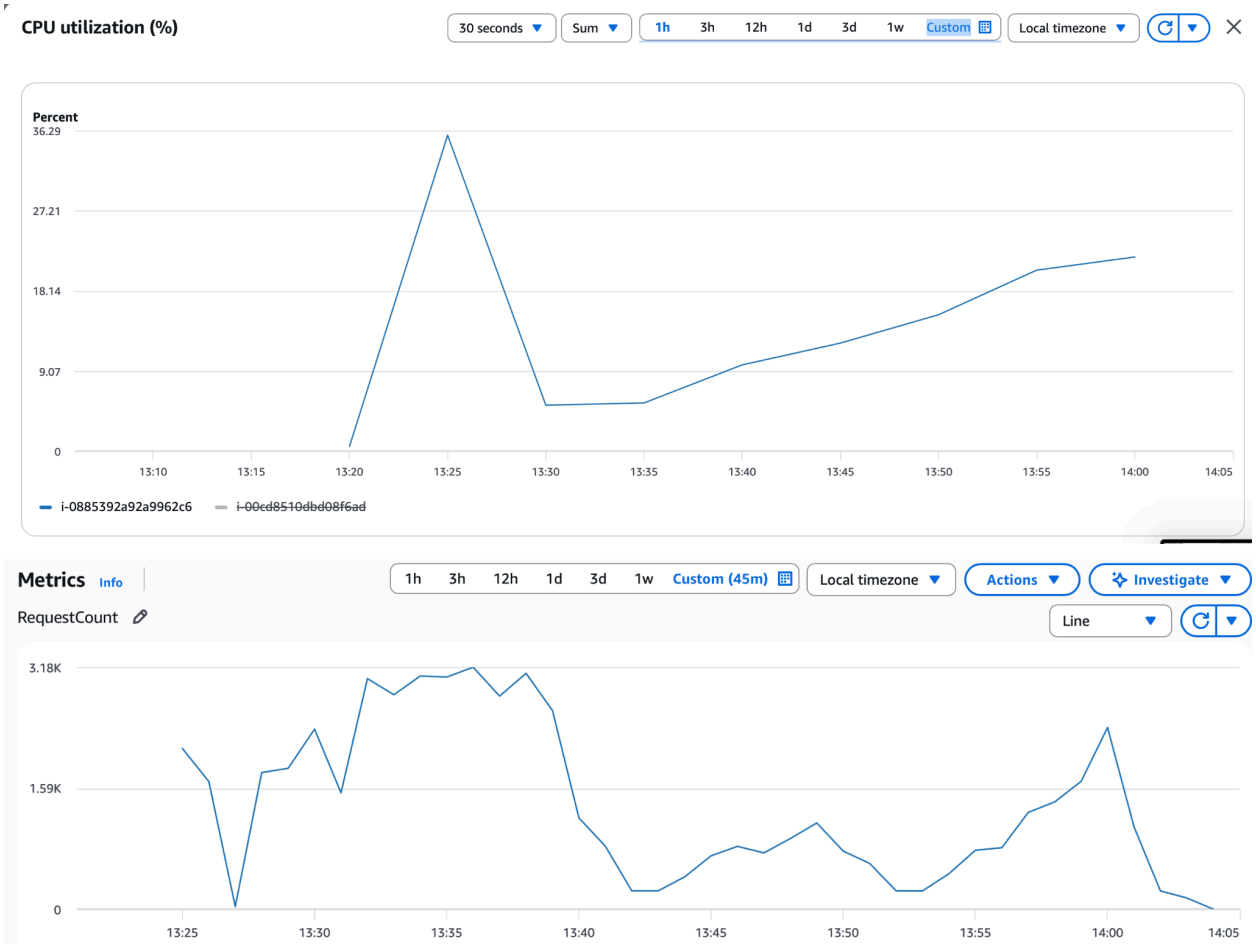

1. What traffic patterns did you see in the load sent to the ELB by the load generator? How did you actually figure out the load pattern? (Please provide appropriate screenshots from the AWS dashboard wherever necessary)

By looking at the two graphs I can determine that the load sent to the ELB appears to be variable in nature. This can be seen in the RequestCount graph, which drops suddenly and peaks suddenly. The evidence can also be seen in the CPUUtilization graph, which also peaked when the requests increased.

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.

The initial configurations that I had started with had a much higher cpu_lower_threshold, around 55-60. However in those configs I was getting both a low max RPS and high instance hours (ih). I experimented with lowering both cpu_lower_threshold and cpu_higher_threshold and while this was reducing the ih it was also reducing the max RPS and average RPS. Eventually I was able to settle at 40 for the cpu_lower_threshold and 60 for the cpu_higher_threshold. Another parameter that had similar issues was cool_down_period_scale_in and cool_down_period_scale_out. If those numbers were too high, it was taking longer to kill and start instances which meant that the max RPS value never reached beyond 35 in the given time frame. After many experimentations I was able to find a value that was able to satisfy all the requirements and successfully complete the experiment.