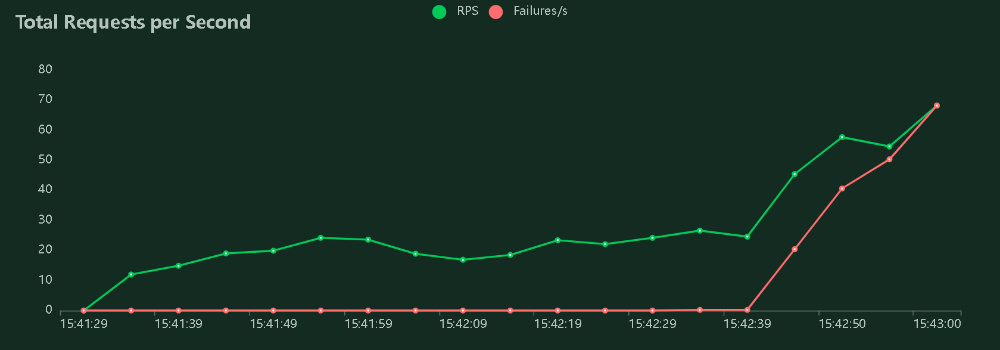
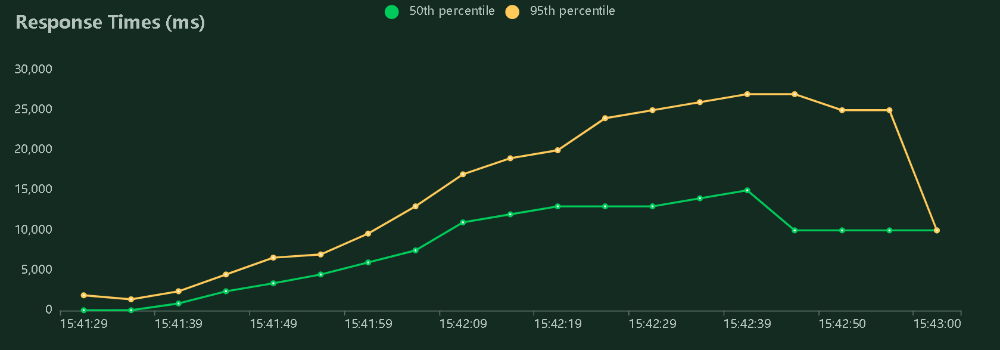
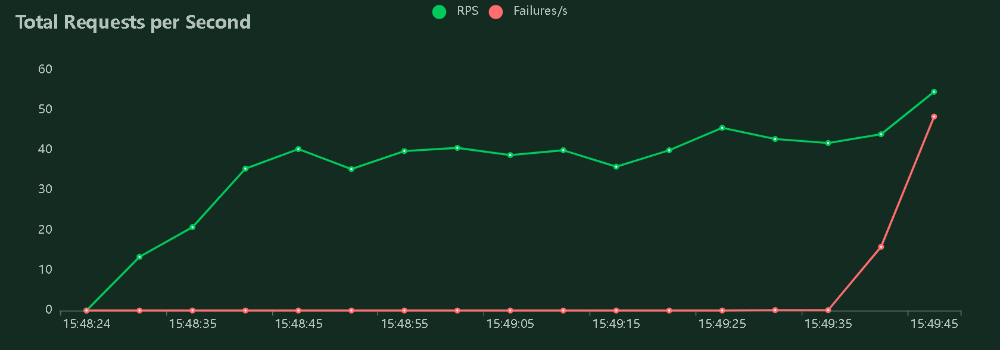
**LOCUST REPORT**

We conducted two tests using an i7 10700 processor and 16GB RAM. In the first test, we used one model's container, and the request per second (RPS) ranged from 12 to 24, depending on the number of users. The maximum number of users we tested was 720, after which the Docker container went down, and all requests failed. The following graph shows the RPS vs. the number of users. As we can see, there is a steady increase in RPS with the number of users.

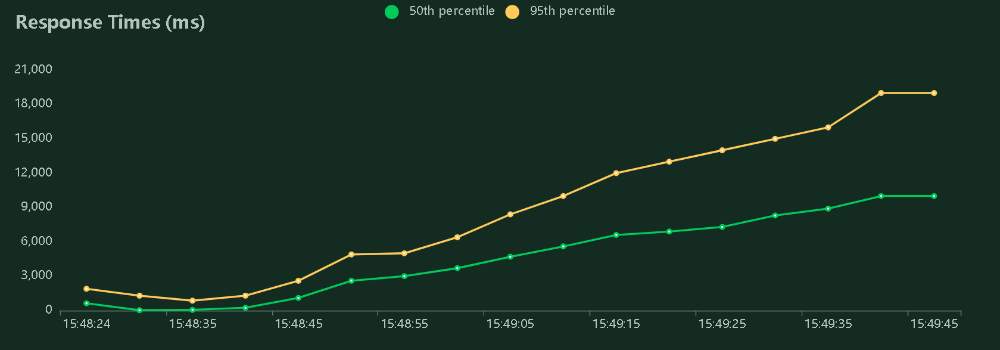
In the next graph, we plotted the response time vs. the number of users. We observed a positive correlation between the number of users and the response time. As the number of users increased, the response time also increased, indicating that the system is taking longer to process each request.



In the second test, we used two model's containers, and the RPS ranged from 13 to 40, depending on the number of users. The maximum number of users we tested was 730, after which the Docker container went down, and all requests failed. The following graph shows the RPS vs. the number of users. We can see that the RPS is higher than in the first test, indicating that the system can handle more requests with two containers.



In the next graph, we plotted the response time vs. the number of users. We can see a positive correlation between the number of users and the response time, similar to the first test. However, we can also see that the response time is more stable and consistent with two containers, indicating better performance.



Based on our results, we can conclude that using more containers of a model generally leads to better performance, as shown in the RPS and response time graphs. However, there is a limit to the number of users that the system can handle, as demonstrated by the maximum number of users tested before the Docker container went down.