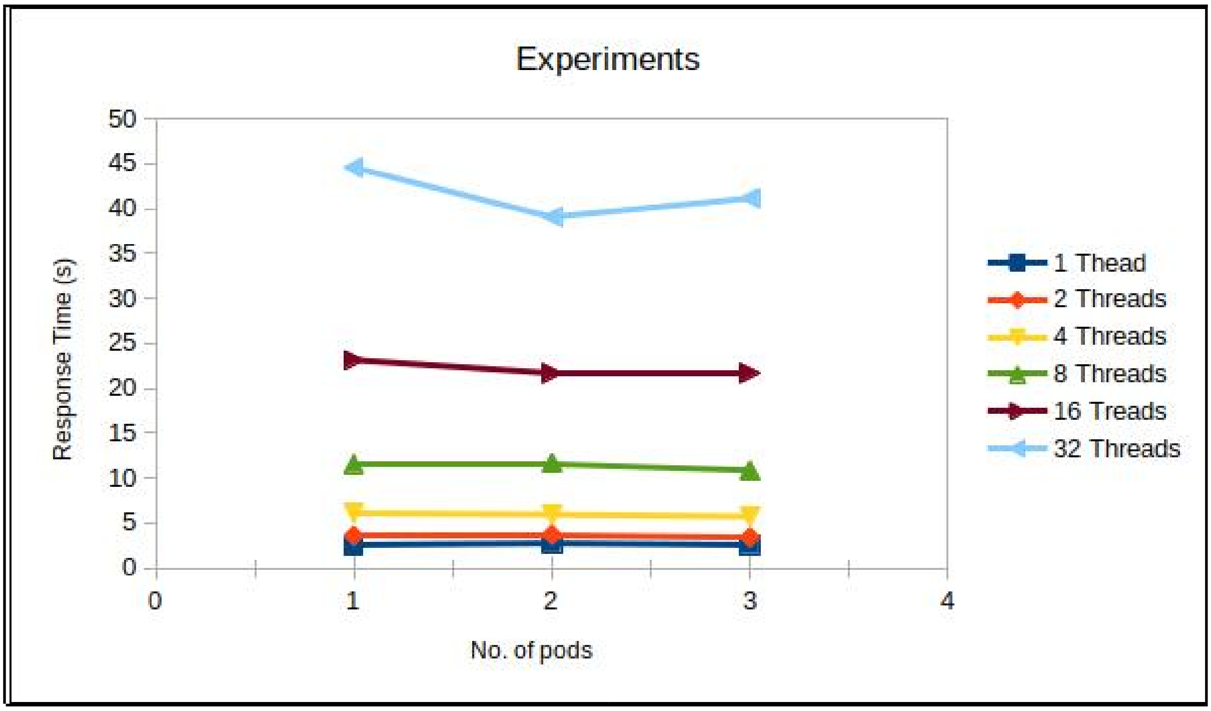
Name: Siddharth Rajesh Student ID: 23852860 Username: sraj50

Report



In this assignment, a RESTful web service was developed and created as a Docker image. A Kubernetes cluster was then created, and the Docker image was used to deploy to the pods created in the cluster. The web service was then exposed as a Kubernetes service upon which a series of experiments was performed. This report describes how the experiments were done and the justification of the test results. The web service can perform multithreading meaning, it can send multiple simultaneous requests to the web service and get the appropriate response back. On the client side we can vary the number of threads to vary the workload by number of simultaneous requests. On the server side, the number of pods is varied i.e. the number of resources in the cluster. The aim of the experiment is to show how the response time changes under varying number of requests (threads) and the resources available (number of pods) in the cluster. The response time is defined as the time taken for the client to send a request to the web service and get back an appropriate response.

The experiments were performed such that first we set up the cluster and specify how many pods will be running. Then the client initiates the responses and varies the number of threads. For example, 1 pod will be set up in the cluster first then the client will send requests starting from 1, 2, 4, 8, 16, 32 threads. The response time of each thread for different workloads is recorded. Each experiment is recorded three times to average out any outliers. This is again repeated for 2 and 3 pods.

As seen in the plot above, for 1, 2, 4 threads, the response time does not change much as the number of pods increases. This is because since the number of threads is relatively low, there are still enough resources for thread execution and threads are not competing for resources hence the response time mostly remains the same. However, this changes as the number of threads increases in 8, 16 and 32 threads. The thread count is high, and more resources need to be allocated and threads are competing for resources. For 32 threads, the difference in response time is more pronounced between 1 pod and 3 pods. In resource constrained environments, the scheduling of thread execution causes the thread execution time to increase as resources are being shared. As more resources are available, threads are scheduled to execute quickly and hence the response time is smaller.

To conclude, the response time is smaller as we increase the amount of resources (pods) for a higher number of threads. For lower thread count, the response time largely remains the same despite increasing the number of resources.