Performance Evaluation Report 2022-2023

Assignment Report

Name: Anemona-Maria Moraru

Group Number: 342 C5

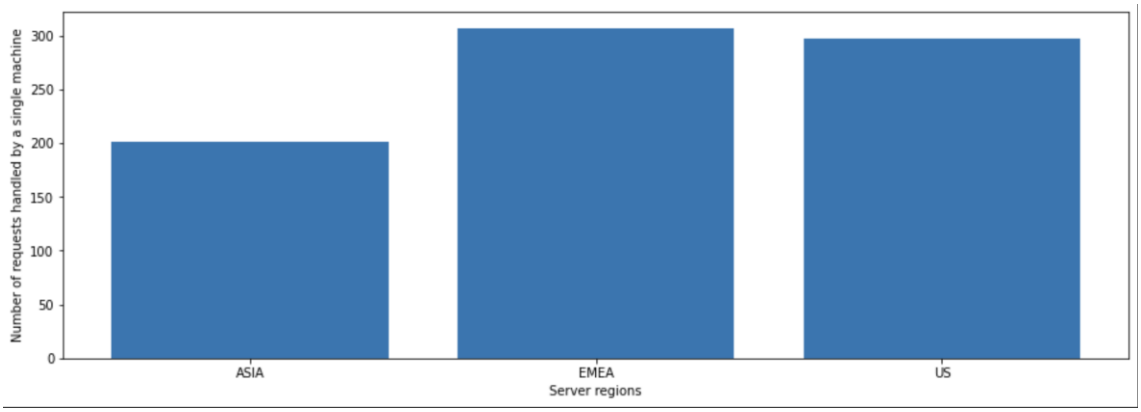
Solved tasks: I, II, III, IV

Considered grade: 85/110

Evaluation - System Limits Analysis

1. How many requests can be handled by a single machine?

When I tested the servers with 500 requests they crashed. The graph below shows the number of requests each server stopped working. The analyse in the graph is made using the api for get requests provided in client.py. After couple of tests made with ping I found out that the servers do not crash with this method even though they were stressed with over 1000 requests each server.



2. What is the latency of each region?

I made several ping requests in each machine and the average latency over each region was:
ASIA - 69.48 ms
EMEA - 75.17 ms
US - 75.25 ms

3. What is the server path with the smallest response time? But the slowest?

Analysing the results of the question above, you would say that the smallest response time is on the path to ASIA region and the slowest on the path for the US, even though there are very small differences between the results. But, analysing the results for each machine, the smallest response time is on the path to the first machine in US region and the slowest is on the path to the second machine in the US region.

4. What is the path that has the greatest loss percentage?

I tested each machine with 100 ping requests and the conclusion is that the path that goes to h5 has the greatest loss percentage (14%).

5. What is the latency introduced by the ***first router*** in our path?

After several tests with 100 ping requests between the computer – router, computer – machines, router – machines the result is that the latency introduced by the first router is around 0.6 ms.

6. Is there any bottleneck in the topology? How would you solve this issue?

Analysing the results from ping requests between all components, I can tell there is a bottleneck in the topology around ASIA region, between r1 – r0 routers. This can be solved using a load balancing policy.

7. What is your estimation regarding the latency introduced?

The estimation regarding the latency introduced is around an average of 44 ms per route to servers.

8. What downsides do you see in the current architecture design?

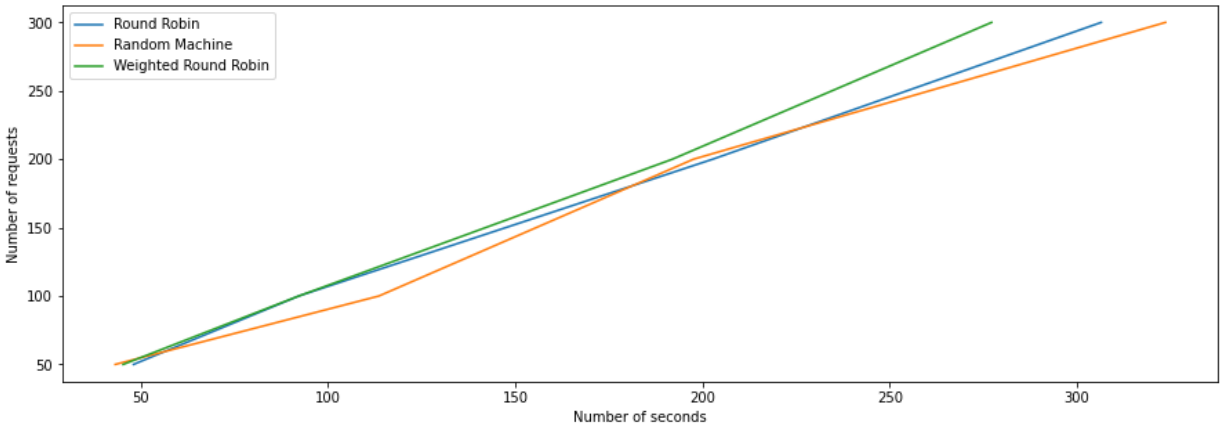
The current architecture has a bottleneck around ASIA region.

Implementation

Round Robin
This implementation distributes the requests evenly between all workers.

Random Machine
This implementation gives random indexes, so we get random servers to make our requests.

Weighted Round Robin
This implementation distributes the request to all regions but with a weight depending on how fast each region is. I put on ASIA 3 requests, on EMEA 4 requests and on US 2 requests.



As a conclusion extracted from the graph above, the load balancing policy Weighted Round Robin does a slightly better job than the other policies. At the beginning Round Robin and Weighted Round Robin seem to be very similar but at 300 requests they are visibly different.