Server Cluster Load Balancing

https://github.com/nikhilsidhu/server-network-simulation

Nikhil Sidhu

December 14, 2022

1 Intro

The motivation for this project comes from the desire to optimize resource utilization in server networks. Although it is possible to make more powerful servers, at the end of the day this becomes expensive and has diminishing returns. Load balancing refers to the distribution of traffic within a network, in an attempt to improve server availability and efficiency of processing.

The set up for this project involves using 5 slow and 3 fast servers handling requests (fig.1), along with two load balancing algorithms that determine where the network traffic flows. The first algorithm is simple, directing requests to the server node with the shortest queue. The more advanced algorithm prefers to send traffic to the faster servers as much as possible to maximize efficiency.

2 Libraries

Ciw - event simulation for open queuing networks, packet-switched networks are of this kind.

numpy - used for calculations

matplotlib - used for creating graphs

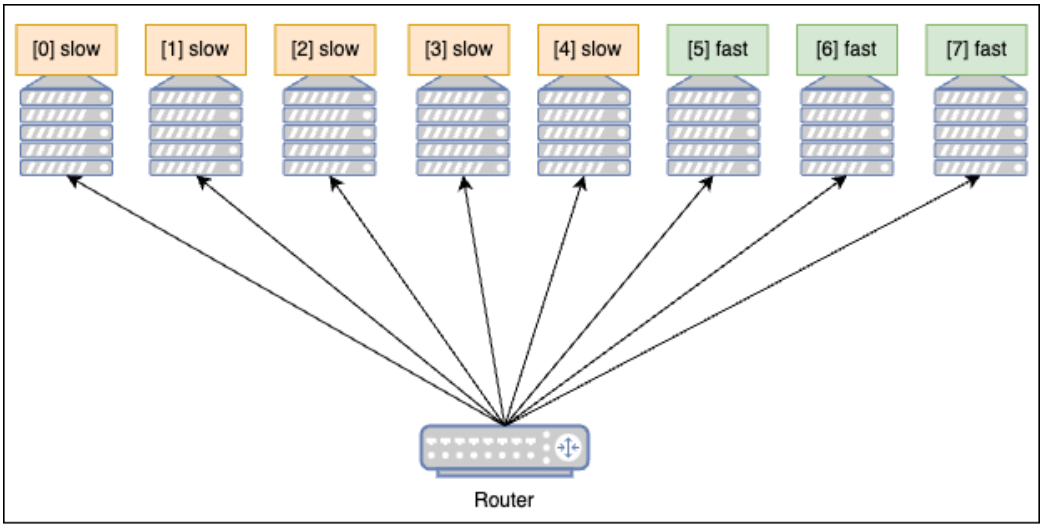

Figure 1: Layout of the server network

3 Instructions

Simply run the simulation with 'python3 Network.py'. Matplotlib will open the graphs for each run at the end of simulation, these can be closed to end the program.

4 Results

The results were as expected, the simple load balancer performed worse than the more advanced algorithm, that preferentially selects faster servers, after convergence (fig. 2).

The server utilization for the simple algorithm is high for all servers (fig.3). When the second algorithm is executed, we see that the slow servers aren't even used in the first block of time that is simulated (fig.4), as they aren't at full load. In the second iteration, the slow servers begin to be used at a slow rate (fig.5). By the last iteration, the algorithm has reached convergence and the numbers stabilize (fig.6).

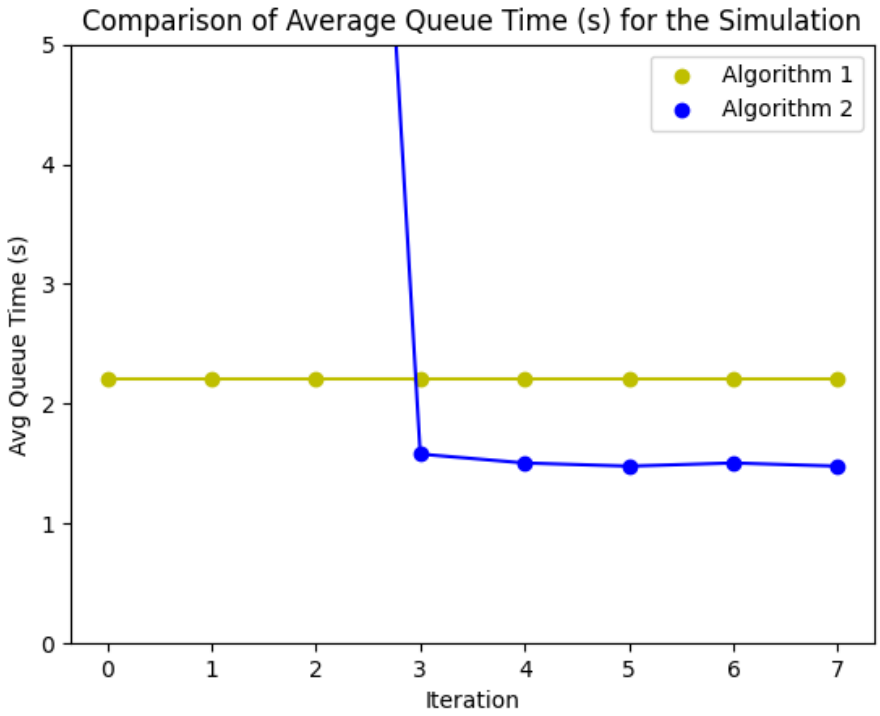

Figure 2: Average queue time in each iteration

5 Discussion

It would be interesting to experiment with larger server clusters and also having servers shut down randomly to simulate failure. There are much more advanced algorithms for load balancing and implementing those in such a setting could be a useful tool when paired with differently structured networks.

6 Papers Referenced

https://sites.pitt.edu/ dtipper/2130/2130_Slides5.pdf

https://ieeexplore-ieee-org.uml.idm.oclc.org/abstract/document/4403185

https://ieeexplore-ieee-org.uml.idm.oclc.org/abstract/document/8368317

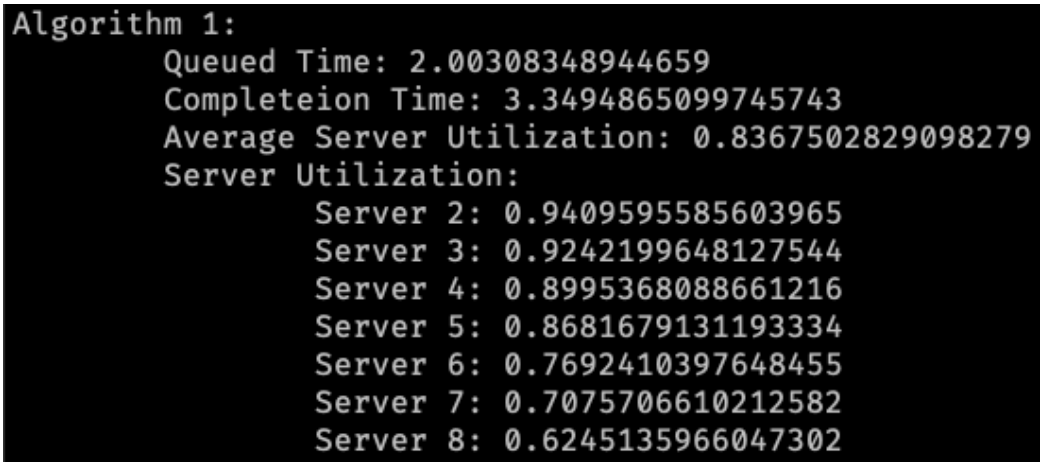

Figure 3: Average server utilization for the simple algorithm

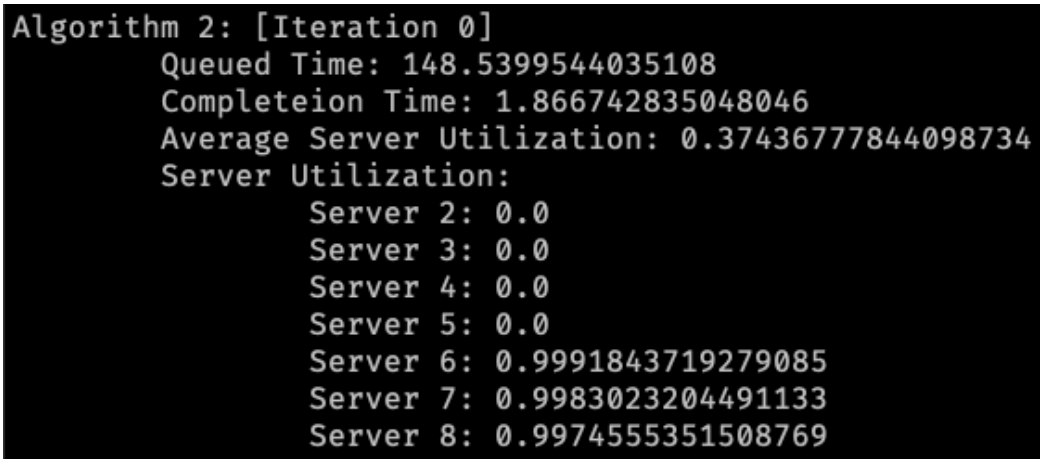

Figure 4: Server utilization in algorithm 2 iteration 0

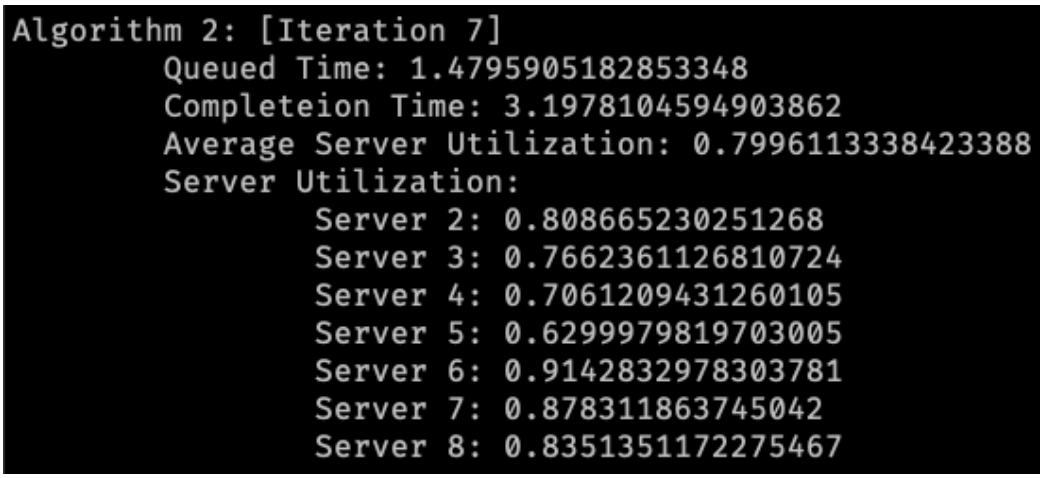

Figure 5: Server utilization in algorithm 2 iteration 7