Assignment 1

Simple Matlab simulation to understand queue behavior

And traffic congestion

V.Pragatheeswaran 150476J

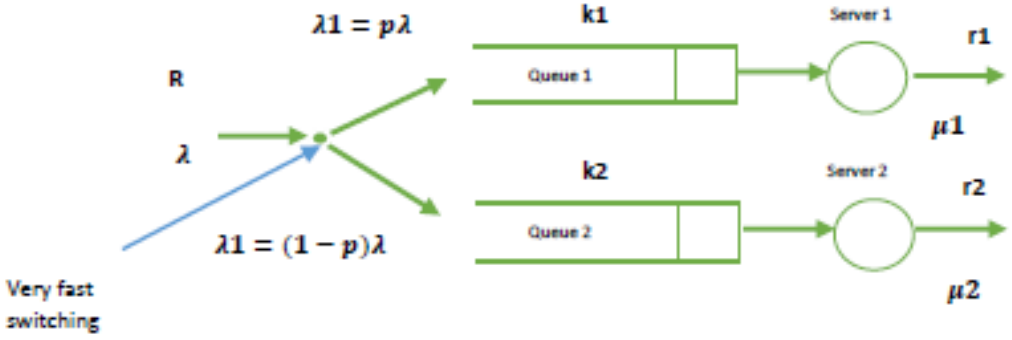
This is submitted as a partial fulfillment for the module
EN3370: Traffic Engineering
Department of Electronic and Telecommunication Engineering
University of Moratuwa

20th of May 2019

Scenario:

- A routing node has a single incoming link of 10 Mbps
- Two outgoing links (4 and 6 Mbps)
- Assume a FIFO simple store

The given scenario can be modelled using two parallel M/M/1/k queues with the addition of blocking as well. In a diagram based representation the following image would be a simplified model.



<u>Theoretical analysis</u>

Let the average packet size is L bits

Let the data rate in arrival link is R bits/second

Hence the packet arrival rate would be = R/L packet/second

Hence arrival rate for queue 1 would be = Rp/L

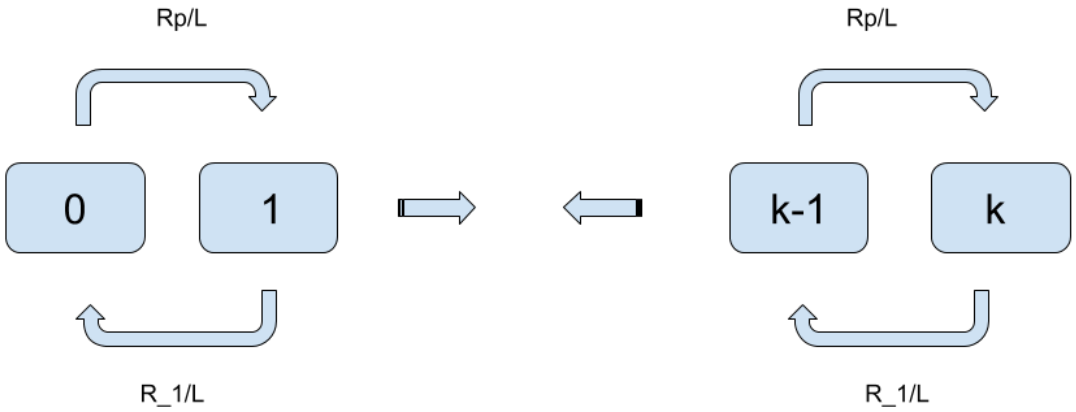
And arrival rate for queue 2 would be = R(1-p)/L ; where p is the probability of a packet reaching queue 1.

Let R_1 and R_2 would be the data rate of departure links from each server respectively, the departure rate or the rate of service would be ,

Service rate for server_1 = R_1/L

Service rate for server_2 = R_2/L

Also since server_1 and server_2 opperations are independent from each other , the derived marko chain for server_1 is as follows ,



Similarly,the marko chain for server_2 would also be similar with their respective arrival and departure rates. Also this is a M/M/1/K1 queue for server 1.

The probability of having m packets which would have to go through server_1 would be,

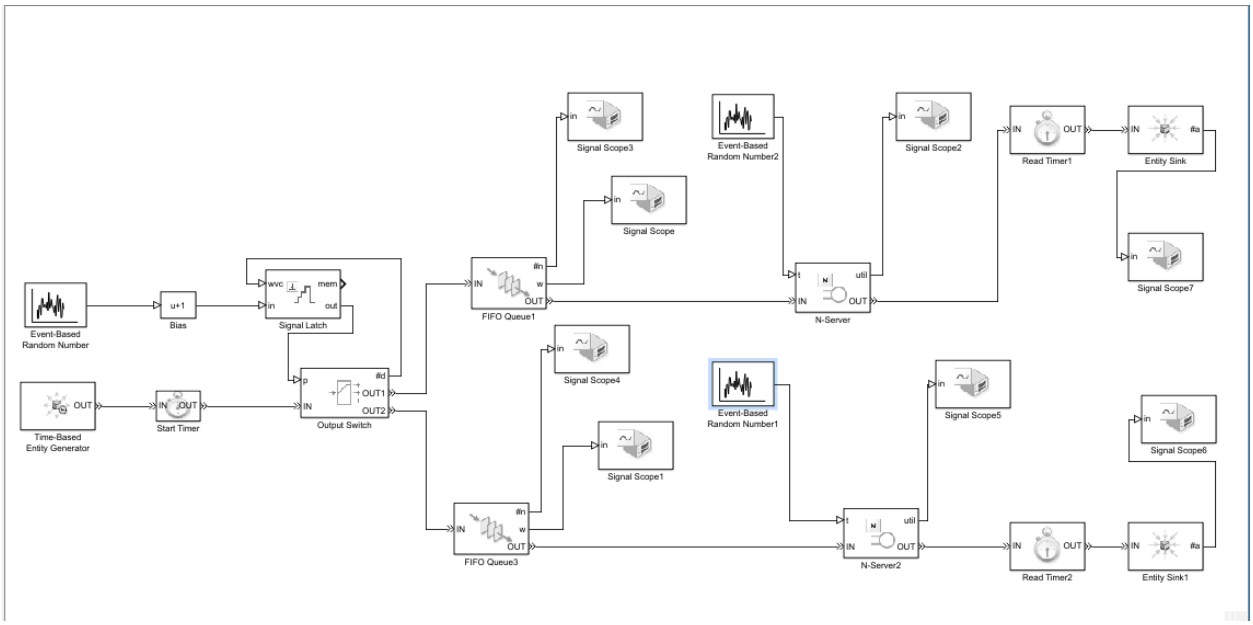
$$P_m = (\rho 1)^m \frac{1-\rho 1}{1-(\rho 1)^{k1+1}}$$

Where þ1 = (Rp/L)/(R_1/L) = Rp/R_1

Similarly for server_2 the R_1 would be replaced with R_2 and number of allowable states might be different.

Simulation and visualization

To visualize and test out the scenario the following matlab model was created in Simulink.



The model basically gives the ability to visualize the given task by controlling the following parameters.

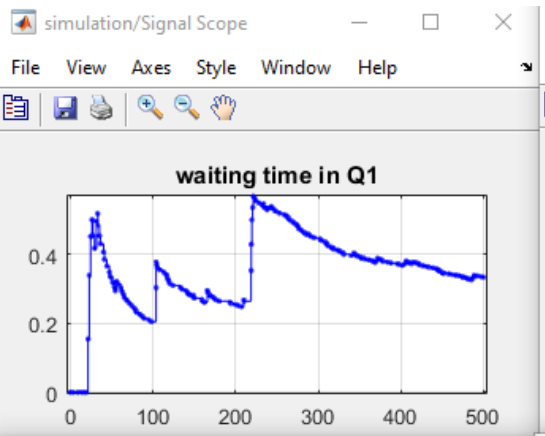
- Queue size
    - Both servers having the same queue size
    - Both the servers having a different queue size
- Service time
    - Both servers having the same mean service time exponential distribution
    - Both servers having different mean service time exponential distribution

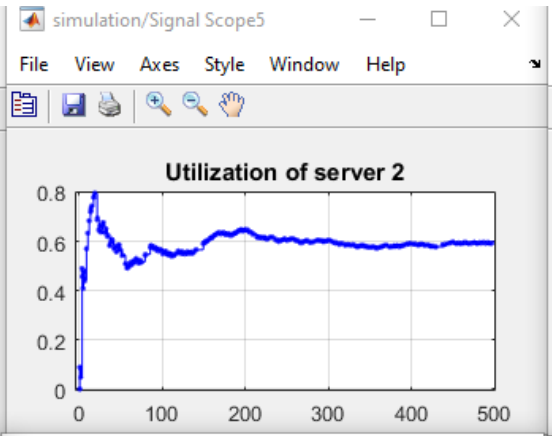
Additionally it is notable that the packet generated interval or the switching probabilities were not changed as it was kept constant to achieve the optimum performance parameters. The entity generated was also kept at a constant with a new entity generated every second. The mean of the time distribution was kept the same as the entity generated rate.

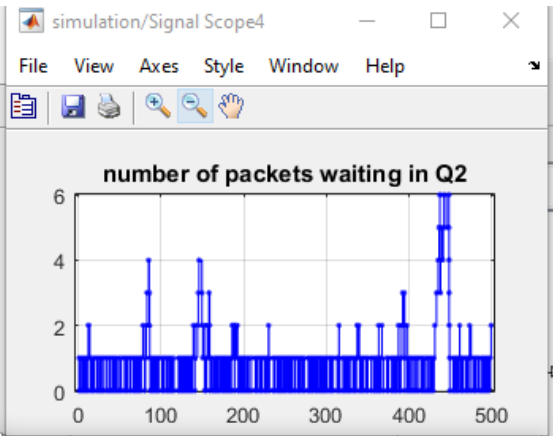
1) **Both servers having the queue size**

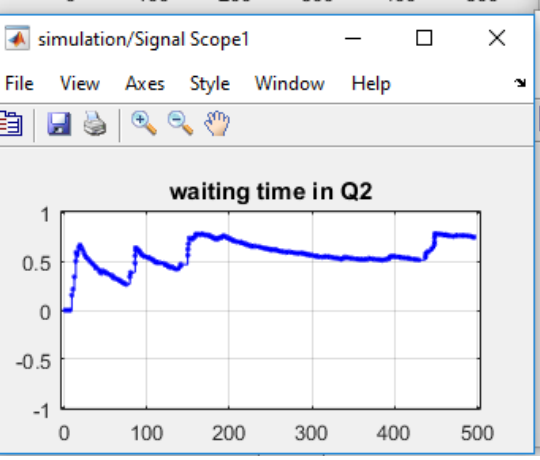
   - **Server queue size 25**

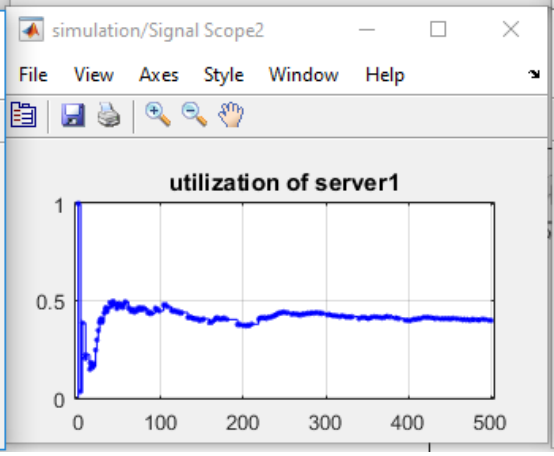
The below shown are the simulation results for a system which has equal queue length, each at 25. The rest of the system was kept as it was mentioned earlier.

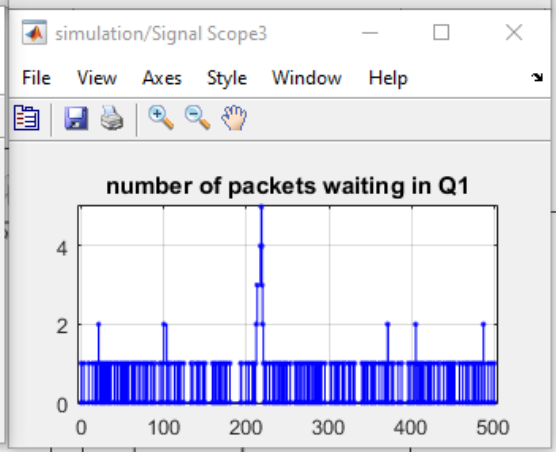












Discussion : -

Looking at the above plots it is very much visible that the server 1 is very much underutilized. Adding on the queues are also pretty much empty all the time with the maximum amount filled at any time being at 20%. The waiting times in both queues are very much low compared to the generated time. From this we can conclude that this system ratios are better for a fast moving system (where service time is very less) , but there is a considerable room for improvement in terms of optimization.

- **Server queue size 10 and 5**

As there was room for development the queue size was reduced to 10 and 5, the entire system was tested.

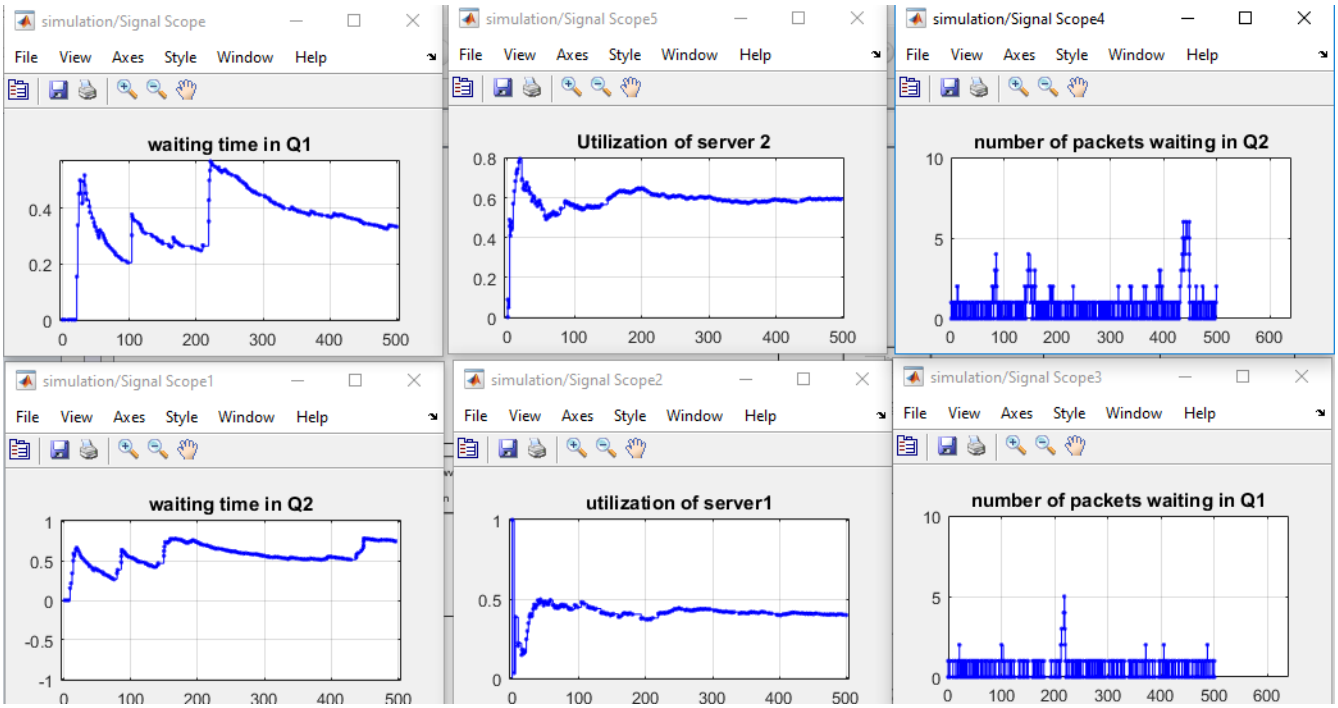
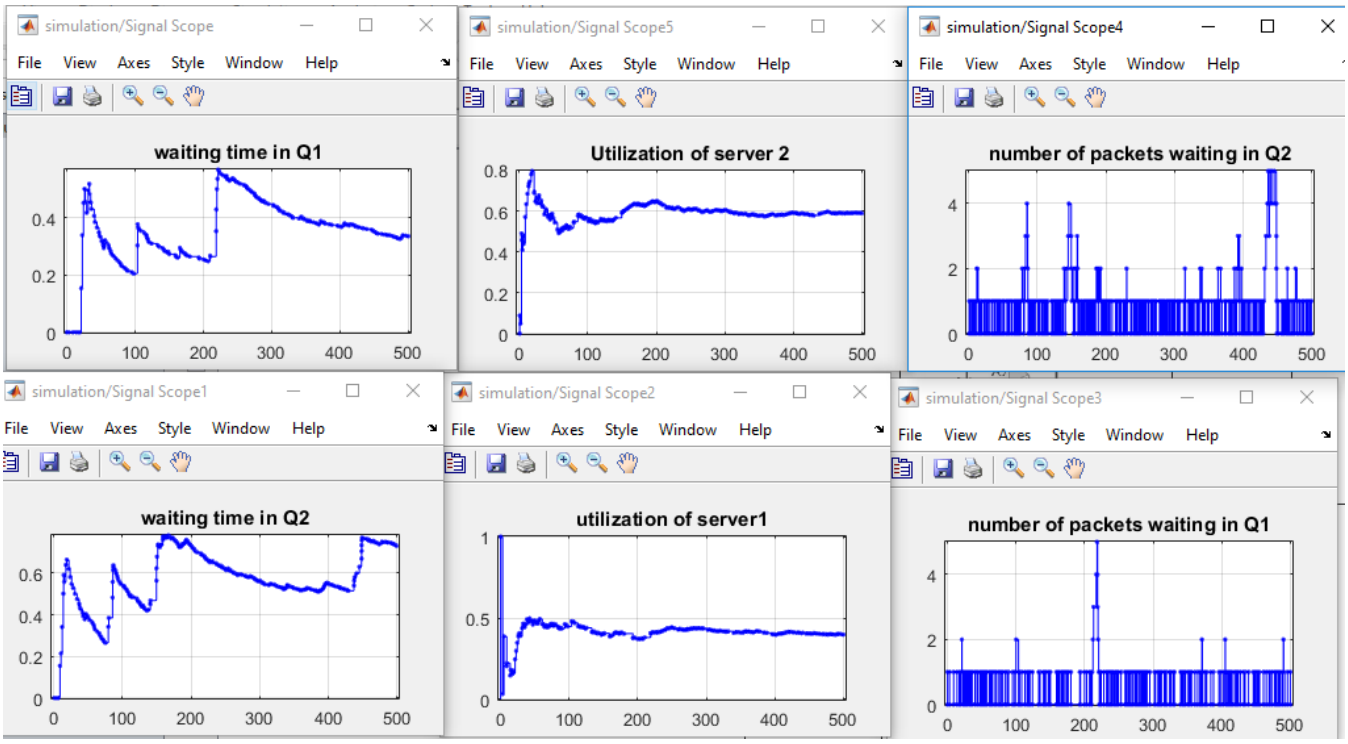


Figure 1.1 : Model results with capacity 10



Discussion :

When compared to the previous results one could simply say that nothing drastically improved in both those capacity levels. Hence it was visible that we cannot actually optimize this procedure by just simply reducing the redundant queue length. As a result of this, as the next step I decided to go with different queue sizes.

2) Both servers having different queue size

○ Queue sizes with 10 and 25

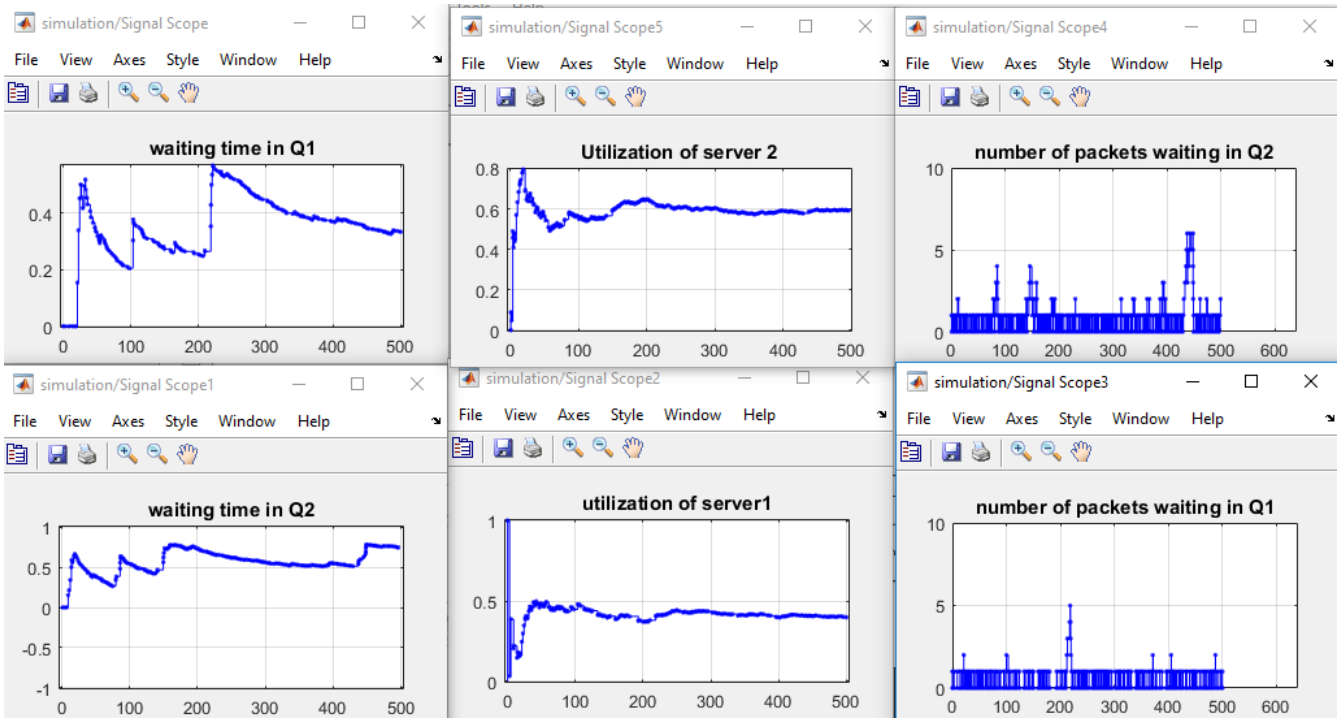


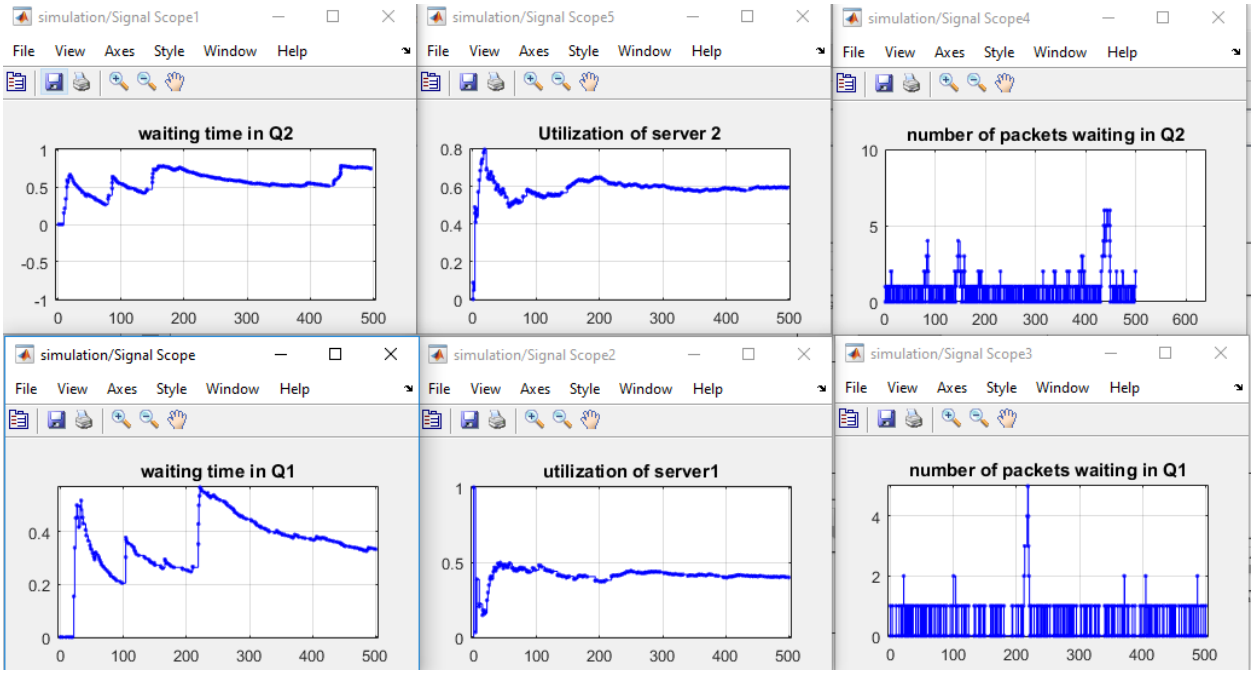
*Figure **Error! No text of specified style in document.**..3: model results with queue size 10 and 25*

With the previously gathered data it was evident that the server 1 needed less queue size while the queue size of the 2. This fluctuation is caused by the difference in data rates. Upon checking with previously found open

*Figure **Error! No text of specified style in document.**1.**Error! No text of specified style in document.**..2 : model results with capacity*

boundary of 10 and 25 was carried out and not much of a significant change in terms of optimization was observed.

- **Server with queue 8 and 12**



Discussion :

The capacitance 8 and 12 were selected with the proportion of the data transfer rates. By reducing the number of empty slots compared to the previous results we also increase the amount of space wastage too. This in real world terms would translate to reduction in cost for storage as well as reduction in complexity.

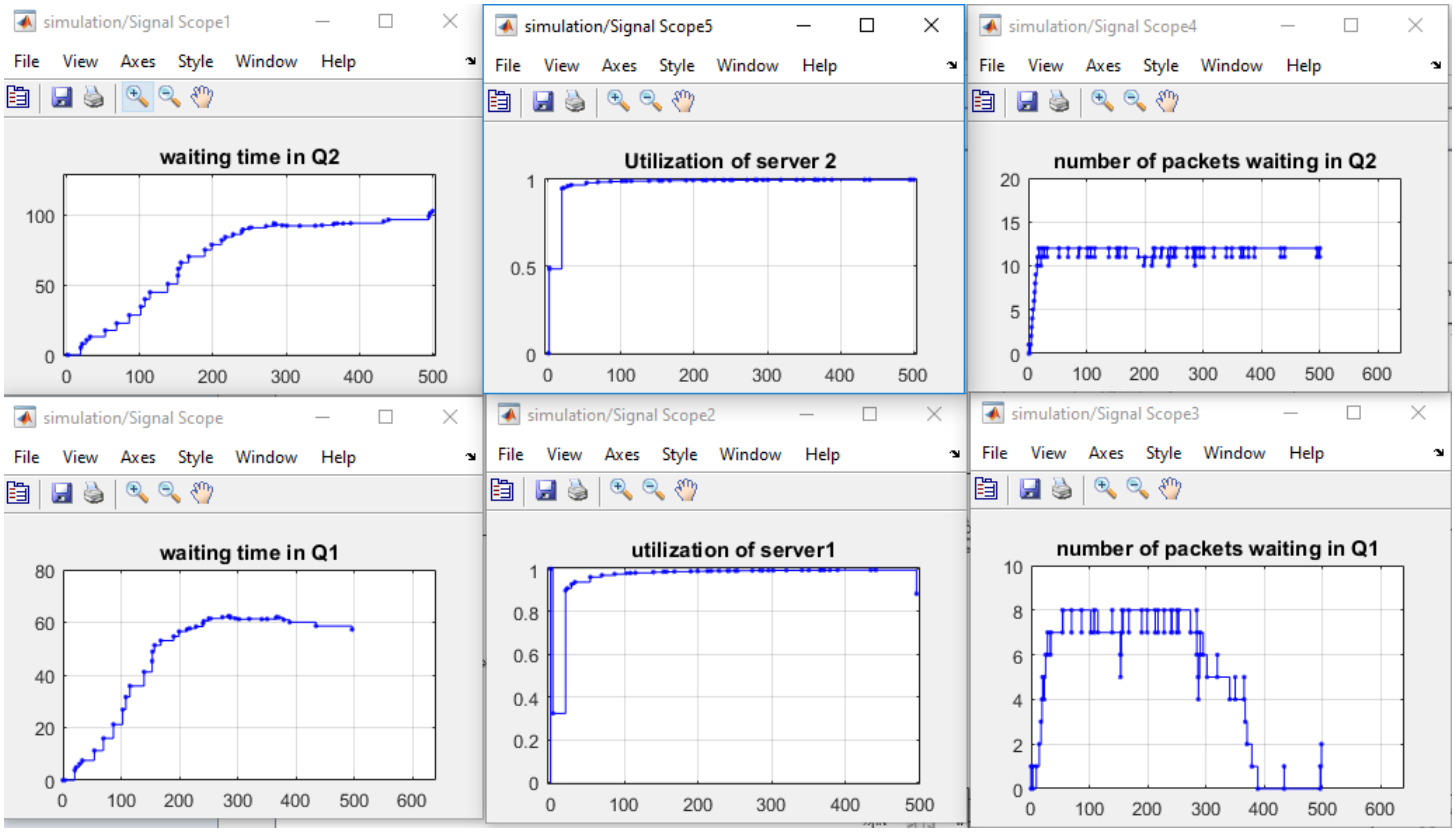
It is visible that utilization of server 1 can be increased further. This can be achieved by increasing the data rate of that path or even using it as fail safe path to users to access when server 2 is having a high work load.

3) Service time – both servers getting packets of the same service time

In this section we will focus on optimizing for the service time changes which the servers may experience. In other words, simulating the entire model for a heavy work requires type of work

- **Service mean time of 10**

Upon tempting to test the model to its limits , the service mean time was set to 10 times the generating time. Which in real worlds could be compared to a task which requires high amount of time.

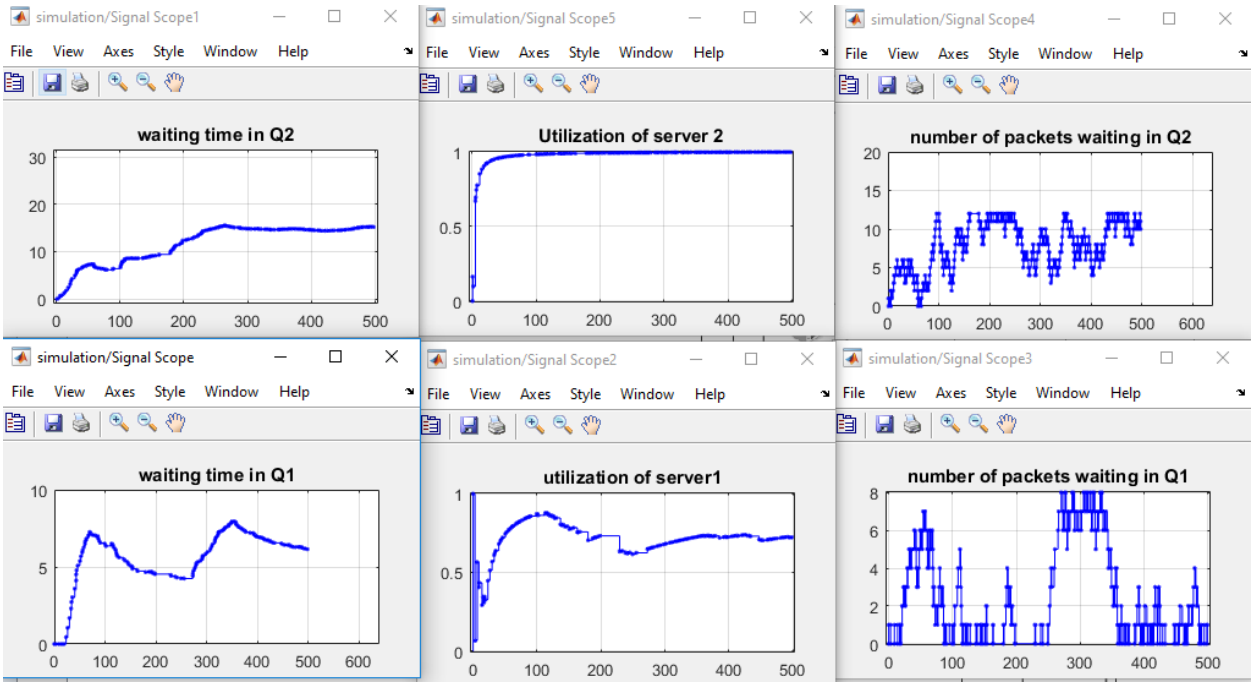


Discussion :

The drastic changes in waiting time is the first thing which raises the question, as of to whether we can expect the system to have a very less waiting time without changing the number of servers. The utilization of both servers have been pushed almost towards its maximum capacity. It is also visible that some of the packets have been dropped because the queue has been full.

- **Service mean time of 2**

Since this is a single server situation it was concluded that ideally it will not be used in heavy work required spaces, where the ideal solution will be a bulk processing . hence the system was evaluated with a service mean time of 2. Still compared to the generating rate it needs two times of rate to sort the requirements.



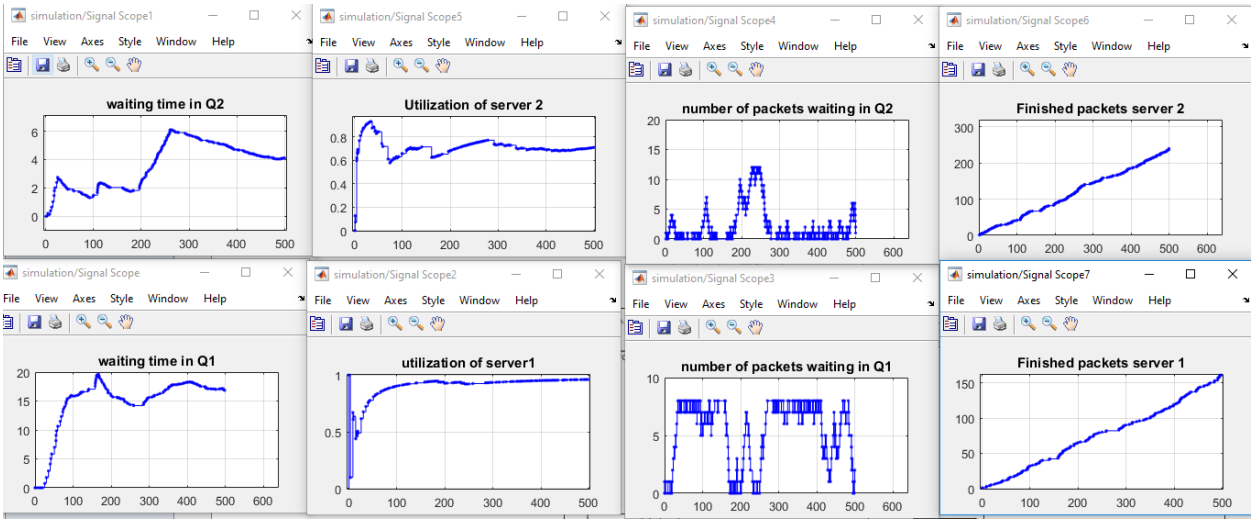
Discussion :

With the results it is visible that both servers are not utilized in a equal manner but server 1 is at its peak usage, which causes two outcomes. First of all waiting time is drastically increased compared to the server 1. The second one is the packets being lost because of the queue being filled. This simply suggests both the servers should be allocated to handle two different work loads to optimize the system. Additionally care needs to be taken towards the number of packets missed over all as well as it directly corresponds to the reliability as wel as the downtime of the system.

4) **Service time – different mean rates**

- **Service rates – 3 and 1.5**

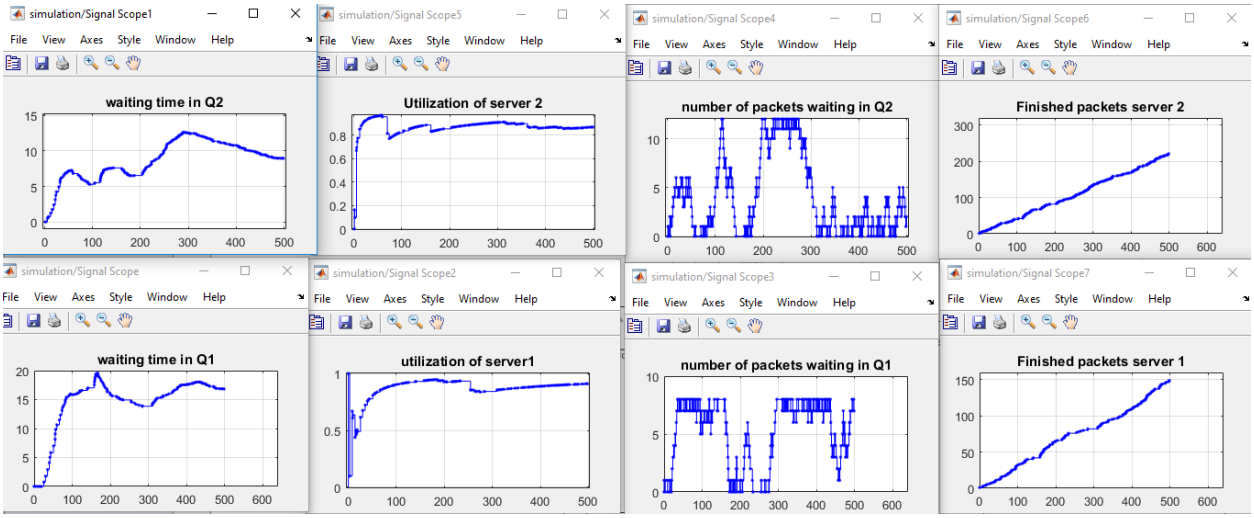
Initially server 1 was allocated packets which needed mean service rates of 3 while server two was given packets which had a service rate of 1.5 These inputs were decided with the previous test results.



Discussion :

It is visible that the server two can be pushed for a bit of utilization. Currently out of the 500 packets which was sent in 220 packets were processed by server 2 and 150 were processed by server 1. Which means 130 packets out of 500 were lost , which is around 26% loss.

- **Service rates of 3 and 2**



Discussion :

With out number of packets lost by a very big margin, we are now able to utilize the model even better by allocating the server 2 with packets which need a mean processing time which is two times the generating rate.

In order to improve the packet loss data, we could simple increase the queue length , with the ratio accordance which was found in section 2.

Summery

Visualizing what this model capable of was one of the key learning outcomes from this assignment, also the path in which optimization process was carried out. The system was tested for both heavy work load packets as well as packets which can be processed at a very small mean time.

From this point the optimization can be carried forward looking at the number of parallel servers in a single FIFO line itself. Increasing the queue length is also a solution which was seen in the last part of this assignment. The time out effect was not accounted for in this assignment , which will also be a factor in a real world environment. The bandwidth capacity of the connecting paths are also a factor in the real world , which has been assumed as 100% allocated bandwidth usage in this assignment.