Serial Queues for Computer Networks and its Equivalent Model

Sriharsha Gangam (h.gangam@iitg.ernet.in)
Roll no. 04010138, B.Tech - VII Semester
MA 402 (Queuing models for performance Analysis) Term Paper
Department Of Mathematics, IIT Guwahati

Abstract – In this report we analyze described the in the paper which proposes a serial queues model to estimate the performance of computer networks with a serial stage. First, based on the relationship between the arrival rate and the service rate, which may be equal or not, we have altogether 13 kinds of working regions. These working regions are obtained from the various conditions on the service rates and Buffer sizes of the individual queues whose equivalent model is to be designed. For each region or case we deduce the system average response time from the equivalent serial queues model according to the arrival rate and the service rate. The average response time has something to do with arrival rates, service rates and buffer size and the relation between these and with the working regions.

Keywords – *System response time, Equivalent Model, Serial Queues.*

INTRODUCTION:
As people use the network service to search for information from Web pages, hence the rapid increase in the number of Web servers. Requesting a particular Web page quantity causes an increment in service requests, and this situation will bring about an increment in the system response time. This is because the loading received by the Web page server increases and causes the servicing time for each HTTP connection to increase. The growth of the servicing time influences the system response time directly. If the number of stages of a Web page servicing in the server cluster is decreased, then the system response time is also decreased. These serial stages of computer networks cause the system response time mainly as result of network latency, transmission time,

DNS lookup and queuing time as well as by service time. But the network transmission time is usually very hard to predict as the transmission time of a network includes a router, switches and network circuits and thus every kind of process and facility .

The Serial Queues Representing the Subtree of Computer Networks :

In a real Internet system, if all personal computers (PCs) want to make use of a network connection, several devices are needed, for example a network card, a modem, a repeater, a router and a switch. Then the PC can deliver the message to the Web server through several stages of a serial connection to the Internet. We assume that messages from the Internet enter into multi-stage switches to reach the Web server. In the sub-tree of the network the

number of switches and serial connections is "*n*", and the service rate can be represented as u1, u2, …. In each of the switches.

We have two serial queues as shown in the figure in the paper and the related parameters are given in the table.

The parameters for analysis are mentioned in table : 1 of the paper. We define response times of a server in three different scenarios.
1)Service rate is less than arrival rate.
2) Service rate is equal to the arrival rate
3) Service rate is more than the arrival rate.

Without considering the stability of the queue length, these parameters are defined in section 3 of the paper. And depending on various values of K1, K2, u1,u2 etc and for each of these cases,
An equivalent queuing model is proposed.

In Case 1:Since K1 < 1, we have E1(t) = B1/u1.
This is because the arrival rate is faster and the queue is never empty and the output is at a constant rate u1 and the output is given at a rate in which the inter arrival times is exponential u1, this is independent of the arrival rate and thus the second queue behaves with K1 >1 and thus the response time is as given in equation 2.
Because the bottleneck is at the first queue, the system response time of the equivalent serial queue is as shown as equation 3.

Thus we can write the total response time as the sum of response times at individual queues and thus we get equation 5.

Equations 6 and 7 are obtained by simply substituting the previous equations in their respective places.

In Case2: We have for queue 1, the arrival and service rates as equal. This can be seen as equivalently as a bottle neck with two services one after the other and this is equivalent to 2u1 service rate with bottle neck and so we use a different expression of the response time as equation 8.

We use the same expression as the total response time being equal to the sum of individual response times and obtain the value of the equivalent service rate. Again because there is a bottleneck in the first queue we assume a bottleneck in the equivalent queue and this leads to equation 10. We finally substitute the values to get the equivalent response times as equation 14.

The author in the paper then proceeds in a similar fashion substituting appropriate response time equations depending on the different conditions the of parameters of individual queues.

The key point to be noted is the different response times given with various variations of the arrival and service rates in the individual queues as given in Section 3 of the paper.

Conclusion:

The simulation has been performed on the QNAT tool. The simulation results show that the single equivalent queue provides a system response time which is accurate enough to simplify the computation of the average response time of the serial stages. The errors are small, below 0.11%, and we can use easy

equations to compute the system response time for the two serial queues. By using these equivalent methods we simplify the complicated queuing network into a couple of major queues; this is an efficient way to estimate the overall performance with respect to the parameters of the queuing network.