Sant'Anna
Scuola Universitaria Superiore Pisa

NETWORK SIMULATION

2019/20

YAMAN PARASHER

PHOTONICS INTEGRATED CIRCUITS, SENSORS & NETOWORKS
(PIXNET 2019-21)

Assignment No.1
Queueing Systems

12 Nov 2019

1 Objective of the Assignment

Given the M/M/1 queue , Service rate (S)= 1 Mb/s , Packet size = 1500 bytes (average). Plot the average delay spent by a packet in the queue as a function of the arrival rate (A).

2 Introduction

Before going directly to the objective of the assignment, I would like to show the process of how I develop the understanding of the whole concept. Starting from developing understanding about the queuing theory , I developed intuitive explanantion of some of the prime components (like the basics about the queuing systems & classification) that paved my way to understand the assignment in a much better way.

A Queuing theory is simply defined as the mathematical study of queuing, or waiting in lines. A general Queue can contain anything like such as people, objects, or information. A basic queuing system consists of an arrival process (how customers arrive at the queue, how many customers are present in total), the queue itself, the service process for attending to those customers, and departures from the system. Such queuing models are often used in software and business to determine the best way of using limited resources.

For the given objective, I start by developing my understanding about the queueing system in detail through Little's Theorem. According to the Little's theorem : The average number of customers (N) in a queuing system are determined from the following equation:

$$N = \frac{\lambda}{T} \tag{1}$$

Here lambda λ is the average customer arrival rate and 'T' is the average service time for a customer.

Hereafter focusing on an intuitive understanding of the fundamental concept behind the Littles theorem I dig deeper into characteristics of a queueing system that impact its performance. I found that the most important characteristics of a queueing system basically includes understanding of three things in general, which are

1. Arrival Process
   The probability density distribution that determines the customer arrivals in the system. In a messaging system, this refers to the message arrival probability distribution.

2. Service Process
   The probability density distribution that determines the customer service times in the system. In a messaging system, this refers to the message transmission time distribution. Since message transmission is directly proportional to the length of the message, this parameter indirectly refers to the message length distribution.

3. Number of Servers
   Number of servers available to service the customers. In a messaging system, this refers to the number of links between the source and destination nodes.

After this, I come to know that a typical queueing systems can be classified by the following convention: A/S/n
,where A is the arrival process, S is the service process and n is the number of servers.

A and S are can be any of the following:

- M (Markov)-Exponential probability density
- D (Deterministic)-All customers have the same value
- G (General)-Any arbitrary probability distribution

In this assignment the objective mainly revolves around the M/M/1 queueing system. From various texts, I found that here the arrival and service time are negative exponentially distributed (Poisson process) where the whole system consists of only one server. I believe that this queueing system can be applied to a wide variety of problems as any system with a very large number of independent customers are usually approximated as a Poisson process.

To understand this whole M/M/1 concept, I studied individually the basic three components that classify a queuing system. I mean the arrival and service process that is assumed to be Markov here and the number of servers which is taken to be 1(or unity) for this kind of system.

Therefore, to get to know about the negative exponential nature of the Poisson process for the arrival service time, I have taken into consideration the probability density distribution for a Poisson process, whcih can be expressed mathematically as,

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \tag{2}$$

In the conetxt of the queuing theory the equation 2 describes the probability of seeing n arrivals in a period from 0 to t. Where: t can be defined as the time interval 0 to t n is the total number of arrivals in the interval 0 to t. λ is the total average arrival rate in arrivals/sec.

To get a thorough understanding of this Negative Exponential arrival, I consider a special case of this distribution, which is "the probability of no arrivals taking place over a given interval by substituting n with 0".

$$P_0(t) = e^{-\lambda t} \tag{3}$$

This equation shows that probability that no arrival takes place during an interval from 0 to t, is negative exponentially related to the length of the interval which can be understood by example of cars on a highway provided The number of cars in the system or the highway is very large.Impact of a single car on the performance of the another is very small, i.e. a single customer consumes a very small percentage of the system resources.Movement of all car are independent, i.e. their decision to use the system are independent of other users.

Similarly the intuitive understanding about the service time was developed through the example of telephone call duration case example. As far as the number of servers is concerned, the suitability of M/M/1 queueing is easy to identify from the server standpoint. For example, a single transmit queue feeding a single link qualifies as a single server and can be modeled as an M/M/1 queueing system.

While digging a bit dipper into this whole queuing system, I came to know about the generic equations that describe an M/M/1 queueing system which are pretty much straight forward and easy to use. The very first equation is

$$\rho = \frac{\lambda}{\mu} \tag{4}$$

,which define the ρ, the traffic intensity (sometimes called occupancy) of the system. It is defined as the average arrival rate (λ) divided by the average service rate (μ).

For a stable system the average service rate should always be higher than the average arrival rate. (Otherwise the queues would rapidly race towards infinity). Thus ρ should always be less than one. Also since the point of consideration is the average rates here, there can be cases where the instantaneous arrival rate may exceed the service rate.

However over a longer time period, the service rate should always exceed arrival rate. Apart from this, the Mean number of customers in the system (N) can be found using the following equation:

$$N = \frac{\rho}{1-\rho} \tag{5}$$

It can be seen from the above equation that as ρ approaches 1, the number of customers would become very large.

This can be easily justified intuitively by understanding that ρ will only approach 1 when the average arrival rate starts approaching the average service rate. In this situation, the server would always be busy hence leading to a queue build up (large N).

Arrival Rate ($Mbits/s$)	Service Time (s)	Average Delay in Queue (s)	Average Delay in System (s)
0.1	0.012	0.001	0.013
0.15	0.012	0.002	0.014
0.2	0.012	0.003	0.015
0.25	0.012	0.004	0.016
0.3	0.012	0.005	0.017
0.35	0.012	0.007	0.019
0.4	0.012	0.008	0.020
0.45	0.012	0.010	0.022
0.5	0.012	0.012	0.024
0.55	0.012	0.015	0.027
0.6	0.012	0.018	0.030
0.65	0.012	0.023	0.035
0.7	0.012	0.029	0.041
0.75	0.012	0.037	0.049
0.8	0.012	0.049	0.061

Tabela 1: Table depicting different set of values of Average delay in Queue and system for a given arrival rate (A) range

Lastly I got to know about the total waiting time (including the service time) by

$$T = \frac{1}{\mu - \lambda} \tag{6}$$

Again I observed that as mean arrival rate λ approaches mean service rate μ, the waiting time becomes very large.

An important lesson that I learn here is that systems should always be designed in a manner that even at peak throughput of the system, resource occupancy should always be a little below 100%. This is required to keep the queue lengths and delays within bounds.

3 Simulation Setup

According to the given objective, our case of analysis is a single server queue with exponentially distributed inter arrival times and service times called M/M/1 queue (also described as Markovian/Markovian/1 server queue). With:

- Service rate (S) = 1 Mbits/s
- Packet size = 1500 bytes

Where we are supposed to: Plot the average delay spent by a packet in the queue & system as a function of the arrival rate (A).

Given the average packet size in bytes, first step is to convert it to bits. Knowing that 1 byte is equal to 8 bits, our average packet size in bits is 12,000 bits. Now with the service rate (S) given and packet size, we can determine the service time:

$$Service\ time = \frac{Packet\ Size\ (bits)}{Service\ Rate(bits/s)} = \frac{12,000\ bits}{1,000,000\ bits/s} = 0.012\ s \tag{7}$$

For the data collection, I got 2 inputs: service time(which is found to be 0.012 s from the given data) and number of packets. Here, I only input different arrival rate values to see the system's behavior.

As far as the number of packets is concerned ,I use 10,000 packets and vary the arrival rate of the packets for a certain range, reporting the outputs in a table depicted through Tabela 1.

The arrival rate (A):

$$A = \frac{Packet\ size(bits)}{interarrival\ time(s)} \tag{8}$$

And to find the delay in the system, I just add the service time to the delay in the queue I obtained from our simulation for each subsequent value of the arrival rate.

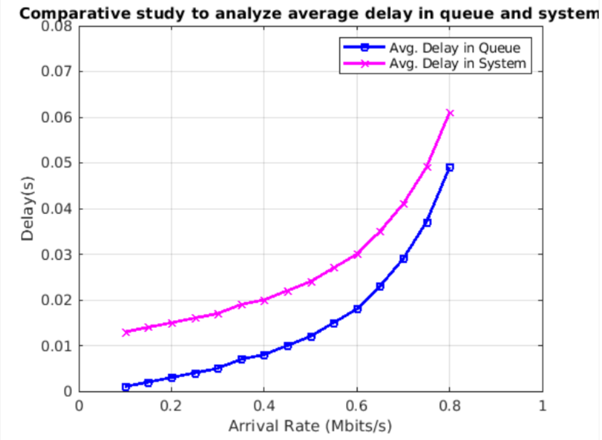

Figura 1: Comparative analysis of Avg. Delay in Queue & System for a range of Arrival rate

4 Conclusion

The randomness of the arrival rate and/or service rate give the delay in a system. For our analysis we do not want a constant delay, so we vary the arrival rate. Therefore, rather than choosing a particular range for interarrival time I decide to choose a certain range of arrival rate for the fixed packet size that was calculated earlier in the above section. From the trend of the graph shown in Fig. 1, I found that when arrival rate increases,the delay in the queue and system also increases exponentially. This can be understood intuitively from the inverse proportionality between the arrival rate and inter arrival time for this case. Thus when the arrival rate was increased, the inter arrival time decreases which actually results in the increment of delay due to large queue length arise from constant service rate.