Applications of Queueing Theory

[1]Rajkumar V
[1]Assistant Professor
[1]Coimbatore Institute of Technology

***Abstract* - Queueing Theory has a wide range of applications, and this article is designed to give an illustration of some of these. It has been divided into 3 main sections, Traffic Flow, Scheduling and Facility Design and Employee Allocation. The given examples are certainly not the only applications where queuing theory can be put to good use, some other examples of areas that queueing theory is used are also given.**

***keywords* - Queuing system, single server model, arrival rate, service rate, infinite and finite models.**

1. Introduction

Queueing Theory is mainly seen as a branch of applied probability theory. Its applications are in different fields, e.g. communication networks, computer systems, machine plants and so forth. For this area there exists a huge body of publications, a list of introductory or more advanced texts on queueing theory is found in the bibliography. Some good introductory books are [1], [2], [3], [4]. The subject of queueing theory can be described as follows: consider a *service center* and a *population* of *customers*, which at some times enter the service center in order to obtain service. It is often the case that the service center can only serve a limited number of customers. If a new customer arrives and the service is exhausted, he has to wait until the service facility becomes available. So we can identify three main elements of a service center: a population of customers, the service facility and the waiting line. Also within the scope of queueing theory is the case where several service centers are arranged in a *network* and a single customer can walk through this network at a specific path for visiting several service centers. As a simple example of a service center consider an airline counter: passengers are expected to check in before they enter into the plane. The check-in is usually done by a single employee; however, there are often multiple passengers. A newly arriving passenger will be directed to the end of the queue, if the service facility (the employee) is busy. This corresponds to a FIFO service (first in, first out). Some examples of the use of queueing theory in networking are the dimensioning of buffers in routers or multiplexers, determining the number of trunks in a central office in POTS, calculating end-to-end throughput in networks and so forth.

Queueing Theory tries to answer questions like e.g. the mean waiting time in the queue, the mean system response time (waiting time in the queue plus service times), mean utilization of the service facility, distribution of the number of customers in the queue, distribution of the number of customers in the system and so forth. These questions are mainly investigated in a stochastic scenario, where e.g. the inter arrival times of the customers or the service times are assumed to be random. The study of queueing theory requires some background in probability theory. Two modern introductory texts are [3] and [5], two really nice “classic” books are [6], [7].

I. Traffic Flow

This is concerned with the flow of objects around a network, avoiding congestion and trying to maintain a steady flow, in all directions.

Queueing on Roads

The study of vehicle flow along a road can be classed as a deterministic flow, because the number of cars flowing within a particular time span can be predicted by averaging previously collected data for that day of the week and time.

For the purposes of analysing the vehicle flow on roads it is normal to monitor the frequency of vehicles in small time groups, over a period of months, and then average out the days in the months, to obtain an average flow for each day of the week. This will have the effect of removing, or at least reducing the effect of, unusual days, when the traffic flow was very different to normal. There is little use in trying to accommodate for these days as the change in traffic flow when something does upset the system is very hard to predict, if it is at all predictable.

I have covered two different types of queues that commonly occur on roads:

Queues Forming at a Motorway Junction

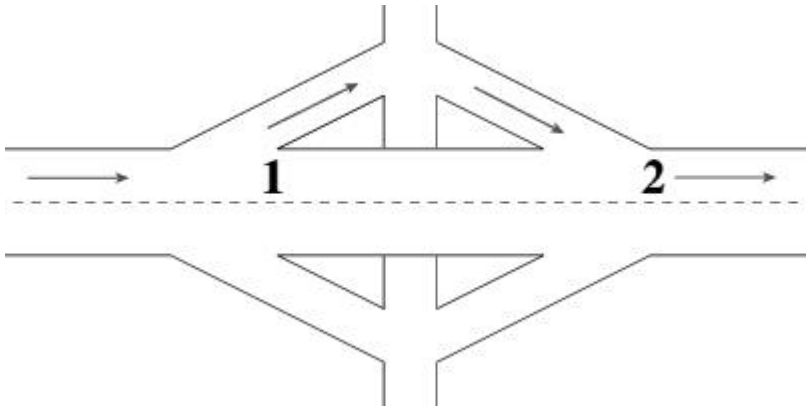

Figure 1

This example is about how a queue caused by traffic trying to enter the motorway can then causes a secondary queue of traffic trying to leave the motorway, as the first queue backs up to pass the exit slip road. Because we are interested how the queue of traffic entering the motorway affects that trying to leave, we say that there are in fact two 'sections' to the queue formed by the merge traffic at point 2:

1. The traffic continuing on the motorway, past point 2
2. The traffic trying to leave the motorway at point 1, as the exit is blocked by the continuing traffic. We shall call this the secondary queueing

Assumptions:

- It is the area immediately downstream of the merge point where the bottle neck occurs (as for all well laid out junctions the merge area is not the cause of the bottle neck)
- The exit slip road is not a bottle neck point (i.e. there is no hold up on the slip road leaving the motorway)
- That the time to get between point 1 and point 2 is constant, regardless of the traffic situation. This is a needed assumption in order to avoid overly complicated expressions

Table 1. Symbols Used

Symbol	Meaning
m	Number of lanes on the motorway
A1m(t)	Cumulative number of through vehicles that would have passed point 1 if there were no queueing
A1s(t)	Cumulative number of exiting vehicles that would have passed point 1 if there were no queueing
A2s(t)	Cumulative number of vehicles from the joining slip road that would have passed point 2 if there were no queuing
A2m(t)	Cumulative number of vehicles from the motorway that would have passed point 2 if there were no queuing
A2(t)	Cumulative number of vehicles that would have arrived at point 2 by time t if there were no queueing
D1m(t)	Cumulative number of vehicles that have passed point one
D2(t)	Cumulative number of vehicles that have actually passed point 2
D2s(t)	Cumulative number of vehicles that have passed point 2, that came from the slip road
D2m(t)	Cumulative number of vehicles that have passed point 2, that came from the motorway
μ	Maximum service rate of motorway at the merge point (i.e. maximum number of cars per hour that can pass point 2)
$\mu_{(t)}$	Service rate of the motorway, at point 2, at time t
μ_s	Maximum service rate of the entry slip road
$\mu_{s(t)}$	Service rate of the entry slip road at time t
t	Time to travel between points 1 and 2, when there is no queue
c	Capacity of the motorway between points 1 and 2 (how many vehicles can queue in that stretch of motorway)

Modeling the traffic flow

As the traffic on the motor way can come from either the entry slip road or from the motorway we can say that A2(t) = A2s(t) + A2m(t). At first sight the number of cars (and hence the length) of the queue would appear to be A2(t) - D2(t), however this is not strictly the case, as the length of the queue is this, *plus* the number of cars that would have been in the area taken up by the queue, had there not been a queue. However the area between D2(t) and A2(t) on a graph of time against cumulative count gives the total delay to all vehicles that pass point 2.

Most motorway junctions are designed so that when there is a large hold up the traffic on the entry slip road will enter the motorway alternately with the traffic on the "slow lane". Thus $\mu_s = \mu/2$, and so the service rate on the motorway for through traffic is $\mu - \mu_s$. This also means that the through traffic on the motorway is occupying m-1/2 lanes of the motorway.

If through traffic on the motorway isn't using all of its service rate then the entry slip road's traffic can use up the slack in the system (up to the maximum limit of the slip road), and thus effectively the joining traffic gets service priority, and vice versa,

when the entry slip road's traffic isn't using its full service rate then motorway through traffic can use up the slack, and so it gets service priority. Most commonly what happens is that a queue builds up on the motorway, but the slip road remains free of queueing traffic, as the slip road operates under its full capacity, μ_s.

It can be said that:

- D2m(t) = D2(t) - D2s(t), as the vehicles to depart from point 2 must have come from either the motorway or the slip road
- D1m(t) = D2m(t) + c
- A1m(t) = A2m(t + *t*), as if there were no queueing then the number of vehicles arriving at point 2 must be equal to the number arriving at point 1, *t* units of time ago, that were going to carry onto point 2

If the queue has not yet reached point 1 then A1m(t) - D2m(t) is the number of vehicles between points 1 and 2, either in a queue or moving along that stretch of motorway. Thus A1m(t) - D2m(t) - c = The number of through cars that have been delayed that have yet to reach point 1.If we presume that there isn't a lane reserved solely for the use of exiting traffic then the exiting traffic will have to share a lane with through traffic, thus the delay for exiting traffic arriving at point 1 at a time T will be the same for a through vehicle arriving at point 1 at time T.

The delay for all through vehicles will be equal to the area between the curve of A1m(t) (which equals A2m(t + *t*)), and D1m(t) (which equals D2m(t) + c). Thus the delay for a single vehicle is the integral from 0 to T of (A2m(t + *t*) - D2m(t) - c) with respect to t / the number of vehicles:

$$Delay\ per\ car = \frac{\int_0^T [A2m(t+t)]dt - \int_0^T [D2m(t+t)+c]dt}{A1m(T) - D2m(T)}$$

In order for there to be no effect on traffic leaving the motorway there would need to be one lane reserved solely for traffic exiting the motorway at point 1, and in which case:

- The exit junction would not affect the delay time in the queue
- The queue would actually be longer, and so the speed of the queue traffic would be higher
- The queue wouldn't affect the exiting of traffic

This is a solution used on many stretches of motorway, where there is enough space.

Queues forming during the Rush Ho

This example discusses how it is possible to model the buildup of traffic in the rush hour, and thus to see when queues form on a straight road, without junctions, relative to the arrival rate of cars on the section of road.

Table 2. Symbols Used

Symbol	Represents
Q(t)	Queue length at time t
A(t)	Cumulative arrivals to the queue at time t
D(t)	Cumulative departures from the queue at time t
μ	The maximum service rate
$\mu_{(t)}$	The service rate at time t
$\lambda_{(t)}$	Arrival rate at time t

Modeling the Queue

In the rush hour the arrival rate, $\lambda_{(t)}$, will rise until it reaches a maximum rate (i.e. the time when the most people reach the point of road, when they are trying to get to work), after this point the arrival rate will decrease. There are four main points in this flow, as follows:

Table 3. Symbols Used

Time label	Significance
t0	The first point at which $\lambda_{(t)}$ equals the maximum service rate, μ
t1	The point where $\lambda_{(t)}$ is a maximum
t2	The point where $\lambda_{(t)}$ has decreased back down to μ
t3	The point where the queue has gone, and so $\mu_{(t)} = \lambda_{(t)}$

The above details tend to imply that a graph of $\lambda_{(t)}$ is quadratic, and so would be differentiable twice at or near t1.This means that $\lambda_{(t)}$ has a Taylor's expansion about the point t1:

$$\lambda(t) = \lambda(t) - \beta(t - t_1)^2$$

For constant β:

$$\beta = -\frac{1}{2}\frac{d^2\lambda(t)}{dt^2}\bigg|_{t=t_1}$$

For small t-t1, i.e. this is valid where t0 < t < t3
This means that t0 and t2 can be estimated, by substitution:

$$\begin{aligned}\mu &= \lambda(t_0)\\ &= \lambda(t_1) - \beta(t_0 - t_1)^2\\ t_0 &= t_1 - \left(\frac{\lambda(t_1) - \mu}{\beta}\right)^2\\ t_2 &= t_1 + \left(\frac{\lambda(t_1) - \mu}{\beta}\right)^2\end{aligned}$$

The length of the queue is the area between the curves of λ(t) and μ(t), i.e.:

$$\begin{aligned}Q(t) &= A(t) - D(t)\\ &= \int_{t_0}^{t}[\lambda(t) - \mu(t)]dt\end{aligned}$$

By the using similar substitutions to those used in the previous derivation we can simplify this to:

$$\begin{aligned}Q(t) &= \int_{t_0}^{t}[\lambda(t) - \mu(t)]dt\\ &= \beta(t - t_0)^2\left(\frac{t_2 - t_0}{2} - \frac{t - t_0}{3}\right)\end{aligned}$$

II. Scheduling
Computer Scheduling
There are many different strategies used for scheduling in computers. This section examines a simple strategy, with no allowance for process prioritization, this is the Batch Processing algorithm, and has the following properties:

- Each process gets as much time as it needs to complete
- There is a single FIFO queue for processes waiting to be processed
- New processes are placed on the back of the queue

The model can be considered as an M/G/1 case. It is clear that the average response time of the system for a process that takes x seconds to complete, T(x), will be the time spent waiting in the queue, plus the time to complete the process:

$$T(x) = \frac{W_0}{1 - \rho} + x$$

For an M/G/1 queue:

$$W_0 = \frac{\lambda\overline{x^2}}{2}$$

Notice how the wait time is totally independent of the job's length, thus batch processing is completely indiscriminate of job length, but this does have the disadvantage that short jobs get no priority and thus batch processing isn't suitable for a time shared system. From the above we can substitute in **W0** to give:

$$T(x) = \frac{\lambda\overline{x^2}}{2(1 - \lambda\bar{x})} + x$$

III. Facility Design and Employee Management
Queues in a Bank
As a rule people dislike queueing in banks, and so in order to improve the level of service a bank manager is interested to find out:

1. The average number of people waiting in the bank (i.e. the queue length)
2. How much of the cashiers' time is spent idle (as a percentage of their time)

Depending on how many tellers she employs during lunchtime. She is prepared to employ up to 5 cashiers, but not less than 1.
Model Assumptions and Details:

- The distribution of the length of time it takes the cashiers to carry out a task is exponential, with mean 2 minutes and standard deviation of 5/4 minutes

- There is virtually no limit on the queue length, as the bank has a large floor area
- Customers arrive in a Poisson distribution, with mean of 25 per hour
- Service is done on a first come first serve basis
- This is an M/M/c queue, where 1 =< c =< 5

Table 4. Symbols Used

Symbol	Meaning, and value if in assumptions
c	Number of cashiers (servers)
Lq	Average length of the queue
W	Percentage of cashiers idle (waiting) time
λ	Mean arrival rate (25 per hour)
μ	1/(mean time to server customers) = 30 per hour
p	Average amount of work for each server, per hour

A Mail Sorting Office

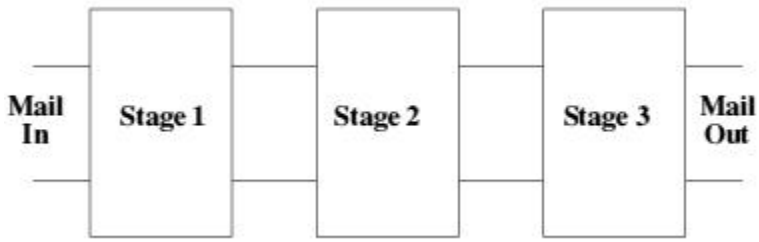

Figure 2

In a mail sorting office the letters to be sorted often come in as a batch, and it is important to organize workers in the way that will get all the post sorted as quickly as possible. Many post companies used to move the bulk of their workers along with the mail, i.e. all workers were initially made to help at the start of the sorting line, to reduce the backlog, and then were gradually moved along the sorting line, as the backlog moved along it. However Queueing Theory can be used to see if this is indeed the quickest way to sort mail, or to suggest the quickest way, if this isn't it. For this example stage refers to a stage in the process of sorting, and server refers to a worker who sorts letters, at any stage.

Assumptions and Simplifications:

The following simplifying assumptions are made:

1. There is a fixed number of sorters
2. All sorters sort at a fixed, continuous rate, which is independent of which stage they are at
3. The cumulative input to the first of the n stages is recorded
4. The transit time of letters between stages and in stages is ignored (i.e., the time letters are in the system is due to queuing between stages)
5. The number of letters in the system doesn't affect the system performance
6. There are so many servers that we can ignore that fact that they are distinct servers

Table 5. Symbols Used

Symbol	Represents
N	Number of stages (the stages are numbered from 1 to n)
Ak(t)	Cumulative number of letters that have arrived at the stage number k
Dks(t)	Cumulative number of letters that have departed from stage number k, into the input queues of the next server
Dkq(t)	Cumulative number of letters that have departed the queue for stage number k
Dk(t)	Cumulative number of letters that have departed from stage number k, this is equivalent to Dkq(t), as all letters that go into a server must also leave the server, and by our assumptions letters spend a negligible time in the server.
μ	The service rate for all servers combined
μ_k	The service rate for the kth stage

Positioning the Workers by the fact that $\mu_k(t)$ is the service rate of the kth stage then:

$$\sum_{j=1}^{n} \mu_k(t) = \mu$$

We want to either:

a. minimize the total delay through the system (i.e., each letter is in the system for the smallest possible time), or:

b. maximize the number of letter that have departed the system after a given time

Both of these can be achieved by maximizing Dn(t). This is constrained by the values of Dk(t) for previous stages, as the final output can't be more that the output from previous stages. Thus:

$$D_n(t) \leqslant D_{n-1}(t) \leqslant \ldots \leqslant D_1(t) \leqslant A_1(t)$$

The achieve either a. or b. we wish to maximize Dn(t), hence:

$$D_n(t) = D_{n-1}(t) = \ldots = D_1(t) = A_1(t)$$

Hence all the stages should have the highest possible service rate, which is thus μ/n, this would yield the highest Dk(t) for all k. It is never good to have a μ(k-1)> μk, as this will create a queue between stages k-1 and k, and the excess service rate used to create the queue could have been better used at the next stage to avoid the queue forming in the first place.

Summary of Results

1. It is best to have the Postal Workers sorting mail distributed evenly between all the stages
2. It is never advantageous to have a μk greater than the next stage.

Some Other Examples

1. Design of a garage forecourt
2. Airports - runway layout, luggage collection, shops, passport control etc.
3. Hair dressers
4. Supermarkets
5. Restaurants
6. Manufacturing processes
7. Bus scheduling
8. Hospital appointment bookings
9. Printer queues
10. Minimising page faults in computing

References

[1] Leonard Kleinrock. *Queueing Systems – Volume 1: Theory*, volume 1. John Wiley and Sons, New York, 1975.

[2] Arnold O. Allen. *Probability, Statistics, and Queueing Theory – With Computer Science Applications*.

[3] Randolph Nelson. *Probability, Stochastic Processes, and Queueing Theory – The Mathematics of* Computer *Performance Modeling*. Springer Verlag, New York, 1995.

[4] Phuoc Tran-Gia. *Analytische Leistungsbewertung verteilter Systeme – eine Einf ¨uhrung*. Springer Verlag, Berlin, 1996.

[5] S. M. Ross. *A First Course in Probability*. Macmillan, fourth edition, 1994.

[6] William Feller. *An Introduction to Probability Theory and Its Applications - Volume II*. John Wiley, New York, second edition, 1968.

[7] William Feller. *An Introduction to Probability Theory and Its Applications - Volume I*. John Wiley, New York, third edition, 1968.

[8] Fundamentals of Queueing Theory, Third Edition, Donald Gross & Carl M. Harris, Wiley- Inter Science, 1998

[9] Applications of Queueing Theory, Second Edition, G. F. Newell, Chapman and Hall, 1982

[10] Queueing Systems Volume II: Computer Applications, Leonard Kleinrock, Wiley-Inter Science, 1976

[11] Introduction to Queueing Theory, Fourth Edition, Robert B. Cooper, Ceep Press Books, 1990