3

Queuing Theory

Contents

3.1 Introduction

The objective of queuing analysis is to offer a reasonably satisfactory service to waiting customers. Unlike the other tools of OR presented in the preceding chapters, queuing theory is not an optimization technique. Rather, it determines the measures of performance of waiting lines, such as the average waiting time in the queue and the productivity of the service facility, which can then be used to design the service installation.

The queuing theory also called waiting line theory, owes its development to A. K. Erlang's efforts to analyze telephone traffic congestion with a view to satisfying the randomly arising demand for the service of the Copenhagen automatic telephone system, in the year 1909. The theory is applicable to situations where the customers arrive at some service station/stations, for some service, wait (occasionally not); and then leave the system after getting the service.

In such 'arrival and departure' problems, the customers might be people waiting to deposit their gas bills at a cash counter, machine waiting to be repaired in factory's repair shop, aeroplanes waiting to land at an airport, patients in hospital waiting for treatment, and so on. The service stations in such problems are the cash counters in gas company's office, a repairman in the shop, runway at the airport, and the doctors attending the patients, respectively. Some more examples of the queuing system are given in *Table 3.1*.

Table 3.1 - Queuing Examples

Situation	Arriving Customer	Service Facility
Passage of customers through a supermarket checkout	Shoppers	Checkout counters
The flow of automobile traffic through road network	Automobiles	Road network
Transfer of electronic messages	Electronic messages	Transmission lines
Banking transactions	Bank patrons	Bank tellers
The flow of computer programmes through a computer system	Computer programmes	Central Processing Unit
Sale of movie tickets	The person going to watch movie	Ticket booking window
The arrival of trucks at marketing yard	Trucks	Loading crews and facilities
Registration of students at job fair	Students	Registration assistants

The queuing theory developed because service to a customer may not be rendered immediately as the customer reaches the service facility. Thus, inadequate service facilities would cause waiting lines of customers to be formed. The only way that the service demand can be met with ease to increase the service capacity (and raising the efficiency of the existing capacity if possible) to a higher level.

The capacity might be built to such a high level as can always meet the peak demand with no queues. But adding to capacity may be costly and uneconomical after a stage because then it shall remain idle to varying degrees when there are no or few customers. A manager, therefore, has to decide on an appropriate level of service which is neither too low nor too high. Inefficient or poor service would cause excessive waiting which has a cost in terms of customer frustration, loss of goodwill in the long run, direct cost of idle employees. On the other hand, too high a service level would result in very high set-up cost and idle time for the service station/stations. Thus the goal of queuing model is the achievement of an economic balance between the cost of providing service and the cost associated with the wait required for that service.

3.2 General Structure of the Queuing System

The general structure of a queuing system is shown in *Fig.3.1*.

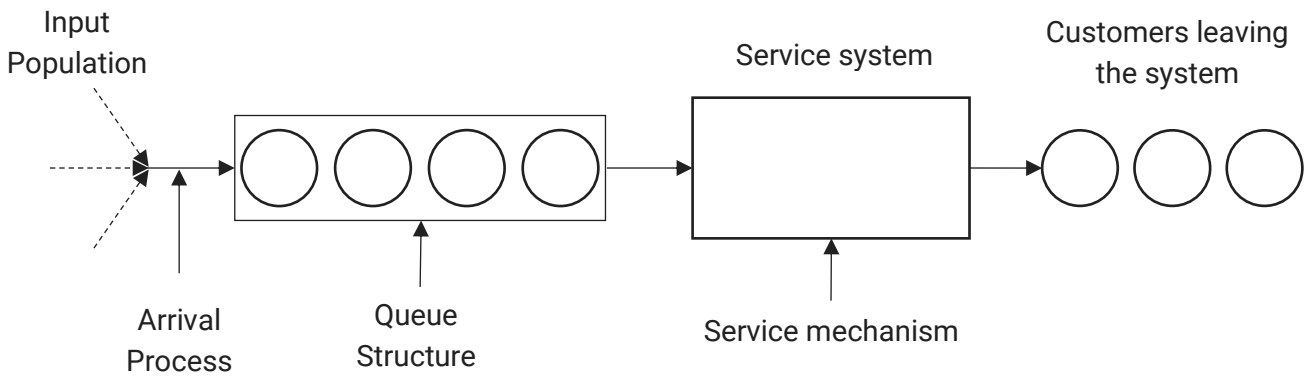

Fig.3.1 – General structure of the queuing system

Components of the queuing system are arrivals process, the customer waiting in the queue, the customer being served, the service facility and the customer leaving the queue after service.

3.2.1 Arrival Process or Input Process

The arrival describes the way in which the customers arrive and join the system. In general customer arrival will be in random fashion, which cannot be predicted because the customer is an independent individual and the service organization has no control over the customer. The characteristics of arrival are shown in *Fig.3.1*.

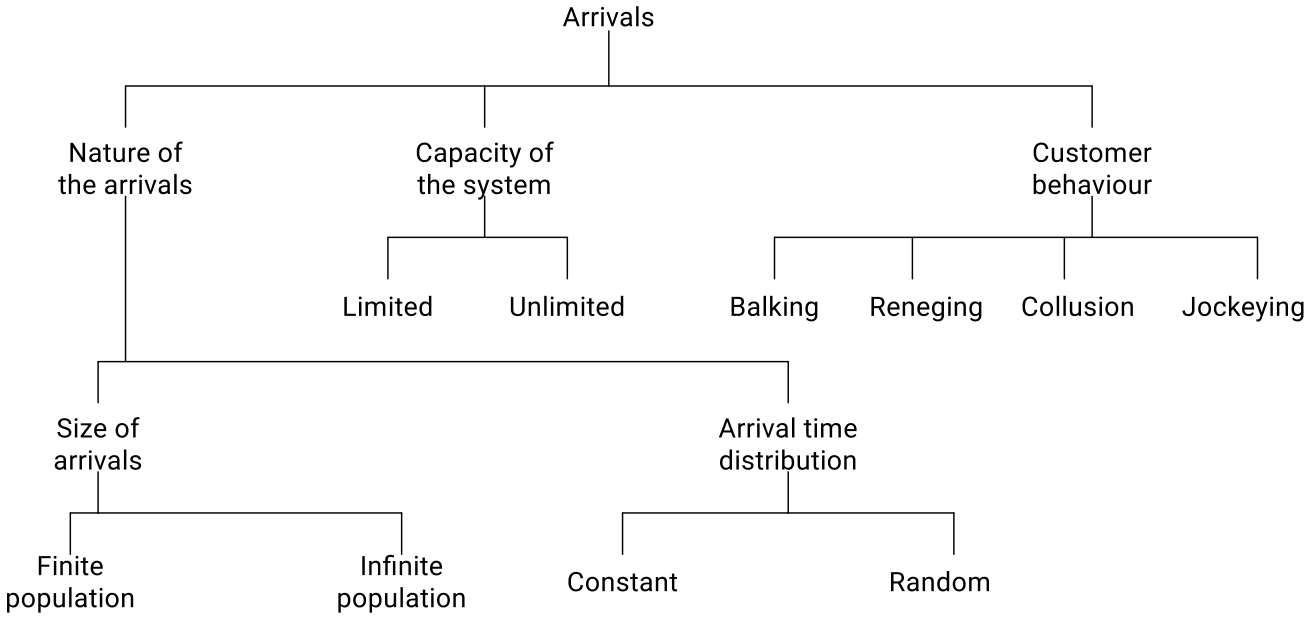

Fig.3.2 – Characteristics of arrivals or input.

Input to the queuing system refers to the pattern of arrival of customers at the service facility. We can see at ticket counters or at petrol pumps or any such service facility that the customer arrives randomly individually or in batches. The input process is described by the following characteristics (as shown in *Fig.3.2*) nature of arrivals, capacity of the system and behavior of the customers.

a) **Size of arrivals:** The source of customers for a queuing system can be infinite or finite. For example, all people in a city or state (and others) could be the potential customers at a supermarket. The number of people being very large, it can be taken to be infinite. On the other

hand, there are many situations in business and industrial conditions where we cannot consider population to be infinite as it is finite. Thus, the ten machines in a factory requiring repairs and maintenance by the maintenance crew would be a typical example of finite population. Removing one machine from a small, finite, population like this will have a noticeable effect on the calls expected to be made (for repairing) by the remaining machines than if there were a large number of machines, say 500.

b) Inter-arrival time: Customers may arrive in the system at known (regular or otherwise) times, or they might arrive in random way. The queuing models wherein customers' arrival times are known with certainty are characterized as deterministic models and are easier to handle. On the other hand, a substantial majority of the queuing models are based on the assumption that the customers enter the system stochastically, at random points in time.

With random arrivals, the number of customers reaching the system per unit time might be described by a probability distribution. Although the arrivals might follow any pattern, the frequently employed assumption, which adequately supports many real-world situations, is that the arrivals are Poisson distributed.

c) Capacity of the service system: In the queuing context, the capacity refers to the space available for the arrivals to wait before taken to service. The space available may be limited or unlimited. When space is limited, length of waiting line crosses a certain limit; no further units or arrivals are permitted to enter the system till some waiting space becomes vacant. This type of system is known as system with finite capacity and it has its effect on the arrival pattern of the system, for example a doctor giving tokens for some customers to arrive at certain time and the present system of allowing the devotees for darshan at Tirupathi by using the token belt system.

d) Customer behaviour

The length of the queue or the waiting time of a customer or the idle time of the service facility mostly depends on the behaviour of the customer. Here the behaviour refers to the impatience of a customer during the stay in the line. Customer behaviour can be classified as:

(i) Balking: This behaviour signifies that the customer does not like to join the queue seeing the long length of it. This behaviour may affect in losing a customer by the organization. Always a lengthy queue indicates insufficient service facility and customers may not turn out next time. For example, a customer who wants to go by train to his destination goes to railway station and after seeing the long queue in front of the ticket counter, may not like to join the queue and seek other types of transport to reach his destination.

(ii) Reneging: In this case, the customer joins the queue and after waiting for certain time loses his patience and leaves the queue. This behaviour of the customer may also cause loss of customer to the organization.

(iii) Collusion: In this case, several customers may collaborate and only one of them may stand in the queue. One customer represents a group of customers. Here the queue length may be small but service time for an individual will be more. This may break the patience of the other customers in the waiting line and situation may lead to any type of worst episode.

(iv) Jockeying: If there are number of waiting lines depending on the number of service stations, for example, Petrol pumps, Cinema theatres, etc. A customer in one of the queues after seeing the other queue length, which is shorter, with the hope of getting the service, may leave the present queue and join the shorter queue. Perhaps the situation may be that other queue which is shorter may be having more number of collaborated customers. In such case the probability of getting service to the customer who has changed the queue may be very less. Because of this character of the customer, the queue lengths may go on changing from time to time.

3.2.2 Service Mechanism or Service Facility

There are two aspects of the service system–(a) structure of the service system, and (b) the speed of service.

a) Structure of the service system

By the structure of the service system, we mean how the service facilities exist. There are several possibilities as discussed below:

1. ***A single service facility***
   A library counter is an example of this. The models which involve a single service facility are called *single server models. Fig.3.3***Error! Reference source not found.** illustrates such a model.

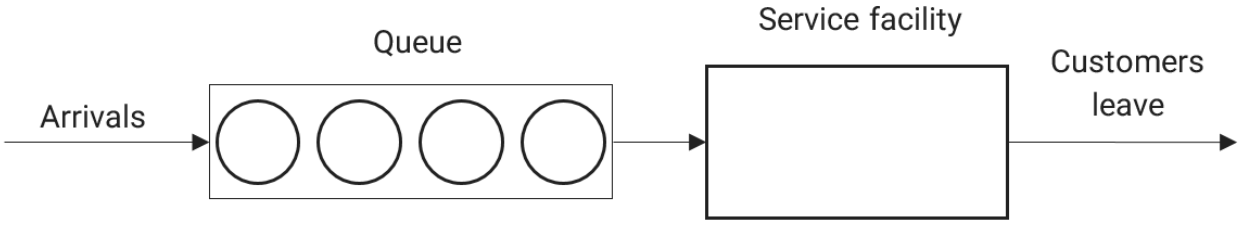

Fig.3.3 – Single Server, Single Queue Model

2. ***Multiple, parallel facilities with a single queue***
   That is, there is more than one server. The term parallel implies that each server provides the same type of facility. Booking at a service station that has several mechanics, each handling one vehicle, illustrates this type of model. It is shown in *Fig.3.4*

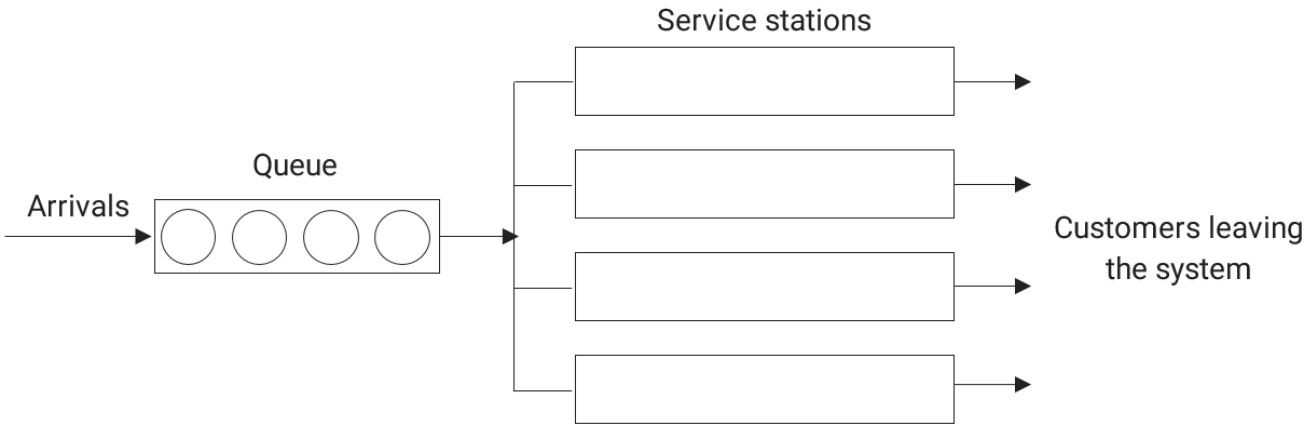

Fig.3.4 – Multiple, Parallel Servers, Single Queue Model

3. ***Multiple, parallel facilities with multiple queues***

This type of model is different from the earlier one only in that each of the servers has a different queue. Different cash counters in an electricity office where the customers can make payment in

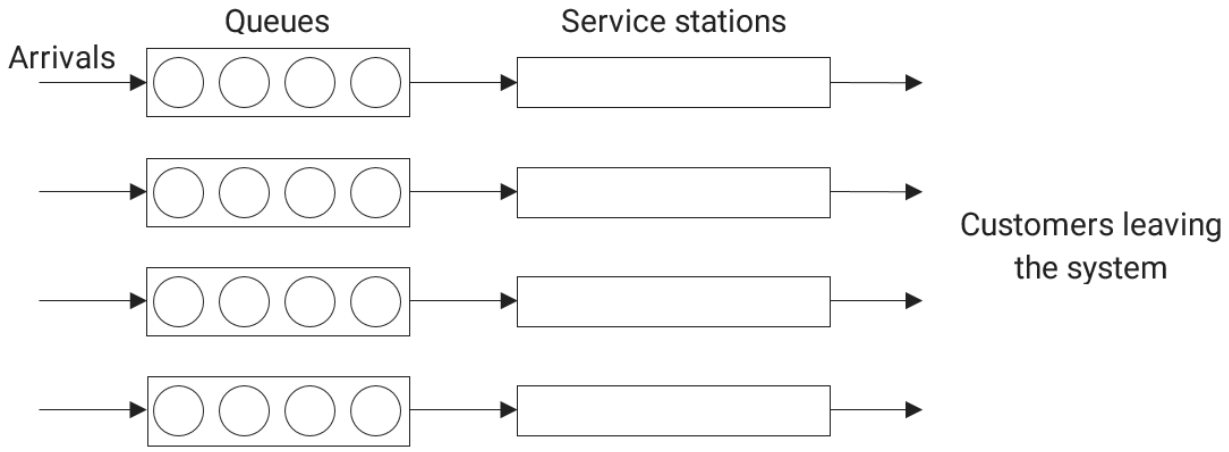

Fig.3.5 – Multiple, Parallel Severs, Multiple Queues Model

respect of their electricity bills provide an example of this type of model. *Fig.3.5* shows such type of model.

4. ***Service facilities in a series***

In this, a customer enters the first station and gets a portion of service and then moves on to the next station, gets some service and then again moves on to the next station……and so on, and finally leaves the system, having received the complete service. For example, machining of a certain steel item may consist of cutting, turning, knurling, drilling, grinding, and packaging operations, each of which is performed by a single server in a series. *Fig.3.6* shows this type of model.

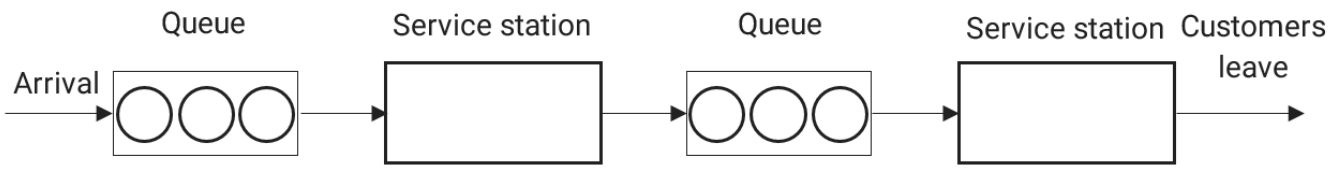

Fig.3.6 – Multiple servers in series

b) Speed of service

In a queuing system, the speed with which service is provided can be expressed in either of two ways–as service rate and as service time. The service rate describes the number of customers serviced during a particular time period. The service time indicates the amount of time needed to service a customer. Service rates and times are reciprocals of each other and either of them is sufficient to indicate the capacity of the facility. Thus, if a cashier can attend, on average, to 10 customers in an hour, the service rate would be expressed as 10 customers/hour and service time would be equal to 6 minutes/customer. Generally, however, we consider service time only.

If these service times are known exactly, the problem can be handled easily. But, as generally happens, if these are different and not known with certainty, we have to consider the distribution of the service times in order to analyse the queuing system. Generally, the queuing models are based on the assumption that service times are exponentially distributed about some average service time.

3.2.3 Queue discipline or Service discipline or Queue structure

Another element of a queuing system is the queue structure. In the queue structure, the important thing to know is the queue discipline which means the order by which customers are picked up from the waiting line for service. There are various ways in which the customer called to serve. They are as follows.

1. ***First In First Out (FIFO) or First Come First Served (FCFS)***
   We are quite aware that when we are in a queue, we wish that the element which comes should be served first, so that every element has a fair chance of getting service. Moreover, it is understood that it gives good morale and discipline in the queue. When the condition of FIFO is violated, there arises the trouble and the management is answerable for the situation.
2. ***Last in First out (LIFO) or Last Come First Served (LCFS)***
   In this system, the element that arrived last will have a chance of getting service first. In general, this does not happen in a system where human beings are involved. But this is quite common in the inventory system. Let us assume a bin containing some inventory. The present stock is being consumed and suppose the material ordered will arrive that is loaded into the bin. Now the old material is at the bottom of the stock whereas fresh arrived material at the top. While consuming the top material (which is arrived late) is being consumed. This is what we call Last come first served). This can also be written as First In Last Out (FILO).
3. ***Service In Random Order (SIRO)***
   Random order of service is defined as whenever a customer is chosen for service, the selection is made in a way that every customer in the queue is likely to be selected. The time of arrival of the customer is, therefore of no consequence in such a case. It is seen to allocate an item whose demand is high and supply is low, also seen in the allocation of shares to the applicants to the company.
4. ***Service by priority***
   The customers in a queue might be rendered service on a priority basis. Thus customers may be called according to some identifiable characteristics for service. As an example, we can quote that in a hospital, when the doctor is treating a patient with stomach pain, suddenly a patient with heart stroke enters the hospital, the doctor asks the patient with stomach pain to wait for some time and give attention to a heart patient. This is the rule of priority.

3.3 Operating Characteristics of the Queuing System

An analysis of a given queuing system involves a study of its different operating characteristics. This is done using queuing models. Some of the more commonly considered characteristics are discussed below:

1. **Queue length:** The average number of customers in the queue waiting to get service. Large queues may indicate poor server performance while small queues may indicate too much server capacity.
2. **System length:** the average number of customers in the system, those waiting to be served and those who are getting service.
3. **Waiting time in the queue:** The average time that a customer has to wait in the queue to get service.
4. **Total time in the system:** The average total time spent by a customer in the system from the moment he arrives till he leaves the system. It is taken to be the waiting time plus the service time.
5. **Traffic intensity (Utilization factor):** It is the proportion of time a server actually spends with the customers.
6. **Server idle time:** The relative frequency with which the service system is idle. Idle time is directly related to cost.

3.4 Kendall's Notation for Representing Queuing Model

D. G. Kendall (1953) and late A. Lee (1966) introduces useful notation for queuing models. The complete notation can be expressed as

(a/b/c) : (d/e/f)

where

a = arrival (or interarrival) distribution
b = departure (or service time) distribution
c = number of parallel service channels in the system
d = service discipline
e = maximum number of customers allowed in the system
f = calling source or population

The following conventional codes are generally used to replace the symbol a, b and d.

Symbols for a and b

M = Markov (or Poisson) arrival or departure distribution (or exponential interarrival or service time distribution),
E_k = Erlangian or gamma interarrival or service time distribution with parameter k,
GI = General independent arrival distribution
G = General departure distribution
D = Deterministic interarrival or service times

Symbols for d

FCFS = First come first served
LCFS = Last come first served
SIRO = Service in random order
GD = General service discipline

The symbols e and f represents a finite (N) or infinite (∞) number of customers in the system and calling source respectively.

For example, $(M/E_k/1)$: (FCFS/N/∞) represents Poisson arrival (exponential interarrival), Erlangian departure, single server, 'first come first served' discipline, maximum allowable customers N in the system and infinite population model.

3.5 Classification of Queuing Models

The various types of queuing models can be classified as follows:

3.5.1 Probabilistic Queuing Models

1. **Model I (Erlang Model):** This model is symbolically represented by (M/M/1) : (FCFS/∞/∞). This represents Poisson arrival (exponential interarrival), Poisson departure (exponential service time), single server, first come, first served service discipline, infinite number of customers allowed in the system and infinite population. Since the Poisson and exponential distribution are related to each other, both of them are denoted by the symbol 'M; due to Markovian property of exponential distribution.
2. **Model II (General Erlang Model):** Though this model is also represented by (M/M/1) : (FCFS/∞/∞), it is a general queuing model in which the arrival and service rate depend upon the length of the queue. Some persons desiring service may not join the queue since it is too long, thus affecting the arrival rate. Similarly, the service rate is also affected by the length of the queue.
3. **Model III:** This model is represented by (M/M/1) : (SIRO/∞/∞). It is essentially the same as model I except the service discipline is SIRO instead of FCFS.
4. **Model IV:** This model is represented by (M/M/l) : (FCFS/N/∞). In this model the capacity of the system is limited or finite, say N. So the number of arrivals cannot exceed N.

5. **Model V:** This model is represented by (M/M/I) : (FCFS/n/M). It is finite population or limited source model. In this model, the probability of arrival depends upon the number of potential customers available to enter the system.
6. **Model VI:** This model is represented by (M/M/c) : (FCFS/∞/∞). This is the same as model I except that there are c service channels working in parallel.
7. **Model VII:** This model is represented by ($M/E_k/1$) : (FCFS/∞/∞). In this model, instead of exponential service time, there is Erlang service time with k phases.
8. **Model VIII:** This model is represented by (M/M/1): (GD /m/n), where m<n. It represents machine repair problem with a single repairman, n is the total number of machines out of which m are broken down and forming a queue. GD represents a general service discipline.
9. **Model IX:** This model is represented by (M/M/c): (GD/m/n), m ≤ n. It is same as model VIII except that there are c repairmen, c < n.
10. **Model X:** This is called the power supply model.

3.5.2 Deterministic Queuing Models

11. **Model XI:** This model is represented by (D/D/1) : (FCFS/∞/∞). In this model interarrival time, as well as service time, are fixed and known with certainty. The model is, therefore, called deterministic model.

3.5.3 Mixed Queuing Models

12. **Model XII:** This model is represented by (M/D/1) : (FCFS/∞/∞). Here, arrival rate is Poisson distributed while the service rate is deterministic or constant.

3.6 Model I. Single channel Poisson Arrivals with Exponential Service, Infinite Population Model [(M/M/1) : (FCFS/∞/∞)]

Let us consider a single-channel system with Poison arrivals and exponential service time distribution. Both the arrivals and service rates are independent of the number of customers in the waiting line. Arrival are handled on first come, first served basis. Also the mean arrival rate λ is less than the mean service rate μ.

The following notifications and equations will be used in connection with queuing models:

λ = mean arrival rate (average number of arrivals per unit of time)

μ = mean service rate (average number of units served per unit of time)

ρ = Traffic Intensity (or utilization factor) which represents the proportion of time the servers are busy

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

L_s = Average number of customers in the system or length of the system

$$L_s = \frac{\lambda}{\mu - \lambda}$$

L_q = Average number of customers waiting in the queue or length of the queue

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

W_s = Average waiting time of a customer in the system both waiting and in service

$$W_s = \frac{1}{\mu - \lambda}$$

W_q = Average waiting time of a customer in the queue

$$W_q = \frac{\lambda}{\mu}\left(\frac{1}{\mu - \lambda}\right)$$

Probability that the system is empty $P_0 = (1 - \rho)$

3.7 Examples

Ex. 3.1 Patrons arrive at a reception counter at an average inter-arrival rate of 2 minutes. The receptionist in duty takes an average of 1 minute per patrons. A) What is the chance that patrons will straightway meet the receptionist? B) For what portion of time the receptionist is busy. C) What is the average queue length? D) What is the average no. of patrons in the system? E) What is the average waiting time of a patron? F) What average time a patron spends in the system? G) Suppose management went to keep a second receptionist when the average waiting time of an arrival excess 1.5 minutes, find what should be the average interarrival time to justify a second receptionist.

Solution:

Ex. 3.2 Arrivals at a telephone booth are considered to be Poisson with an average time of 10 minutes between one arrival and then next. The length of the phone calls is assumed to be distributed exponentially with a mean of 3 minutes. (a) What is the probability that a person arriving at the booth will have to wait? (b) What is the average length of the queue that is formed from time to time? (c) The telephone company will install a second booth when convinced that an arrival would have to wait at least three minutes for the phone to be free. By how much flow of arrivals is increased in order to justify a second booth?

Solution: