**(Unit 2) Queuing Theory**

The mechanism of a queuing process is very simple. Customers arrive at services counter are attended by one or more of the servers. As soon as a customer is served, he departs from the system. Thus a queuing system can be described as composed of customers arriving for service, waiting for service if it is not immediate, and if having wanted for service, leaving the system after being served.



**16.3  Component of a Queuing System**

A queuing system can be described by the following components:

**16.3.1  Input process (or Arrival pattern)**

This is considered with the pattern in which the customers arrive and join the system. An input source is characterized by

a)      Size of the calling population.

b)      Pattern of arrivals at the system.

c)      Behaviour of the arrivals.

Customers requiring service are generated at different times by an input source, commonly known as population. The rate at which customers arrive at the service facility is determined by the arrival process.

***16.3.1.1  Size of the calling population***

The size represents the total number of potential customers who will require service. The source of customers can be either finite or infinite. It is considered infinite if the number of people being very large e.g. all people of a city or state (and others) could be the potential customers at a milk parlour. Whereas there are many situations in industrial conditions where we cannot consider the population to be infinite�it is finite. The customers may arrive for service individually or in groups. Single arrivals are illustrated by a customer visiting a milk parlour, students arriving at a library counter etc. On the other hand, families visiting restaurants, ships discharging cargo at a dock are examples of bulk or batch arrivals.

***16.3.1.2  Pattern of arrivals at the system***

Customers arrive in the system at a service facility according to some known schedule (for example one patient every 15 minutes or a candidate for interview every half hour) or else they arrive randomly. Arrivals are considered at random when they are independent of one another and their occurrence cannot be predicted exactly. The queuing models wherein customers� arrival times are known with certainty are categorized as deterministic models and are easier to handle. On the other hand, a substantial majority of the queuing models are based on the premise that the customers enter the system stochastically, at random points in time. The arrival process (or pattern) of customers to the service system is classified into two categories: **static**and **dynamic**. These two are further classified based on the nature of arrival rate and the control that can be exercised on the arrival process.

In static arrival process, the control depends on the nature of arrival rate (random or constant). Random arrivals are either at a constant rate or varying with time. Thus to analyze the queuing system, it is necessary to describe the probability distribution of arrivals. From such distributions average time between successive arrivals, is obtained also called inter-arrival time (time between two consecutive arrivals), and the average arrival rate (i.e. number of customers arriving per unit of time at the service system).

The dynamic arrival processis controlled by both service facility and customers. The service facility adjusts its capacity to match changes in the demand intensity, by either varying the staffing levels at different timings of service, varying service charges (such as telephone call charges at different hours of the day or week) at different timings, or allowing entry with appointments. Frequently in queuing problems, the number of arrivals per unit of time can be estimated by a probability distribution known as the Poisson distribution, as it adequately supports many real world situations

**16.3.2  Service Mechanism (or Service Pattern)**

The service is provided by a service facility (or facilities). This may be a person (a bank teller, a barber, a machine (elevator, gasoline pump), or a space (airport runway, parking lot, hospital bed), to mention just a few. A service facility may include one person or several people operating as a team. There are two aspects of a service system

a)      the configuration of the service system

b)      the speed of the service.

***16.3.2.1  Configuration of the service system***

The customers� entry into the service system depends upon the queue conditions. If at the time of customers� arrival, the server is idle, then the customer is served immediately. Otherwise the customer is asked to join the queue, which can have several configurations. By configuration of the service system we mean how the service facilities exist. Service systems are usually classified in terms of their number of channels, or numbers of servers.

i)      **Single Server � Single Queue**-- The models that involve one queue � one service station facility are called single server models where customer waits till the service point is ready to take him for servicing. Students arriving at a library counter are an example of a single server facility.

ii)    **Single Server � Several Queues**� In this type of facility there are several queues and the customer may join any one of these but there is only one service channel.

iii)  **Several (Parallel) Servers � Single Queue**� In this type of model there is more than one server and each server provides the same type of facility. The customers wait in a single queue until one of the service channels is ready to take them in for servicing.

iv)  **Several Servers � Several Queues**� This type of model consists of several servers where each of the servers has a different queue. Different cash counters in an electricity office where the customers can make payment in respect of their electricity bills provide an example of this type of model. Different ticket issue encounters in a trade fair and different boarding pass encounters at an airport are also other possible examples of this type of model.

v)    **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 in a milk plant packaging of milk pouches consist of boiling, pasteurization, cooling and packaging operations, each of which is performed by a single server in a series.

***16.3.2.2  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 and the service time indicates the amount of time needed to service a customer. Service rates and times are reciprocal of each other and either of them is sufficient to indicate the capacity of the facility. Thus if a cashier can attend, on an average 5 customers in an hour, the service rate would be expressed as 5 customers/hour and service time would be equal to 12 minutes/customer. Generally, we consider the 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 analyze the queuing system. Generally, the queuing models are based on the assumption that service times are exponentially distributed about some average service time.

**16.3.3  Queue discipline**

In the queue structure, the important thing to know is the queue discipline. The queue discipline is the rule determining the formation of queue, manner in which customers form the queue are selected for service. There are a number of ways in which customers in the queue are served. Some of these are:

***16.3.3.1 Static queue disciplines***

Theseare based on the individual customer's status in the queue. The most common queue disciplines are:

i)      **First-Come-First-Served (FCFS):**If the customers are served in the order of their arrival, then this is known as the FCFSservice discipline. For example, this type of queue discipline is observed at a milk parlour, railway station etc. FCFS is also known as First In First Out (FIFO).

ii)    **Last-Come-First-Served (LCFS):** Sometimes, the customers are serviced in the reverse order of their entry so that the ones who join the last are served first and the system is referred to as LCFS. For example, in a big godown the items which come last are taken out first. Similarly, the people who join an elevator last are the first ones to leave it.

***16.3.3.2 Dynamic queue disciplines***

Theseare based on the individual customer attributes in the queue. Few of such disciplines are:

i)      **Service in Random Order (SIRO):** Under this rule customers are selected for service at random, irrespective of their arrivals in the service system. In this, every customer in the queue is equally likely to be selected. The time of arrival of the customers is, therefore, of no relevance in such a case.

ii)    **Priority Service:** Under this rule customers are grouped in priority classes on the basis of some attributes such as service time or urgency or according to some identifiable characteristic, and FCFS rule is used within each class to provide service. Treatment of VIPs in preference to other patients in a hospital is an example of priority service.

**16.3.4  Customer�s behaviour**

Another thing to consider in the queuing structure is the behaviour or attitude of the customers entering the queuing system. On this basis, the customers may be classified as being patient, or impatient. If a customer, on arriving at the service system stays in the system until served, no matter how much he has to wait for service is called a patient customer whereas the customer, who waits for a certain time in the queue and leaves the service system without getting service due to certain reasons such as a long queue in front of him is called an impatient customer. The customers generally behave in four ways

i)           **Balking:** A customer may leave the queue because the queue is too long or the estimated waiting time is too long or waiting space is inadequate, for desired service and may decide to return for service at a later time. In queuing theory this is known as **balking**.

ii)         **Reneging:** A customer, after joining the queue, waits for some time and leaves the service system due to intolerable delay or due to impatience.

iii)       **Jockeying:** A customer who switches from one queue to another, hoping to receive service more quickly, is said to be jockeying.

iv)       **Priorities:** In certain applications some customers are served before others regardless of their order of arrival. These customers have priority over others.

The ultimate objective of the analysis of queuing systems is to understand the behavior of their underlying processes so that informed and intelligent decisions can be made in their management. In a specified queuing system the problem is to determine the probability distribution of queue length, waiting time of customers and the busy period. Queuing theory uses queuing models to represent the various types of queuing systems that arise in practice. Formulae for each model indicate how the corresponding queuing system should perform, including the average amount of waiting time under a variety of circumstances. Therefore, these queuing models are helpful in determining how to operate a queuing system in the most effective way. Providing too much service capacity to operate the system involves excessive costs. But not providing enough service capacity results in excessive waiting and all its unfortunate consequences. The models enable finding an appropriate balance between the cost of service and the amount of waiting.

**17.2  Characteristics of A Queuing System**

Queuing models enable the analyst to study the effect of manipulating decision variables on the operating characteristics of a service system. The most commonly used characteristics are stated as under:

**17.2.1  Queue length**

The average number of customers in the queue waiting to get service is known as �queue length�. �Short queues� could mean either good customer service or large waiting space while �long queues� could indicate low service efficiency or a little waiting space.

**17.2.2  System length**

It is the average number of customers in the system waiting to be served and those being served. Long queues imply congestion, potential customer dissatisfaction and need for more capacity.

**17.2.3  Waiting time in queue**

Waiting time is the average time that a customer has to wait in the queue to get service. Long waiting times may indicate a need to adjust the service rate of the system or change the arrival rate of customers.

**17.2.4  Total time in system**

The average time that customer spends in the system from entry in the queue to completion of service. If this time is more then there may be a need to change the priority discipline, increase productivity or adjust the capacity.

**17.2.5  Server idle time**

The relative frequency with which the service system is idle which is directly related to cost. Queuing theory analysis involves the study of systems� behavior over time.

**17.2.6  Transient and Steady States**

When a service system is started, it progresses through a number of changes. However, it attains stability after some time. Before the service operations start, it is very much influenced by the initial conditions (number of customers in the system) and the elapsed time. This period of transition is termed as **transient state**. A system is said to be in transient-state when its operating characteristics are dependent on time.

However, after sufficient time has passed, the system becomes independent of the initial conditions and of the elapsed time (except under very special conditions) and enters a **steady state**condition. A steady state condition is said to prevail when the behavior of the system becomes independent of time. Let Pn(t) denote the probability that are **n** units in the system at time **t**. We know that the change of Pn(t) with respect to **t** is described by the derivative . Then the queuing system is said to be stable eventually, in the sense that the probability Pn(t) is independent of time, i.e. remains the same as time passes (t → ∞). Mathematically, in a steady state,

 

This implies that

  

 From the practical point of view period of the steady state behavior of the system, queuing system under the existence of steady state condition are being considered.

**17.3  Notations and Symbols**

The notations used in the analysis of a queuing system are as follows:

|  |  |
| --- | --- |
|       n = | Number of customers in the system (waiting and in service) |
| �Pn�(t) = | Transient state probability that n calling units are in the queuing system at time t |
|      En= | The state in which there are n calling units in the system |
|      Pn�= | Steady state probability of having n units in the system |
| λ = | Average (expected) customer arrival rate or average number of arrivals per unit of time in the queuing system |
|  μ = | Average (expected) service rate or average number of customers served per unit time at the place of service |
|  ρ = | Traffic intensity or server utilization factor |
|  s = | Number of service channels (service facilities or servers) |
| N = | Maximum number of customers allowed in the system. |
| Ls= | Average (expected) number of customers in the system (waiting and in service) |
| Lq= | Average (expected) number of customers in the queue (queue length) |
| Lb= | Average (expected) length of non-empty queue |
| Ws= | Average (expected) waiting time in the system (waiting and in service) |
| Wq= | Average (expected) waiting time in the queue |
| Pw= | Probability that an arriving customer has to wait |

**17.3.1  Kendalls notation for representing queuing models**

Generally queuing model can be specified by the symbolic representation (a|b|c) : (d|e)� where,

|  |  |
| --- | --- |
| a : | Probability distribution of the arrival (or inter-arrival) time |
| b : | Probability distribution of the service time. |
| c : | Number of channels  (or service stations) |
| d : | Capacity of the system |
| e : | Queue discipline |

The first three characteristics (a|b|c) in the above notation were introduced by D. Kendall in 1953.  Later in 1966, A. Lee added the fourth (d) and fifth (e) characteristics to the notation. Traditionally, the exponential distribution in queuing problems is denoted by M. Thus, (M|M|1): (∞|FIFO) indicates a queuing system when the inter-arrival times and service times are exponentially distributed having one server in the system with first in first out discipline and the number of customers allowed in the system can be infinite.

**17.4  Traffic Intensity (or Utilization factor)**

An important measure of simple queue is its traffic intensity, where



 

The unit of traffic intensity is Erlang.

A necessary condition for a system to have settled down to steady state is that

  

i.e. arrival rate < service rate. If ρ > 1, the arrival rate is greater than the service rate and consequently, the number of units in the queue tends to increase indefinitely as the time passes on, provided the rate of service is not affected by the length of queue.

**17.5  Queuing Models**

The queuing models are categorized as �deterministic� or �probabilistic�. If each customer arrives at known intervals and the service time is known with certainty, the queuing model will be deterministic in nature. When both arrival and service rate are unknown and assumed to be random variable then this type of queuing model is known as probabilistic.

**17.6  Probability Distributions in Queuing systems**

The arrival of customers at a queuing system varies between one system and another, but in practice one pattern of completely random arrivals is observed.

**17.6.1  Distribution of arrivals �the Poisson process� (pure birth process)**

The models in which only arrivals are counted and no departure take place are called pure birth models. In terms of queuing, birth-death process that is increased by birth or arrival in the system and decreased by death or departure of serviced customer from the system. If the arrivals are completely random, then the probability distribution of number of arrivals in a fixed time�interval, follows Poisson probability distribution with parameter (mean) λt. Thus,

 

**17.6.2  Distribution of Inter�Arrival times (Exponential Process)**

Inter�arrival times are defined as the time intervals between two successive arrivals. It will be proved that if the arrival process follows Poisson distribution, an associated random variable defined as the time between successive arrivals (inter-arrival time) follows an exponential distribution f(t) =� λe-λtand vice-versa.

The expected (or mean) time of inter arrival is given by �where λ is the mean arrival rate. Thus, t has the exponential distribution with mean 1/λ, we would intuitively expect that if the mean arrival rate is λ, then the mean time between arrival is 1/λ. Conversely, it can also be independent and have the exponential distribution then the arrival rate follows the Poisson distribution.

**17.7  Classification of Queuing Models (Listing of Four Models)**

The queuing models are classified as follows:

|  |  |  |
| --- | --- | --- |
| Model I | : | (M|M|1): (∞|FCFS) |
| Model II | : | (M|M|s): (∞|FCFS) |
| Model III | : | (M|M|1): (N|FCFS) |
| Model IV | : | (M|M|s): (∞|FCFS) |

In this lesson Model I and II have been discussed as Model III and IV are beyond the scope of this course.

**17.7.1  Model I (single channel queuing model with Poisson arrivals and exponential service times)**

This model is symbolically represented as by (M|M|1): (∞|FCFS). This denotes Poisson arrival (exponential inter-arrival), Poisson departure (exponential service time), single server, Infinite capacity and �First come, First served� service discipline. This model is also called �birth and death model�. This model is one of the most widely used and simplest models. It assumes the following conditions:

(i)           Arrivals are served on a �first come-first served� (FCFS) basis.

(ii)         Every arrival waits to be served regardless of the length of the line.

(iii)       Arrivals are independent of preceding arrivals, but the average number of arrivals does not change over time.

(iv)       Arrivals are described by a Poisson probability distribution and come from an infinite population.

(v)         Service times also vary from one customer to the next and are independent of one another, but their average rate is known.

(vi)       Service times occur according to the negative exponential probability distribution.

(vii)     The average service rate is greater than the average arrival rate.

***17.7.1.1  To obtain the system of steady- state equations***

The probability that there will be n units (n>0) in the system at time (t+∆t) may be expressed as the sum of three independent compound properties by using the fundamental properties of probability, Poisson arrivals, and of exponential service times.

(i) The product of three probabilities (Fig. 17.1)

a)      that there are n units in the system at time t = Pn(t)

b)      that there is no arrival in time ∆t = P0 (∆t) = 1 - λ∆t

c)      that there is no service in time ∆t = φ∆t (0) = 1 - μ∆t is given by





**Fig. 17.1**

****

**Fig. 17.2**

****

**Fig. 17.3**

(ii) The product of three probabilities (Fig. 17.2)

a)      that there are (n-1) units in the system at time t = Pn-1(t)

b)      that there is one arrival in time ∆t = P1 (∆t) = λ∆t

c)      that there is no service in time ∆t = φ∆r(0) = 1 - μ∆t is given by



(iii) The product of probabilities (see Fig. 17.3)

a)      that there are (n+1) units in the system at time t = Pn+1(t)

b)      that there is no arrival in time ∆t = P0 (∆t) = 1 - λ∆t

c)      that there is one service in time ∆t = φ∆t (1) = μ∆t is given by



Now, by adding above three independent compound probabilities, we obtain the probability of n units in the system at time (t + ∆t), i.e.,

����������� Pn (t+∆t) = Pn (t) [1- (λ+ μ) ∆t] + Pn-1 (t) λ∆t + Pn+1 (t) μ∆t +O(∆)� �          (Eq. 17.1)

Where O(∆t) = O1 (∆t) + O2 (∆t) + O3 (∆t)

The equation may be written as

 ���������� 

Now, taking limit as ∆t → 0 on both sides,

����������� 

����������� or

����������� Eq. 17.2)

In  a similar fashion, the    probability that there will be no unit (i.e. n = 0) in   the system at time (t + ∆t) will be the sum of the following two independent probabilities:

����� (i) P[that there is no unit in the system at time t and no arrival in time

������� 

(ii)   P[that there is one unit in the system at time t, one unit serviced in ∆t  and no arrival in time ∆t] = P1(t) μ∆t (1 - λ∆the) ≡ P1 (t) μ∆t + O(∆t)

��� Now, adding these two probabilities we get

����������� P0 (t+∆t) = P0 (t) [1- λ∆t] + P1(t) μ∆t + O(∆)����        (Eq. 17.3)

                          

Now, taking limit as ∆t→0 on both sides,

�����������         (Eq. 17.4)

Consequently the equations (17.2) and (17.4) can be written as

�����������            ������������ (Eq. 17.5)

�����������              (Eq. 17.6)
Equations (17.5) and (17.6) constitute  the system of steady state difference equations for the model.

***17.7.1.2  To solve the system of difference equations***

We solve the difference equations given in (17.5) and (17.6) by the method of successive substitution. Since P0 = P0 and putting n=0 in equation (17.5) we get

����������� � ���(Eq. 17.7)

Now using the fact that 

��������������� 

����������� ������������������ �������� ����������������������������������������������������������������������� ����������� .........(Eq. 17.8)

Now, substituting the value of P0from(17.8) in (17.7), we get

 ���������� �      ���������������������   ��������������������������������������������������������������������  ������������ ����������������������������������������������������������������������������������� ��� (Eq. 17.9)

The equations (17.8) and (17.9) give the probability distribution of queue length

***17.7.1.3  Measure of model 1***

**Expected (average) number of units in system (Ls) is given by**

**�����������**

**�����������**���           ���������������������������������� ������������ ����������������������������������������������������������������������������������� ��� (Eq. 17.10)

**Expected (average) queue length (Lq) is given by:**

Since there are (n-1) units in the queue excluding one being serviced

����������� 
Substituting the value of P0from equation 17.8

����������� ���������������������������������������� ����������������������������������������� ����������������������������������������������������������������������� ���(Eq. 17.11)

**Expected (average) waiting time in the queue (excluding service time) (Wq) is given by:**

**�����������**�                                       ��������������������������������������������������������� ������������������������������������������������� ����������������������������������������������������������������������� ��� (Eq. 17.12)

**Expected (average) waiting time in the system (including service time) (Ws) is given by:**

��������������� �������������������������������������������������������������������������������� �������������������������������������������������� ����������������������������������������������������������������������� ���� (Eq. 17.13)

**Expected (average) length of non-empty queue, (L/L > 0)** **is given by:**

**�����������**
**Expected variance of queue length is given by:**

     �������� 

**Example 1**

Customers arrive at a milk parlour being manned by a single Individual at rate of 25 per hour. The time required to serve a customer has exponential distribution with a mean of 30 per hour. Discuss the various characteristics of the queuing system, assuming that there is only one server.

**Solution**

Arrival rate (λ) = 25 per hour,  Service rate (μ) = 30 per hour

��������������� 

Expected number of units in system (Ls) = 

Expected queue length (Lq) = 

Expected waiting time in the queue �

�**Example 2**

In a service department manned by one server, on an average 8 customers arrive every 5 minutes while the server can serve 10 customers in the same time assuming Poisson distribution for arrival and exponential distribution for service rate. Determine:

a)      Average number of customers in the system.

b)      Average number of customers in the queue.

c)      Average time a customer spends in the system.

d)     Average time a customer waits before being served.

**Solution**

Arrival rate (λ) = 8/5 =1.6 customers per minute.

Service rate (μ) = 10/5 = 2 customers per minute.

��������������� 

a)      Average number of customer in the system.



b)      Average number of customer in the queue.



c)      Average time a customer spends in the system.

     

d)      Average time a customer waits before being served



**17.7.2  Model II (A) general erlang queuing model (Birth-Death Process)**

This model is also represented by (M|M|1): (∞|FCFS), but this is a general model in which the rate of arrival and the service depend on the length n of the line.

***17.7.2.1  To obtain the system of steady state equation***

Let arrival rate λ = λn service rate μ = μn ; [depending upon n]

Then, by the same argument as for equations 17.1 and 17.3

����������� Pn(t+∆t) = Pn(t) [1- (λn + μn) ∆t] + Pn-1(t) λn-1 ∆t + Pn+1(t) μn+1 ∆t + O(∆t), n>0���������������������������� �������� ����������������������������������������������������������������������� ���(Eq. 17.14)

����������� P0(t+∆t) = P0(t) [1- λ0 ∆t] + P1(t) μ1 ∆t + O(∆t), n = 0             ������������������������������������������������ ����������������� ����������������������������������������������������������������������� ���� (Eq. 17.15)

Now dividing equations (17.14) and (17.15) by ∆t, taking limits as ∆t → 0 and following the same procedure as in Model I, obtain

�����������        ������������������������ ������������������������������������������������� ����������������������������������������������������������������������� ����(Eq. 17.16)

����������� ��������������������������������������������������������������������������������������������������� ������ ����������������������������������������������������������������������� ����(Eq. 17.17)

The above written two equations are differential equations which could be solved if a set of initial values P0 (0), P1 (0),�, is given. Such a system of equations can be solved if the time dependent solution is required. But, for many problems it suffices to look at the steady state solution.

In the case of steady state, the boundary conditions are

��������������� 

So the equations (17.16) and (17.17) become,

�����������           ��������������������������������������������������������������������� ����� ����������������������������������������������������������������������������������� ��� (Eq. 17.18)

�����������                 ��������������������������������������������������������������������� ��������������������������������������� ����������������������������������������������������������������������������������� ���� (Eq. 17.19)

The equations (17.18) and (17.19) constitute the system of steady state difference equations for this model.

To solve the system of difference equations.

Since P0 = P0

��������������� 

���������������

����������� 

����������� �        �        �

�����������                                     ����������������������������������������������������������������������� ������������������������������� ����������������������������������������������������������������������������������� ����(Eq. 17.20)

Now, in order to find P0, use the fact that

����������� 

Where��������   ������������������������������������������������������������������������������������������������������������ ������� ����������������������������������������������������������������������������������� ����(Eq. 17.21)

 The result obtained above is a general one and by suitably defining μn and λn many interesting cases could be studied. Now two particular cases may arise:

**Case 1**
**�����������**

In this case, the series S becomes

����������� 

Therefore, from equation 17.21 and 17.20

��������������� 

Here, it is observed that this is exactly the case of Model 1.

**Case 2**
**�����������**

The case, in which the arrival rate λndepends upon n inversely and the rate of service μnis independent of n, is called the case of �Queue with Discouragement�. In this case, the series S becomes

����������� 

The equation 17.21 gives

��������������� 

����������� 

����������� 

����������� ��        �        �

�����������  

It is observed in this case that Pn follows the Poisson distribution, where λ/μ = ρ is constant, however ρ ˃ 1 or ρ < 1 but must be finite. Since the series S is convergent and hence sum able in both the cases.

**17.7.3  Model II (B) : (M**|**M**|**1):(∞** |**SIRO)**

This model is actually the same as Model I, except that the service discipline follows the Service in Random Order (SIOR) rule in place of FCFS rule. Since the derivation of Pn in Model I is independent of any specific queue discipline, so for the SIOR rule also

��������������� 

Consequently, the average number of customers in the customers in the system will be the same whether queue discipline follows SIRO rule or FCFS rule