4. QUEUEING THEORY

One of the most useful areas of application of probability theory is that of queueing theory or the study of waiting line phenomena. Queues are found everywhere in our day-to-day life. For example in industries, schools, hospitals, libraries, banks, post offices, theaters, ticket booking for trains and buses etc. Queues are also common in computer waiting systems, queues of enquiries waiting to be processed by an interactive computer system. Queues of data base requests, queues of input/output requests etc. Queueing problem arises in the following cases.

i) the demand for service is more than the capacity to provide service.
   For example: Ticket booking counters in railway stations, queues are always formed.

ii) the demand for service is less than the capacity to serve so that there is lot of idle facility time or too many facilities.
   For example: In a petrol bunk, if there is no vehicle for refilling petrol then the system is idle, both the pump and the workers are idle.

4.1 Use of Queueing theory

If customers are arriving to service facility in such a way that either the customer or the service facilities have to wait, then we have a queueing problem.

Queueing theory is used to achieve an optimum balance between the cost associated with waiting time of customers and idle time of service facilities so that the profit is maximized.

A customer may be a person, a machine, letter, a ship, a computer job to be processed etc.

4.2 Queueing system or Queueing Model

A queueing system can be described as consisting of customers arriving for service to a service facility, waiting for service, if it is not available immediately, and leaving the service centre

after being serviced.

There are many types of queueing systems. But all of them can be completely described by the following characteristics.

4.3 Characteristics of queueing system

1. **The input pattern or arrival pattern**

   The input pattern represents the manner in which customers arrive for service and join the queueing system. The actual time of arrival of a customer cannot be predicted or observed. The number of arrivals in a time period or the interval between two successive arrivals can not be a constant, but a random variable. Hence the arrival pattern of customers is expressed by means of probability distribution of the number of arrivals per unit time or of inter arrival time. The number of customers arriving per unit time is called the **arrival rate**, which is a random variable.

   The arrival pattern usually used is the Poisson distribution with parameter λ, where λ is the average arrival rate. Then the time interval between consecutive arrivals follow an exponential distribution with mean $\frac{1}{\lambda}$

2. **Service pattern (or departure pattern)**

   The service pattern is represented by the probability distribution of the number of customers serviced per unit of time (i.e. service rate) or the inter-service time. This rate assumes that the service channel to be busy always, that is no idle time is allowed.

   A typical assumption used is that the service time is a random variable following exponential distribution with mean rate of service μ. Sometimes Poisson distribution is also used.

   Note: Exponential distribution is usually used to describe random arrivals or departures because it has memoryless property and so one event does not influence the other.

3. **Service discipline(or Queue discipline**

   Service discipline or order of service is a rule by which customers are selected for service from the queue. The most common discipline is 'first come, first served'(FCFS) or 'first in, first out' (FIFO) according to which the customers are served in the order of their arrival. Example: cinema ticket counters, railway booking counters etc.

   Another discipline is **'last come, first served'(LCFS)** or **'Last in, first out' (LIFO)**.

   Example: In a voting booth a VVIP is chosen as priority to vote as soon as he comes.

   If **service** is given in **random order** then we have a **SIRO** discipline. In this case every customer in the queue has the same probability of being selected for service.

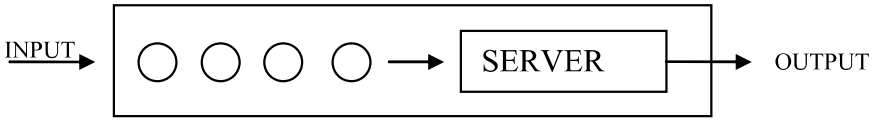

Figure 4.1: Single server model

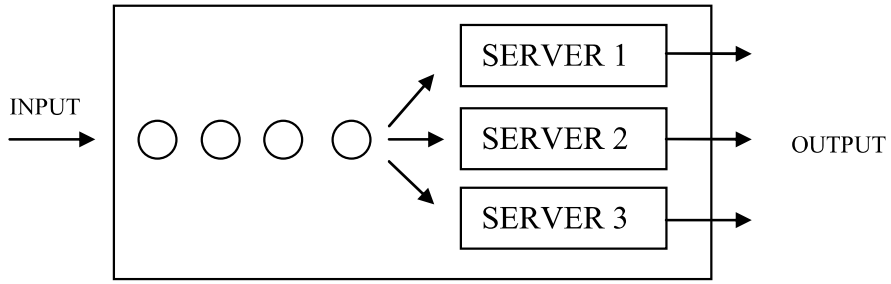

Figure 4.2: Multi server model

4. **Maximum Queueing system Capacity**

Maximum number of customers in the system can be either finite or infinite. In some facilities only limited number of customers are allowed in the system, the new arrivals are not allowed to join the queue. In some system the queue capacity is assumed to be infinite, if every arriving customer is allowed to wait until service is provided.

5. **Calling source or population**

The arrival pattern of the customers depends upon the source from which they come. An infinite source system is easier to describe mathematically than a finite source. In a finite source system, the number of customersin the system affects the arrival rate. For, if every potential customer is already in the queue the arrival rate drops to zero. Whereas for an infinite population system, the number of customers in the queue has no effect on the arrival pattern. So, if the customer population is finite but large, we assume it to be infinite. Infact, in practice if the number of potential customers is over 40 or 50 it is usually said to be infinite.

6. **Customer behaviour**

Generally a customer behaves in the following ways.

(i) **Balking:** A customer who refuses to enter queueing system because the queue is too long is said to be **balking**.

(ii) **Reneging:** A customer who leaves the queue without receiving service because of too much waiting (or due to impatience) is said to have **reneged.**

(iii) **Jockeying:** When there are parallel queues a customer who jumps from one queue to another with shorter length to reduce waiting time is said to be **jockeying**.

7. **State of the System**

In a queueing model, the probability distributions of arrivals, waiting time distribution and service time distribution of customers are in general, functions of time. In the long run it may happen that these characteristics are independent of time. Generally, the states of the queueing system are classified as (i) Transient state, (ii) Steady state and (iii) Explosive state.

Transient state: A queueing system is said to be in transient state when its operating characteristics like input, output, mean queue length etc. are dependent on time. This type of state always occur the beginning of the functioning of the queueing system.

Steady state: A queueing system is said to be in steady state when its operating characteristics are independent of time. If $P_n(t)$ is the probability that there are n customers in the system at time t then the system reaches a steady state if

$$\lim_{t\to\infty} P_n(t) = P_n \text{ or } \lim_{t\to\infty} P_n'(t) = 0$$

Explosive state: If the arrival rate of customers is greater than the service rate, then the queue length will go on increasing with time and as $t \to \infty$, the length queue $\to \infty$. This state is said to be explosive state.

4.3.1 Kendal's notation for representing queueing models

David Kendal introduced a special notation given below to describe a queueing system. The notation has the form

$$(\ a \mid b \mid c) : (\ d \mid e\ )$$

where,

a = arrival (or inter arrival) probability law or distribution.

b = service time probability distribution

c = number of servers (or channels)

d = capacity of the system

e = queue discipline(or service discipline)

Symbols for a and b are the following:

M: Markov(Poisson) arrival and departure distribution or exponential distribution.

E_k: Erlangian(or gamma) inter arrival or service time distribution with parameter k.

G: General service time distribution or general departure distribution.

Symbol for e: FCFS=first come first served.(or)

FIFO=first in first out.

SIRO=service in random order.

We shall now consider some queueing models. We consider here only queueing models under steady state conditions.

The following are usual notations in the discussions.

n = number of customers in the system (i.e., waiting for service in the queue + being served).

λ = mean arrival rate (i.e., average number of customers arriving per unit time)

μ = mean service rate(i.e., average number of customers served per unit time)

ρ = traffic intensity

c = number of parallel service channels

L_q = mean length of the queue (i.e., average(or expected) number of customers waiting in the queue)

L_s = mean length of the system (i.e., average(or expected) number of customers waiting for service in the queue + in service)

W_q = mean waiting time in the queue

W_s = mean waiting time in the system

$P_n(t)$ = transient state probability of exactly n customers in the system

P_n = steady state probability of n customers in the system)

Little's formula

1. $L_s = \lambda.W_s$
2. $W_q = W_s - \dfrac{1}{\mu}$
3. $L_q = \lambda.W_q$
4. $L_s = L_q + \dfrac{\lambda}{\mu}$

$(M\|M\|1):(\infty\|FIFO)$ model	$(M\|M\|1):(K\|FIFO)$ model
Mean arrival time= λ Mean inter(interval) arrival time= $\frac{1}{\lambda}$ Mean service rate= $\frac{1}{\mu}$ $W_s = \frac{1}{\mu-\lambda}$ $W_q = W_s - \frac{1}{\mu}$ $L_q = \lambda W_q$ $L_s = \lambda W_s$	Mean arrival time= λ Mean inter(interval) arrival time= $\frac{1}{\lambda}$ Mean service rate= $\frac{1}{\mu}$ $L_s = \begin{cases} \frac{\left(\frac{\lambda}{\mu}\right)}{\left(1-\left(\frac{\lambda}{\mu}\right)\right)} - (k+1)\frac{\left(\frac{\lambda}{\mu}\right)^{k+1}}{\left(1-\left(\frac{\lambda}{\mu}\right)^{k+1}\right)} & \text{if } \lambda \neq \mu; \\ \frac{k}{2} & \text{if } \lambda = \mu \end{cases}$ Overall Effective arrival rate: $\lambda' = \mu(1-P_0)$ $L_q = L_s - \frac{\lambda'}{\mu}$ $W_q = \frac{L_q}{\lambda'}$ $W_s = \frac{L_s}{\lambda'}$
Probability that the system is empty (or) the system is idle(or) no customers in the system: $P_0 = 1 - \frac{\lambda}{\mu}$ Probability of n customers: $P_n = \left(\frac{\lambda}{\mu}\right)^n P_0$ Probability that an arrival has to wait or system busy or P(channel busy)= $\frac{\lambda}{\mu}$ Probability that there will be more than k customers in the system: $P(N \geq k) = \left(\frac{\lambda}{\mu}\right)^k$ P(waiting time in the system$> t$)=$e^{-(\mu-\lambda)t}$ Probability density function of waiting time in the system is given by $f(w) = (\mu-\lambda)e^{-(\mu-\lambda)w},\ w \geq 0$ Probability density function of waiting time in the queue is given by $g(w) = \frac{\lambda}{\mu}(\mu-\lambda)e^{-(\mu-\lambda)w},\ w \geq 0$	$P_0 = \begin{cases} \frac{\left(1-\frac{\lambda}{\mu}\right)}{1-\left(\frac{\lambda}{\mu}\right)^{k+1}} & \text{if } \lambda \neq \mu; \\ \frac{1}{k+1} & \text{if } \lambda = \mu \end{cases}$ P(Customers turned away)=$P_k = \left(\frac{\lambda}{\mu}\right)^k P_0$

Problem 1 *Customers arrive at a one window drive in bank according to Poisson distribution with mean 10 per hour. Service time per customer is exponential with mean 5 minutes. The space infront of the window including that for a serviced car can accomadate a maximum of 3 cars others can wait outside this space.*

1. *What is the probability that an arriving customer can drive directly to the space in front of the window?*
2. *What is the probability that an arriving customer will have to wait outside the indicated space?*
3. *How long is an arriving customer expected to wait before being served?*

Solution: It is $M|M|1$ model. Given $\lambda = 10$ customer per hour and $\frac{1}{\mu} = 5$ minutes per customer= $\frac{5}{60}$hour= $\frac{1}{12}$hour$\Rightarrow \mu = 12$ customer per hour.

1. An arriving customer can go directly if 0 customer or 1 customer or 2 customers are there in front of the window.

$$\begin{aligned}\therefore \text{Probability} &= P_0 + P_1 + P_2 \\ &= \left[1 - \left(\frac{\lambda}{\mu}\right)\right] + \frac{\lambda}{\mu}\left[1 - \left(\frac{\lambda}{\mu}\right)\right] + \left(\frac{\lambda}{\mu}\right)^2\left[1 - \left(\frac{\lambda}{\mu}\right)\right] \\ &= \left(1 - \frac{\lambda}{\mu}\right)\left[1 + \left(\frac{\lambda}{\mu}\right) + \left(\frac{\lambda}{\mu}\right)^2\right] \\ &= \left(1 - \frac{10}{12}\right)\left[1 + \left(\frac{10}{12}\right) + \left(\frac{10}{12}\right)^2\right] = 0.42\end{aligned}$$

2. Probability that an arriving customer has to wait outside the indicated space$= 1 - 0.42 = 0.58$
3. $W_q = W_s - \frac{1}{\mu} = \frac{1}{12-10} - \frac{1}{12} = \frac{5}{12}hrs= \frac{5}{12} \times 60 = 25$ minutes.

Problem 2 *A super market has a single cashier. During the peak hours, customers arrive at a rate of 20 customers per hour. The average number of customers that can be processed by the cashier is 24 per hour. Find (i) the average number of customers in the queue. (ii) the average number of customers in the system (iii) the average time a customer spends in the system and (iv) in the queue.*

Solution: The super market has only one cashier and so the number of service channels is 1 and any number of customers can enter. So it is $(M|M|1) : (\infty|FIFO)$ model. Given arrival rage $\lambda = 20$ per hour.

Service rate $\mu = 24$ per hour

$\therefore \quad \frac{\lambda}{\mu} = \frac{20}{24} = \frac{5}{6}$ We know $W_s = \frac{1}{\mu - \lambda} = \frac{1}{24-20} = \frac{1}{4}$ hour.

$\therefore L_s = \lambda W_s = 20.\frac{1}{4} = 5$

(i) Average number of customers in the queue is

$$L_q = L_s - \frac{\lambda}{\mu} = 5 - \frac{5}{6} = 4.16$$

(ii) Average number of customers in the system is $L_S = 5$

(iii) Average waiting time in the system is

$$W_s = \frac{1}{4} \text{ hour} = \frac{1}{4} \times 60 \text{ minutes} = 15 \text{ minutes.}$$

(iv) Average waiting time in the queue is

$$W_q = \frac{L_q}{\lambda} = \frac{25}{6} \cdot \frac{1}{20} = \frac{5}{24} \text{ hour} = \frac{5}{24} \times 60 \text{ minutes} = 12.5 \text{ minutes.}$$

Problem 3 *Customers arrive at a one-man barbershop according to a Poisson process with a mean interarrival time of 12 minutes. Customers spend an average of 10 minutes in the barber's chair.*

(i) What is the expected number of customers in the barbershop and in the queue?

(ii) How much time can a customer expect to spend in the barber's shop?

(iii) What is the average time customers spend in the queue?

(iv) What is the percentage of customers who have to wait before getting into the barber's chair?

(v) What is the probability that the waiting time in the system is greater than 30 min?

(vi) What is the probability that more than 3 customers are in the system.

Solution: Since it is a one−man barber shop, it is a $(M|M|1) : (\infty|FIFO)$ model.

Given mean inter arrival time= 12 minutes.

Since it follows exponential, $\frac{1}{\lambda} = 12 \Rightarrow \lambda = \frac{1}{12}$ customer per minute.

Service rate is exponential with mean 10 minutes.

$\Rightarrow \frac{1}{\mu} = 10$ minutes$\Rightarrow \mu = \frac{1}{10}$ customer per minute.

(i) Now, $W_s = \frac{1}{\mu - \lambda} = \frac{1}{\frac{1}{10} - \frac{1}{12}} = \frac{1}{\frac{12-10}{12 \times 10}} = \frac{1}{\frac{2}{120}} = 60$ minutes.

The expected number of customers in the barbershop= $L_s = \lambda W_s = \frac{1}{12} \times 60 = 5$ customers.

and expected number of customers in the queue= $L_q = L_s - \frac{\lambda}{\mu} = 5 - \frac{10}{12} = 4.17$ customers.

(ii) Expected spending time in the barber's shop is $W_s = 60$ minutes.

(iii) The average time customer spend in the queue is $W_q = W_s - \frac{1}{\mu} = 60 - 10 = 50$ minutes.

(iv) Percentage of customers who have to wait before getting into the barber's chair

$$= (1 - P_0) \times 100 = 1 - \left(1 - \frac{\lambda}{\mu}\right) \times 100$$

$$= 1 - \left(1 - \frac{\frac{1}{12}}{\frac{1}{10}}\right) \times 100 = \frac{5}{6} \times 100 = 83.33\%$$

(v) Probability that the waiting time in the system is greater than 30 min

$$= P(W_T > 30) = e^{-(\mu-\lambda)t} = e^{-\left(\frac{1}{10}-\frac{1}{12}\right)30} = e^{-(1/2)} = 0.6065.$$

(vi) Probability that more than 3 customers are in the system

$$= P(N > 3) = P(N \geq 4) = \left(\frac{\lambda}{\mu}\right)^4 = \left(\frac{10}{12}\right)^4 = 0.4823.$$

Problem 4 *Arrivals at a telephone booth are considered to be poisson with an average time 12 minutes between one arrival and the next. The length of telephone call is assumed to be distributed exponentially with mean 4 minutes.*

a) (i) Find the average number of persons waiting in the system.

(ii) What is the probability that a person arriving at the booth has to wait in the queue?

(iii) Also estimate the fraction of the day when phone will be in use.

b) What is the probability that it will take more than 10 minutes for a person to wait and complete his call?

c) The telephone department will install a second booth when convinced that an arrival would expect to wait atleast 3 minutes for the phone. By how much should the flow of arrivals increase in order to justify a second booth?

Solution: It is a $(M|M|1) : (\infty|FIFO)$ model.

Given mean inter arrival time$= 12$ minutes.

$\Rightarrow \frac{1}{\lambda} = 12 \Rightarrow \lambda = \frac{1}{12}$ per minute.

Mean Service time is $\frac{1}{\mu} = 4$ minutes$\Rightarrow \mu = \frac{1}{4}$ per minute.

Now, $W_s = \frac{1}{\mu - \lambda} = \frac{1}{\frac{1}{4} - \frac{1}{12}} = \frac{1}{\frac{12-4}{12\times 4}} = \frac{1}{\frac{8}{48}} = 6.$

a) (i) the average number of persons waiting in the system

$$= L_s = \lambda W_s = \frac{1}{12} \times 6 = 0.5 \text{ person}$$

(ii) Probability that a person arriving at the booth has to wait in the queue

$$\text{P(channel is busy)} = \frac{\lambda}{\mu} = \frac{\frac{1}{12}}{\frac{1}{4}} = \frac{1}{3}$$

(iii) P(phone will be in use)

$$= \frac{\lambda}{\mu} = \frac{\frac{1}{12}}{\frac{1}{4}} = \frac{1}{3}$$

b) Probability that it will take more than 10 minutes for a person to wait and complete his call

$$= P(W_T > 10) = e^{-(\mu-\lambda)t} = e^{-\left(\frac{1}{4}-\frac{1}{12}\right)^{10}} = e^{-(5/3)} = 0.1889.$$

c) The second phone will be installed if $E(W) > 3 \Rightarrow W_q > 3$

$$\text{i.e., if } \frac{\lambda}{\mu(\mu-\lambda)} > 3$$
$$\text{i.e., if } \lambda > 3\mu(\mu-\lambda)$$
$$\text{i.e., if } \lambda > 3\frac{1}{4}(\frac{1}{4}-\lambda)$$
$$\text{i.e., if } \lambda > (\frac{3}{16} - \frac{3\lambda}{4})$$
$$\text{i.e., if } \lambda + \frac{3\lambda}{4} > \frac{3}{16}$$
$$\text{i.e., if } \frac{7\lambda}{4} > \frac{3}{16}$$
$$\text{i.e., if } \lambda > \frac{3 \times 4}{16 \times 7}$$
$$\text{i.e., if } \lambda > \frac{3}{28}$$

Hence the increase in arrival rate should be atleast

$$= \frac{3}{28} - \frac{1}{12} = \frac{1}{42} \text{ per minute.}$$

So, the increase in arrival should be atleast $\frac{1}{42}$ per minute.

Problem 5 *A one-man barber shop can accommodate a maximum of 5 people at a time, 4 waiting and 1 getting hair cut. Customers arrive following poisson distribution with an average of 5 per hour and service is rendered according to exponential distribution at an average rate of 15 minutes,*

(i) What is the probability of idle time?

(ii) Probability of a potential customer turned away.

(iii) Expected number of customer in the queue.

(iv) Expected time spent in the shop.

Solution: Since one-man barber shop with maximum capacity 5, $\therefore$ it is a $(M|M|1) : (k|FIFO)$ model.
Here $k = 5$, $\lambda = 5$ per hour and $\frac{1}{\mu} = 15$ minutes$\Rightarrow \mu = \frac{1}{15} = \frac{60}{15} = 4$ per hour

$$\frac{\lambda}{\mu} = \frac{5}{4} \neq 1$$

(i) Probability of idle time$= P_0 = \dfrac{\left(1 - \frac{\lambda}{\mu}\right)}{1 - \left(\frac{\lambda}{\mu}\right)^{k+1}} = \dfrac{\left(1 - \frac{5}{4}\right)}{1 - \left(\frac{5}{4}\right)^{6}} = \dfrac{0.25}{2.8147} = 0.0888$

(ii) P(a potential customer turned away)$= P(N = 5) = \left(\frac{\lambda}{\mu}\right)^{5} P_0 = \left(\frac{5}{4}\right)^{5} \times 0.0888 = 0.271$

(iii) Expected number of customer in the queue is $L_q = L_s - \dfrac{\lambda'}{\mu}$

where $\dfrac{\lambda'}{\mu} = (1 - P_0) = 1 - 0.0888 = 0.9112$

$$\begin{aligned}
\text{But } L_s &= \frac{\left(\frac{\lambda}{\mu}\right)}{\left(1 - \left(\frac{\lambda}{\mu}\right)\right)} - (k+1)\frac{\left(\frac{\lambda}{\mu}\right)^{k+1}}{\left(1 - \left(\frac{\lambda}{\mu}\right)^{k+1}\right)} \\
&= \frac{\left(\frac{5}{4}\right)}{\left(1 - \left(\frac{5}{4}\right)\right)} - (6)\frac{\left(\frac{5}{4}\right)^{6}}{\left(1 - \left(\frac{5}{4}\right)^{6}\right)} \\
&= -5 + 8.1317 = 3.1317
\end{aligned}$$

$$\therefore\ L_q = L_s - \frac{\lambda'}{\mu} = 3.1317 - 0.9112 = 2.22 \text{ customers}$$

(iv) Expected time spent in the shop is $W_s = \dfrac{L_s}{\lambda'} = \dfrac{L_s}{\mu(1 - P_0)}$

$$\begin{aligned}
&= \frac{3.1317}{4 \times 0.9112} \\
&= 0.8592 \text{ hours} \\
&= 0.8592 \times 60 \text{ minutes} = 51.55 \text{ minutes.}
\end{aligned}$$