10 CHAPTER Queuing Models

The queuing theory or waiting line theory owes its development to A.K. Erlang. He, in 1903, took up the problem on congestion of telephone traffic. The difficulty was that during busy periods, telephone operators were unable to handle the calls the moment they were made, resulting in delayed calls. A.K. Erlang directed his first efforts at finding the delay for one operator and later on the results were extended to find the delay for several operators. The field of telephone traffic was further developed by Molins (1927) and Thornton D-Fry (1928). However, it was only after World War II that this early work was extended to other general problems involving queues or waiting lines.

Waiting lines or queues are omnipresent. Businesses of all types, industries, schools, hospitals, cafeterias, book stores, libraries, banks, post offices, petrol pumps, theatres — all have queuing problems. Queues are also found in industry — in shops where machines wait to be repaired, in tool cribs where mechanics wait to receive tools and in telephone exchanges where incoming calls wait to be handled by the operators. Further examples of queues, though less apparent are: waiting for a telephone operator to answer, a traffic light to change, the morning mail to be delivered and the like.

Waiting line problems arise either because

(*i*) there is too much demand on the facilities so that we say that there is an excess of waiting time or inadequate number of service facilities.

(*ii*) there is too less demand, in which case there is too much idle facility time or too many facilities.

In either case, the problem is to either *schedule arrivals* or *provide proper number of facilities or both* so as to obtain an optimum balance between the costs associated with waiting time and idle time.

Operations research can quite effectively analyse such queuing or congestion phenomena. However, a sound understanding of queuing theory combined with imagination is required to apply the theory to practical situations.

10.1 APPLICATIONS OF QUEUING MODELS

Waiting line or queuing theory has been applied to a wide variety of business situations. All situations where customers are involved such as restaurants, cafeterias, departmental stores, cinema halls, banks, post offices, petrol pumps, airline counters, patients in clinics, etc., are likely to have waiting lines. Generally, the customer expects a certain level of service, whereas the firm providing service facility tries to keep the costs minimum while providing the required service.

Waiting line theory is also widely used by manufacturing units. It has been popularly used in the area of tool cribs. There is a general complaint from the foremen that their workmen wait too long in line for tools and parts. Though the management wants to reduce the overhead charges, engaging more attendants can actually reduce overall manufacturing costs, since the workers will be working instead of standing in line.

Another problem that has been successfully solved by waiting line theory is the determination of the proper number of docks to be constructed for trucks or ships. Since both dock costs and

demurrage costs can be very large, the number of docks should be such that the sum of the two costs is minimized.

Queuing methods have also been used for the problem of machine breakdowns and repairs. There are a number of machines that breakdown individually and at random times. The machines that breakdown form a waiting line for repairs by maintenance personnel and it is required to find the optimum number of repair personnel which makes the sum of the cost of repairmen and the cost of production loss from downtime, a minimum.

Queuing theory has been extended to decide wage incentive plans. For example, some workers are asked to operate, say, two machines while the others, four machines. Since the machines are identical, the base rate of payment is same for all workers. However, the incentive bonus for production in excess of quota is half as much per unit for operators with four machines as for those with two machines. Apparently, the arrangement appears to be fair. However, a study of downtime for repairs shows that while the two machines run by one man would have 12 per cent downtime, four machines run by one man would have 16% downtime. The reason is that two (or more) machines can breakdown at once in the four-machine group which is generally not true for two-machine group. Thus the worker operating four machines would have to operate at a higher efficiency than his counterpart in order to earn the same incentive. The problem was solved by paying the operators of the four-machine group a higher base rate determined by using the probabilities computed from queuing theory.

Queuing theory has also been applied for the solution of problems such as

1. Scheduling of mechanical transport fleets.
2. Scheduling distribution of scarce war material.
3. Scheduling of jobs in production control.
4. Minimization of congestion due to traffic delay at tool booths.
5. Solution of inventory control problems.

10.2 INTRODUCTION

Waiting lines or queues are familiar phenomena, which we observe quite frequently in our daily life. *The basic characteristics of a queuing phenomenon are*

1. Units arrive, at regular or irregular intervals of time, at a given point called the service centre. For example, trucks arriving a loading station, customers entering a department store, persons arriving a cinema hall, ships arriving a port, letters arriving a typist's desk, etc. All these units are called *entries* or *arrivals of customers.*
2. One or more *service channels* or *service stations* or *service facilities* (ticket windows, salesgirls, typists, docks, etc.) are assembled at the service centre. If the service station is empty (free), the arriving customer(s) will be served immediately; if not, the arriving customer(s) will wait in line until the service is provided. Once service has been completed, the customer leaves the system. Whenever we have customers coming to a service facility in such a way that either the customers or the facilities have to wait, we have a queuing problem. Figure. 10.1 shows the major *constituents* of a queuing system (or delay phenomenon). They are

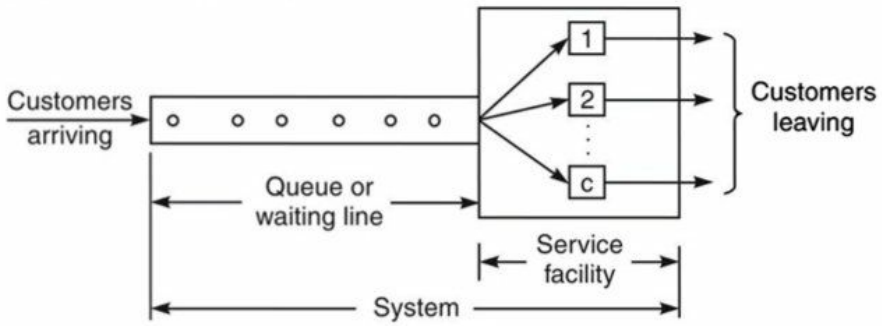

Fig. 10.1

1. **Customer:** The arriving unit that requires some service to be performed. As already described, the customers may be persons, machines, vehicles, parts, etc.
2. **Queue (Waiting line):** The number of customers waiting to be serviced. The queue does not include the customer(s) currently being serviced.
3. **Service Channel:** The process or facility which is performing the services to the customer. This may be single or multi-channel. The number of service channels is denoted by the symbol c.

10.3 ELEMENTS OF A QUEUING SYSTEM (STRUCTURE OF A QUEUING SYSTEM)

A queuing system is specified completely by seven main elements:

1. Input or arrival (inter-arrival) distribution
2. Output or departure (service) distribution
3. Service channels
4. Service discipline
5. Maximum number of customers allowed in the system
6. Calling source or population
7. Customer's behaviour.

1. Arrival Distribution. It represents the *pattern* in which the number of customers arrive at the service facility. Arrivals may also be represented by the *inter-arrival* time, which is the period between two successive arrivals.

Arrivals may be separated by equal intervals of time or by unequal but definitely known intervals of time or by unequal intervals of time whose probabilities are known; these are called random arrivals.

The rate at which customers arrive to be serviced, *i.e.*, number of customers arriving per unit of time is called *arrival rate*. When the arrival rate is random, the customers arrive in no logical pattern or order over time. This represents most cases in the business world.

When arrivals are random, we have to know the probability distribution describing arrivals, specifically *the time between arrivals*. Management scientists have demonstrated that random arrivals are often best described by the Poisson distribution, which was discussed in some details in chapter 8. Of course, arrivals are not always Poisson, and we need to be certain that the assumption of Poisson distribution is appropriate before we use it. Mean value of arrival rate is represented by λ. It may be noted that the Poisson distribution with mean arrival rate λ is equivalent to the (negative) exponential distribution of inter-arrival times with mean inter-arrival time $1/\lambda$.

2. Service (Departure) Distribution. It represents the *pattern* in which the number of customers leave the service facility. Departures may also be represented by the *service* (inter-departure) time, which is the time period between two successive services.

Service time may be constant or variable but known or random (variable with only known probability).

If service times are randomly distributed, we have to find out what probability distribution best describes their behaviour. In many cases where service times are random, management scientists have found that they are best described by the *exponential probability distribution.* If service times are exponentially distributed and arrivals Poisson distributed, the mathematics necessary to study waiting line behaviour is somewhat easier to develop and use. Fig. 10.2 illustrates an exponential probability distribution of service times; from this we find that the probability of long service times is rather small.

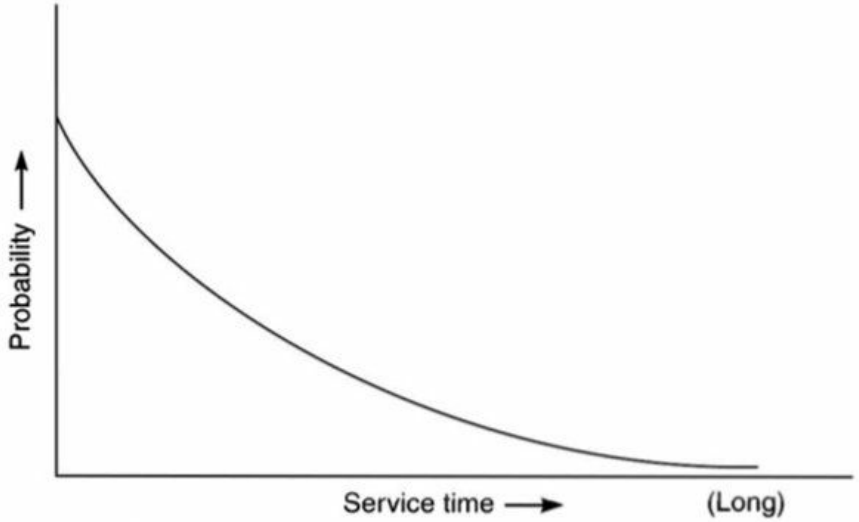

Fig. 10.2 Exponential distribution of service times.

The rate at which one service channel can perform the service, *i.e.,* number of customers served per unit of time is called *service rate.* This rate assumes the service channel to be always busy, *i.e.,* no idle time is allowed. Mean value of service rate is represented by μ. In business problems more cases of uniform service rate will be found than of uniform arrival rates.

3. Service Channels. The queuing system may have a single service channel. Arriving customers may form one line and get serviced, as in a doctor's clinic. The system may have a number of service channels, which may be arranged in parallel or in series or a complex combination of both. In case of parallel channels, several customers may be serviced simultaneously, as in a barber shop. For series channels, a customer must pass successively through all the channels before service is completed, *e.g.,* a product undergoing different processes over different machines or students, during admissions going through one counter after another before all admission formalities are complete. A queuing model is called *one server model,* when the system has one server only and a *multi-server model* when the system has a number of parallel channels each with one server.

Sometimes several service channels may feed into one subsequent service channel; for example, several ticket booths in a theatre may send all the ticket holders to a single ticket collector at the entrance of the theatre. On the other hand, sometimes, a single service channel may disperse customers among several channels that come after it; for example, an enquiry clerk in an office.

4. Service Discipline. Service discipline or order of service is the rule by which customers are selected from the queue for service. The most common discipline is 'first come, first served', according to which the customers are served in the order of their arrival, *e.g.,* cinema ticket windows, railway stations, banks, etc. The other discipline is 'last come, first served', as in a big godown, where the items arriving last are taken out first. Still other disciplines include 'service in random order (SIRO)' and 'priority'. 'Priority' is said to occur when an arriving customer is chosen for service ahead of some other customers already in the queue. A unit (customer) is said to have 'pre-emptive' priority if it not merely goes to the head of the queue but displaces any unit already being served when it arrives. Provided that the order of service is not related to service time, it does not affect the queue length or average waiting time but it does affect the time an individual customer has to wait. The service discipline, therefore, affects the derivation of equations used for analysis. In this text only the most common service discipline 'first come, first served' will be assumed for further discussion.

5. Maximum Number of Customers allowed in the System (Capacity of the System). Maximum number of customers in the system can be either finite or infinite. In some facilities, only a limited number of customers are allowed in the system and new arriving customers are not allowed to join the system unless the number becomes less than the limiting value.

6. Calling Source or Population. The arrival pattern of the customers depends upon the source which generates them. If there are only a few potential customers, the calling source (population) is called finite. If there are a large number of potential customers (say, over 40 or 50), it is usually said to be infinite. There is still another rule for categorising the source as finite or infinite. A finite source exists when an arrival affects the probability of arrival of potential future customers. For example, a battery of M running machines is a finite source, as far as machine repair situation is concerned. Before any machine breaks down, the calling source consists of M potential customers. As soon as a machine breaks down, it becomes a customer and hence cannot generate another 'call' until it gets serviced (repaired). An infinite source is said to exist when the arrival of a customer does not affect the rate of arrival of potential future customers.

7. Customer's behaviour: The customer's behaviour is also very important in the study of queues. If a customer decides not to enter the queue since it is too long, he is said to have *balked.* If a customer enters the queue, but after sometime loses patience and leaves it, he is said to have *reneged.* When there are two or more parallel queues and the customers move from one queue to the other, they are said to be *jockeying.*

10.4 OPERATING CHARACTERISTICS OF A QUEUING SYSTEM

Analysis of a queuing system involves a study of its different operating characteristics. Some of them are

1. *Queue length* (L_q) – the average number of customers in the queue waiting to get service. This excludes the customer(*s*) being served.
2. *System length* (L_s) – the average number of customers in the system including those waiting as well as those being served.
3. *Waiting time in the queue* (W_q) – the average time for which a customer has to wait in the queue to get service.
4. *Total time in the system* (W_s) – 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. *Utilization factor* (ρ) – it is the proportion of time a server actually spends with the customers. It is also called *traffic intensity.*

10.5 WAITING TIME AND IDLE TIME COSTS

In order to solve a queuing problem, service facility must be manipulated so that an optimum balance is obtained between the cost of waiting time and the cost of idle time.

The cost of waiting customers generally includes either the indirect cost of lost business (because people go somewhere else, buy less than they had intended to, or do not come again in future) or direct cost of idle equipment and persons; for example, cost of truck drivers and equipment waiting to be unloaded or cost of operating an airplane or ship waiting to land or dock. The cost of lost business is not easy to assess. For example, vehicle drivers wanting petrol will avoid pumps having long queues. To determine how much business is lost, some type of experimentation and data collection is required.

The cost of idle service facilities is the payment to be made to the servers (engaged at the facilities) for the period for which they remain idle.

By increasing the investment in labour and equipment (service facilities), waiting time and the losses associated with it can be decreased. It is desirable, then, to obtain the minimum sum of these two costs; costs of investment and operation, and costs due to waiting. This optimum balance of costs can be obtained by *scheduling* the *flow* of *units* requiring service and/or *providing proper number of facilities.* If the facilities are not under control, flow of units may be scheduled to minimize the sum of waiting time and idle time costs. If the flow is not subject to control, that amount of equipment and personnel be employed which minimizes the overall costs of operation. If

both can be controlled, one should schedule the input as well as provide facilities which minimize the overall cost.

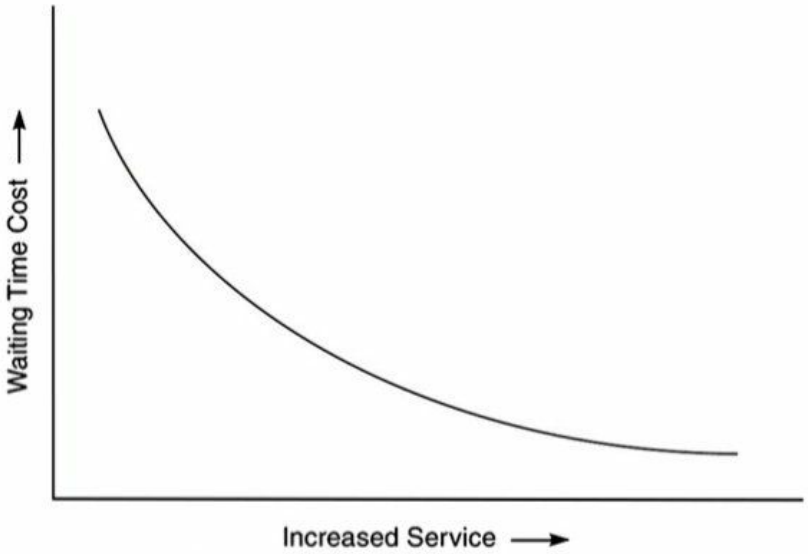

Fig. 10.3 Relationship between level of service and waiting time cost.

Fig. 10.3 illustrates the relationship between the level of service provided and the cost of *waiting time*. It is observed that as the level of service is increased (as more servers are provided), the cost of waiting time decreases.

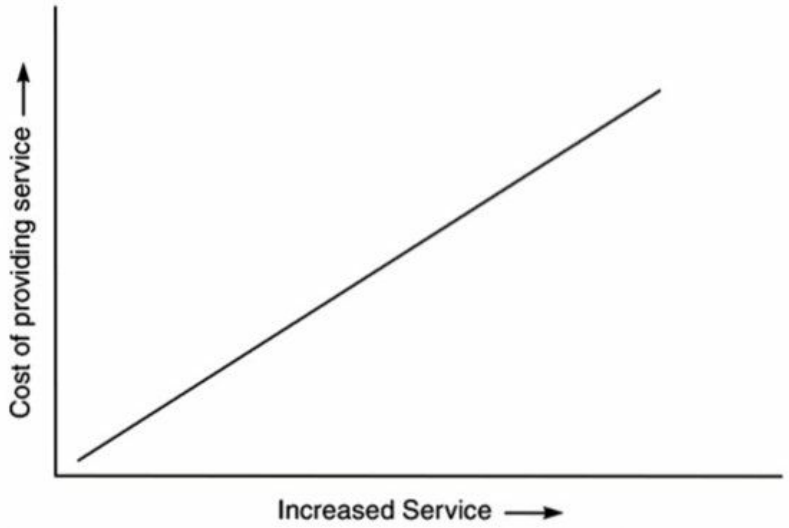

Fig. 10.4 Relationship between level of service and cost of providing service.

In Fig. 10.4 is illustrated the relationship between the level of service and the cost of *providing that* service. It is observed that as the level of service increases, so does the cost of providing that increased service.

In Fig. 10.5 the waiting time cost is added to the cost of providing service to establish a *total expected cost*. We see that the total expected cost is minimum at a service level denoted by point S. Thus the objective of the techniques explained in the remainder of this chapter is really to determine that particular level of service which minimizes the total cost of providing service and waiting for that service.

Let C_w = expected waiting cost/unit/unit time,

L_s = expected (average) number of units in the system,

and C_f = cost of servicing one unit.

Then expected waiting cost per unit time (period) = $C_w \cdot L_s = C_w \cdot \dfrac{\lambda}{\mu - \lambda}$, and expected service cost per unit time (period) = $C_f \cdot \mu$.

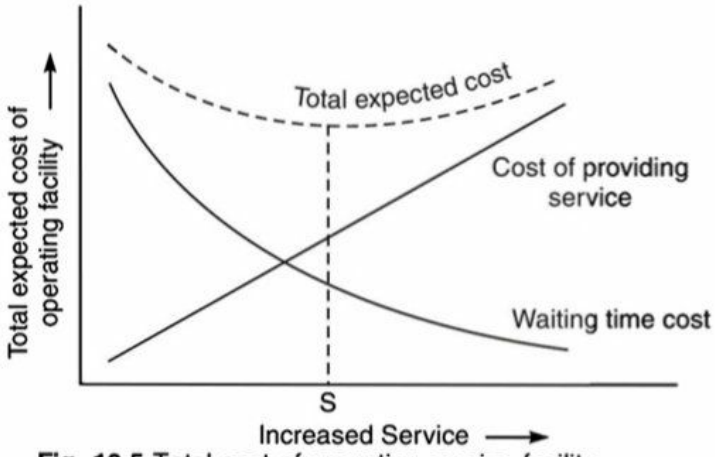

Fig. 10.5 Total cost of operating service facility.

$$\therefore \qquad \text{Total cost, } C = C_w \frac{\lambda}{\mu - \lambda} + \mu C_f$$

This will be minimum if $\frac{d}{d\mu}(C) = 0$

or if $-C_w \cdot \frac{\lambda}{(\mu - \lambda)^2} + C_f = 0$, which gives $\mu = \lambda \pm \sqrt{\frac{C_w}{C_f}}\lambda$.

Note that a plus and minus sign appear before the square root sign. A negative value of μ is not a possible answer in real life problems. μ given by the above equation is called *minimum cost service rate*.

EXAMPLE 10.5-1

Consider a situation in which the mean arrival rate is one customer every 4 minutes and the mean service time is $2\frac{1}{2}$ minutes. If the waiting cost is ₹ 5 per unit per minute and the minimum cost of servicing one unit is ₹ 4, find the minimum cost service rate.

Solution

Here, $\lambda = \frac{1}{4} = 0.25.$

$$\mu = \lambda \pm \sqrt{\frac{C_w}{C_f}.\lambda} = 0.25 \pm \sqrt{\frac{5 \times 0.25}{4}} = 0.25 \pm \sqrt{0.3125} = 0.25 \pm 0.56.$$

$\therefore$ $\mu = 0.81$ units/minute. ($\because$ $\mu = -0.32$ is not a feasible solution.)

10.6 TRANSIENT AND STEADY STATES OF THE SYSTEM

Queuing theory analysis involves the study of system's behaviour over time. If the operating characteristics (behaviour of the system) vary with time, it is said to be in *transient state*. Usually a system is transient during the early stages of its operation, when its behaviour still depends upon the initial conditions (number of customers in the system) and the elapsed time. However, it is the 'long-run' behaviour or the *steady state condition* of the system which is more important. A system is said to be in steady state condition if its behaviour becomes independent of its initial conditions and of the elapsed time.

An essential condition for reaching a steady state is that the total elapsed time since the start of the operation must be sufficiently large (theoretically, it should tend to infinity). However, this is not the sufficient condition as the parameters of the system also affect its state e.g., number of customers at the counter of a post office within 15 minutes of its opening.

For example, if the average arrival rate is less than average service rate and both are constant, the system eventually settles down to a steady state and the probability of finding a particular length of queue will be same at any time. If the rates are not constant, the system will not reach a steady state, but it could remain stable. If the arrival rate is greater than service rate, the system cannot attain a steady state (regardless of the length of elapsed time); it is rather unstable, queue length increases steadily with time and theoretically, it could build up to infinity. Such state of the system is called *explosive state*. Evidently, imposing a limit on the maximum length of the queue (so that further arrivals are not accepted) automatically ensures stability. Queuing situations which are unstable for a limited time are common in practice– rush-hour traffic is an example. In this text we shall consider the steady state analysis; transient and explosive states require complex mathematical tools for analysis and will not be touched upon.

10.7 KENDALL'S NOTATION FOR REPRESENTING QUEUING MODELS

D.G. Kendall (1953) and later A. Lee (1966) introduced 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 symbols *a, b* and *d:*

Symbols for a and b

M = Markovian (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 represent a finite (N) or infinite (∞) number of customers in the system and calling source respectively. For instance, $(M/E_k/1) : (FCFS/N/\infty)$ 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.

10.8 CLASSIFICATION OF QUEUING MODELS

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

(a) 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

predicts the number of arrivals in a given time. The Poisson distribution involves the probability of occurrence of an arrival. Poisson assumption is quite restrictive in some cases. It assumes that arrivals are random and independent of all other operating conditions. The mean arrival rate (*i.e.*, the number of arrivals per unit of time) λ is assumed to be constant over time and is independent of the number of units already serviced, queue length or any other random property of the queue.

Since the mean arrival rate is constant over time, it follows that the probability of an arrival between time t and $t + dt$ is $\lambda \, . \, dt$.

Thus probability of an arrival in time $dt = \lambda \, dt$. ...(10.1)

The following characteristics of Poisson distribution are written here without proof :

Probability of n arrivals in time $t = \dfrac{(\lambda t)^n \, . \, e^{-\lambda t}}{n!}$, $n = 0, 1, 2, \ldots\ldots$, ...(10.2)*

Probability density function of inter-arrival time (time interval between two consecutive arrivals)

$$= \lambda \, . \, e^{-\lambda t}.$$ ...(10.3)

Finally, Poisson distribution assumes that the time period dt is very small so that $(dt)^2$, $(dt)^3$, etc. $\rightarrow 0$ and can be ignored.

Service time is the time required for completion of a service *i.e.*, it is the time interval between beginning of a service and its completion. The mean service rate is the number of customers served per unit of time (assuming the service to be continuous throughout the entire time unit), while the average service time $1/\mu$ is the time required to serve one customer. The most common type of distribution used for service times is exponential distribution. It involves the probability of completion of a service. It should be noted that Poisson distribution cannot be applied to servicing because of the possibility of the service facility remaining idle for some time. Poisson distribution assumes fixed time interval of continuous servicing, which can never be assured in all services.

Mean service rate μ is also assumed to be constant over time and independent of number of units already serviced, queue length or any other random property of the system. Thus probability that a service is completed between t and $t + dt$, provided that the service is continuous

$$= \mu dt.$$

Under the condition of continuous service, the following characteristics of exponential distribution are written, without proof :

Probability of n complete services in time $t = \dfrac{(\mu t)^n \, . \, e^{-\mu t}}{n!}$. ...(10.4)

Probability density function (*p.d.f*) of interservice time, *i.e.*, time between two consecutive services $= \mu \, . \, e^{-\mu t}$. ...(10.5)

Probability that a customer shall be serviced in more than time $t = e^{-\mu t}$. ...(10.6)

EXAMPLE 10.9-1.1

On an average, 6 customers reach a telephone booth every hour to make calls. Determine the probability that exactly 4 customers will reach in 30-minute period, assuming that arrivals follow Poisson distribution.

Solution

Here $\lambda = 6$ customers/hour,

$t = 30$ minutes $= 0.5$ hour,

$n = 4$,

$\lambda t = 6 \times 0.5 = 3$ customers.

$\therefore$ Probability of 4 customers arriving in 0.5 hour

*See appendix B.1 for derivation of the expression.

A self-service store employs one cashier at its counter. Nine customers arrive on an average every 5 minutes while the cashier can serve 10 customers in 5 minutes. Assuming Poisson distribution for arrival rate and exponential distribution for service time, find

1. *Average number of customers in the system.*
2. *Average number of customers in the queue or average queue length.*
3. *Average time a customer spends in the system.*
4. *Average time a customer waits before being served.*

[*J.N.T.U. Hyderabad B.Tech. (Mech.) May, 2012; May, 2011; P.T.U. B.E., 2001; Karn. U. B.E. (Mech.) 1998, 95*]

Solution

Arrival rate $\lambda = 9/5 = 1.8$ customers/minute,
service rate $\mu = 10/5 = 2$ customers/minute.

1. Average number of customers in the system,

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{1.8}{2 - 1.8} = 9.$$

2. Average number of customers in the queue,

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{\lambda}{\mu} \cdot \frac{\lambda}{(\mu - \lambda)} = \frac{1.8}{2} \times \frac{1.8}{2 - 1.8} = 8.1.$$

3. Average time a customer spends in the system,

$$W_s = \frac{1}{\mu - \lambda} = \frac{1}{2 - 1.8} = 5 \text{ minutes.}$$

4. Average time a costumer waits in the queue,

$$W_q = \frac{\lambda}{\mu}\left(\frac{1}{\mu - \lambda}\right) = \frac{1.8}{2}\left(\frac{1}{2 - 1.8}\right) = 4.5 \text{ minutes.}$$

EXAMPLE 10.9-4.2

A person repairing radios finds that the time spent on the radio sets has exponential distribution with mean 20 minutes. If the radios are repaired in the order in which they come in and their arrival is approximately Poisson with an average rate of 15 for 8-hour day, what is the repairman's expected idle time each day ? How many jobs are ahead of the average set just brought in ?

[P.U.B.E. (T.&I.T.) Nov., 2004; B.E. (Mech.) 2002 ; P.T.U. B. (Tech.) 2010; 2000; MBA May, 2002 ; IGNOU MBA, 2000; G.J.U. B.E. (Mech.) 1996]

Solution

$$\text{Arrival rate } \lambda = \frac{15}{8 \times 60} = \frac{1}{32} \text{ units/minute,}$$

$$\text{service rate } \mu = \frac{}{20} \text{ units/minute.}$$

Number of jobs ahead of the set brought in = Average number of jobs in the system,

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{1/32}{1/20 - 1/32} = \frac{5}{3}.$$

Number of hours for which the repairman remains busy in an 8-hour day

$$= 8\,\frac{\lambda}{\mu} = 8 \times \frac{1/32}{1/20} = 8 \times \frac{20}{32} = 5 \text{ hours.}$$

$\therefore$ Time for which repairman remains idle in an 8-hour day

$$= 8 - 5 = 3 \text{ hours.}$$

EXAMPLE 10.9-4.3

A branch of Punjab National Bank has only one typist. Since the typing work varies in length (number of pages to be typed), the typing rate is randomly distributed approximating a Poisson distribution with mean service rate of 8 letters per hour. The letters arrive at a rate of 5 per hour during the entire 8-hour work day. If the typewriter is valued at ₹ 1.50 per hour, determine.

1. *Equipment utilization.*
2. *The per cent time that an arriving letter has to wait.*
3. *Average system time.*
4. *Average cost due to waiting on the part of typewriter i.e., it remaining idle.*

[H.P.U. B. Tech. (Mech.) Nov., 2007; Nagpur U.B.E. (Mech.) 2003]

Solution

Arrival rate, $\lambda = 5$ per hour,

service rate, $\mu = 8$ per hour.

1. Equipment utilization, $\rho = \frac{\lambda}{\mu} = \frac{5}{8} = 0.625.$
2. The per cent time an arriving letter has to wait

   = per cent time the typewriter remains busy

   = 62.5%.

3. Average system time,

$$W_s = \frac{1}{\mu - \lambda} = \frac{1}{8-5} = \frac{1}{3} \text{ hr.} = 20 \text{ minutes.}$$

4. Average cost due to waiting on the part of the typewriter per day
$= 8 \times (1 - 5/8) \times ₹\ 1.50 = ₹\ 4.50.$

EXAMPLE 10.9-4.4

The milk plant at a city distributes its products by trucks, loaded at the loading dock. It has its own fleet of trucks plus trucks of a private transport company. This transport company has complained that sometime its trucks have to wait in line and thus the company loses money paid for a truck and driver that is only waiting. The company has asked the milk plant management either to go in for a second loading dock or discount prices equivalent to the waiting time. The following data are available:

Average arrival rate (all trucks) = 3 per hour,
average service rate = 4 per hour.

The transport company has provided 40% of the total number of trucks. Assuming that these rates are random according to Poisson distribution, determine

1. *The probability that a truck has to wait.*
2. *The waiting time of a truck that waits.*
3. *The expected waiting time of company trucks per day.*

[*P.U.B.E. (Mech.) Nov., 2002; Kuru. U. M.Tech. May, 1988*]

Solution

1. The probability that a truck has to wait for service = utilization factor $= \rho = \frac{\lambda}{\mu} = 3/4 = 0.75.$
2. The waiting time of a truck that waits

$$= W_n = \frac{1}{\mu - \lambda} = \frac{1}{4-3} = \frac{1}{1} = 1 \text{ hour.}$$

3. Total expected waiting time of company trucks per day = Trucks/day × % of company trucks × expected waiting time per truck

$$= (3 \times 8) \times (0.40) \times \frac{\lambda}{\mu(\mu - \lambda)}$$

$$= 24 \times 0.40 \times \frac{3}{4(4-3)} = 7.2 \text{ hours/day.}$$

EXAMPLE 10.9-4.5

Arrival rate of telephone calls at a telephone booth are according to Poisson distribution, with an average time of 9 minutes between two consecutive arrivals. The length of telephone call is assumed to be exponentially distributed, with mean 3 minutes.

(a) Determine the probability that a person arriving at the booth will have to wait.
(b) Find the average queue length that is formed from time to time.
(c) The telephone company will install a second booth when convinced that an arrival would expect to have to wait at least four minutes for the phone. Find the increase in flow rate of arrivals which will justify a second booth.
(d) What is the probability that an arrival will have to wait for more than 10 minutes before the phone is free ?
(e) What is the probability that he will have to wait for more than 10 minutes before the phone is available and the call is also complete ?