Queuing Theory

Introduction:

A group of items waiting to receive service, Including those receiving the service is known as a waiting line or queue. Queuing theory involves the mathematical study of **“queues”** or **“waiting lines”.**

The formation of queue or waiting line is a common phenomenon whenever the current demand for a service exceeds the current capacity to provide that service.

The queues of people may be seen at the cinema ticket window, bus – stop, reservation office , customers at super market etc.

The person waiting in the queue and waiting for service is known as the **customer** and the person by whom he is serves is called the **server**.

The customers arriving for service may form one queue and be serviced through only one station, as in a doctor’s clinic, they may form one queue an served through several stations, as the check out counters of a super market.

The Basic Queuing Process (system) and its Characteristics:

The basic queuing process can be described as a process in which the customer arrive for service at a service counter (or station), wait for their turn in the queue if the server is busy in the service of other customer and served when the server gets free. Finally the customer leave the system as soon as he served.

Characteristics of Queuing System:

The characteristics of a queuing system are as follows:

a. The input (or arrival pattern)
b. Queue (or waiting line)
c. The service discipline (or queue discipline)
d. The service mechanism (or service pattern)

The Input or Arrival Pattern:

The input describes the pattern in which the customers arrive for service. Since the units for service (customers) arrive in a random fashion therefore, their arrival pattern can be describe in terms of probabilities. Here we assume that they arrive according to a Poisson Process i.e., the number of units arriving until any specific time has a poisson distribution. This is the case where arrivals to the queuing system occur at random, but at a certain average rate.

Queue (or Waiting Line):

The units requiring service enter the queuing system on their arrival and join the "queue" which is characterized by the maximum permissible number of units that it can contain called the **capacity of the system**. A queue is called finite if the number of units in it is **finite** otherwise it is called **infinite**.

The Service Discipline:

The service discipline refers to the manner in which the members in the queue are chosen for service. The following service discipline are seen in common practice.

First Come First Serve (FCFS):

According to this discipline the customers are served in the order of their arrival. This service discipline may be seen at a cinema ticket window, at a railway ticket window etc. Here we will deal with the queuing models in which the service discipline is FCFS.

Last Come First Serve (LCFS):

This is discipline may be seen in big godowns where the units (items) which come last are taken out (served) first.

Service in Random Order:

Service On Some Priority – Procedure:

Some customers are served before the others without considering their order of arrival i.e., some customers are served on priority basis.

The Service Mechanism (or service pattern):

The service mechanism refers to.

(i) The pattern according to which the customers are served#
(ii) Facilities given to the customers
(iii) Single Channel
(iv) Multi - Channel

Customers Behaviour in a Queue:

(i) Reneging:

A customer may leave the queue due to impatience.

(ii) Balking:

A customer may not like to wait in the queue due to lack of time space or time or otherwise.

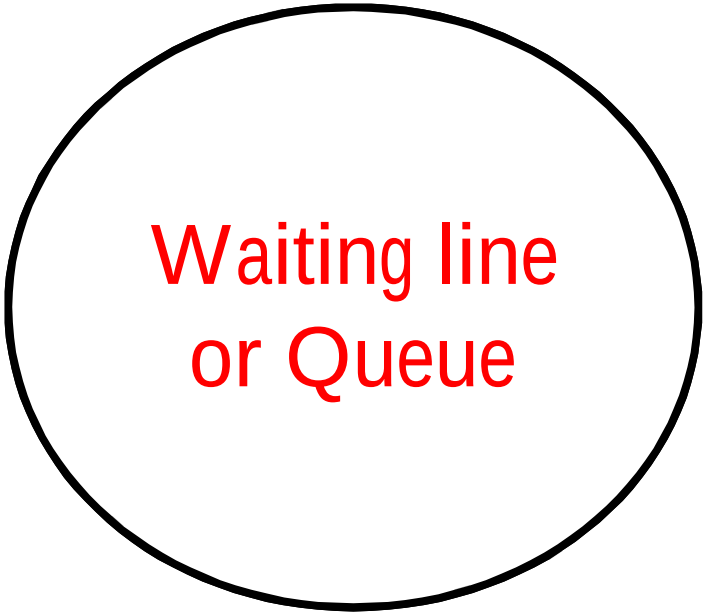
(iii) Collusion:

Some customers may collaborate and only one of them may join the queue.

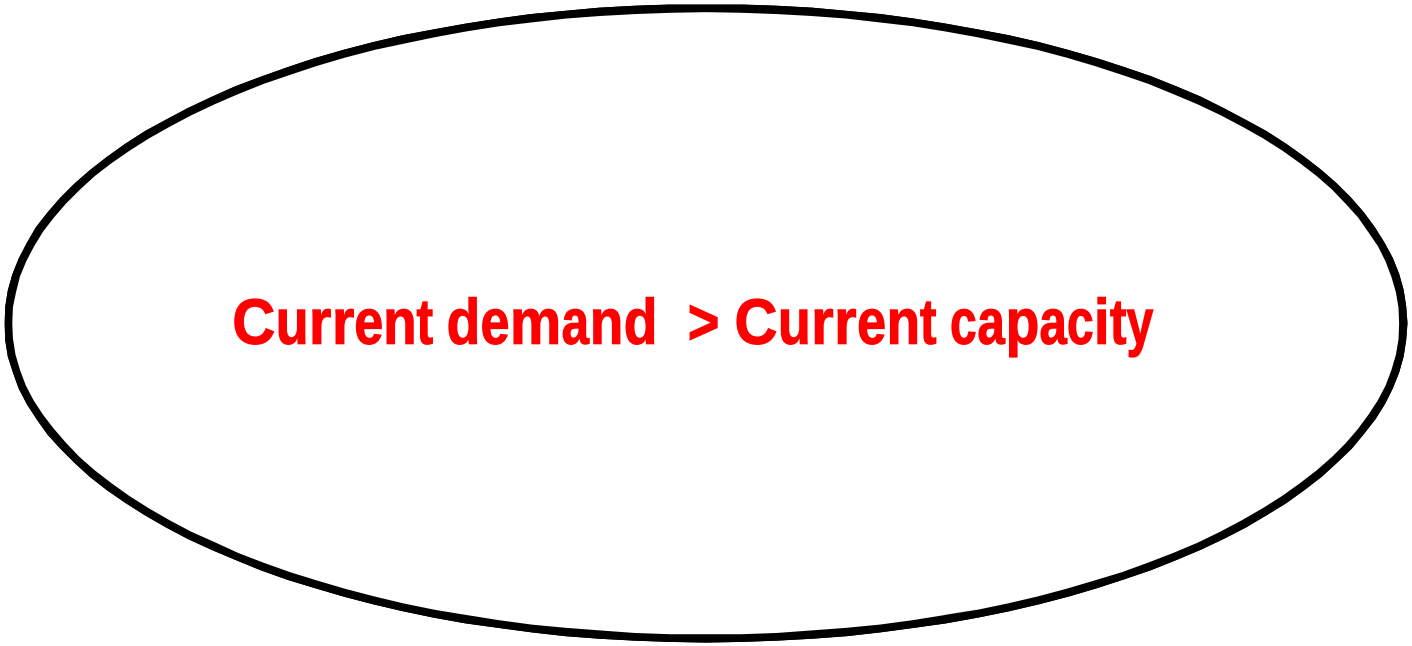
(iv) Jockeying:

If there are more than one queue customer may leave the one queue and join other.

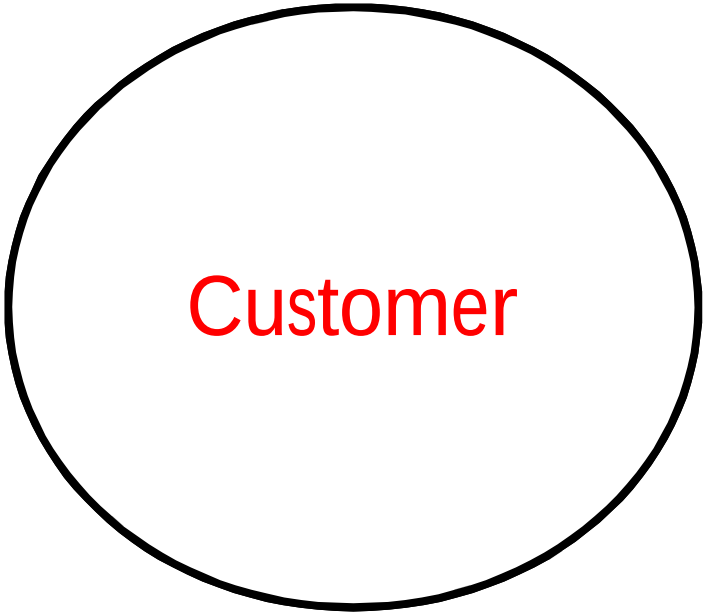
A group of items waiting to receive service, Including those receiving the service is known as;



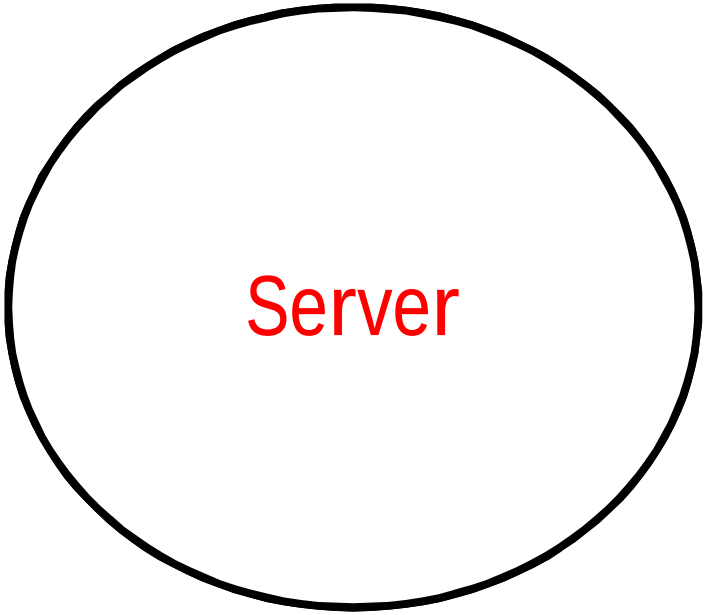
The formation of queue or waiting line is a common phenomenon whenever;



The person waiting in the queue and waiting for service is known as;



The person by whom customer is served is called;



What are the characteristics of a queuing system?

The input (or arrival pattern)

Queue (or waiting line)

The service discipline (or queue discipline)

The service mechanism (or service pattern)

The input describes in a queuing pattern?

The pattern in which the customers arrive for service

Customer's arrival pattern can be describe in terms of?

Poisson Process

Capacity of the queuing system can be defined as;

The maximum permissible number of units that system can contain

Some customers may collaborate and only one of them may join the queue. This behaviour is known as;

Collusion

If there are more than one queue customer may leave the one queue and join other. This behaviour is known as;

Jockeying

Some Important Definitions:

(i) Queue Length:

Queue length is defined by the number of persons (customers) waiting in a line at any point of time.

(ii) Average Length of a line:

Average length of a line (or queue) is defined by the number of customers in the queue per unit time.

(iii) Waiting Time:

It is the time up to which a unit has to wait in the queue before it is taken in to service.

(iv) Servicing Time:

The time taken for servicing of a unit is called its servicing time.

(v) Busy Period:

Busy – period of a server in the time during which he remains busy in servicing. Thus this is the time between the start of a service of the first unit to the end of the service of the last unit in the queue.

(vi) Idle Period:

When all the units in the queue are served. The idle period of the server begins and it continues up – to the time of arrival of the unit (customer). The ideal period of a server is the time during which he remains free because there is no customer present in the system.

(vii) Mean Arrival Rate:

The mean arrival rate in a waiting – line situation is defined as the expected number of arrivals occurring in a time interval of length unity.

(viii) Mean servicing Rate:

The mean servicing rate for a particular servicing station is defined as the expected number of services completed in a time interval of length unity, given that the servicing is going on throughout the entire time unit.

(ix) Traffic Intensity:

In case of a simple queue the traffic intensity is the ratio of mean arrival rate and the mean servicing rate.

$$\text{Traffic Intensity} = \frac{\text{Mean arrival rate}}{\text{Mean servicing rate}} \qquad \rho = \frac{\lambda}{\mu}$$

The number of persons (customers) waiting in a line at any point of time defines;

Queue Length

The time up to which a unit has to wait in the queue before it is taken in to service is known as;

Waiting Time

The mean arrival rate in a waiting – line situation is defined as the expected number of arrivals occurring in a time interval of length unity.

Mean Arrival Rate

Traffic Intensity is represented by;

$$\text{Traffic Intensity} = \frac{\text{Mean arrival rate}}{\text{Mean servicing rate}}$$

The State of The System:

A basic concept in the analysis of the queuing theory is that of a state of the system it involves the study of a system behaviour over a time. The state of the system may be classified as follows:

(i) Transient State:

A system is said to be in transient state when its operating characteristics are dependent on time. Thus a queuing system is in transient state when the probability distribution of arrivals, waiting time and servicing time of the customer dependent on time. This state occurs at the beginning of the operation of the system:

(i) Steady State:

A system is said to be in steady state when its operating characteristics becomes independent of time. Thus a system is said to be in steady state when the probability distribution of arrivals, waiting time and servicing time of the customers are independent of the time. This state occurs in the long run of the system.

Let $P_n(t)$ denote the probability that there are n units in the system at time t, then the system acquires steady state as t tends to ∞ if :

$\lim_{n\to\infty} P_n(t) = P_n$ (Independent of time t)

In most of the queuing problems, steady – state solutions exist independent of the initial state of the queue.

(i) Explosive State:

If the arrival rate of the system in more than its servicing rate, the length of the queue will go on increasing with the time and will tend to infinity as t tends to infinity. This state of the system is said to be explosive state.

If the arrival rate of the system in more than its servicing rate the system is in;

Steady State

Transient State

Explosive State

The Poisson Process:

In Poisson process the probability of n arrivals during time intervals of length t is given by

$$P_n(t) = \frac{(\lambda t)^n . e^{-\lambda t}}{n!} \text{.....(1)}$$

Where λ is a parameter.

Case (i): When, n = 0

$$P_0(\Delta t) = \frac{(\lambda \Delta t)^0 . e^{-\lambda \Delta t}}{0!} = e^{-\lambda \Delta t}$$

$$= 1 - \lambda \Delta t + \frac{\lambda^2}{2!}(\Delta t)^2 - \ldots$$

$$= 1 - \lambda \Delta t + 0(\Delta t)$$

Where 0 (Δt) denotes a quantity which is of smaller order of magnitude then Δt.
If Δt is very small then 0 (Δt) = 0.

$P_0(\Delta t) = 1 - \lambda \Delta t$
i.e., the probability of no arrival in time $\Delta t = 1 - \lambda \Delta t$

Case (2): When, n = 1

$$P_1(\Delta t) = \frac{(\lambda \Delta t)^1 . e^{-\lambda \Delta t}}{1!}$$

$$= \lambda \Delta t \left\{ 1 - \lambda \Delta t + \frac{\lambda^2}{2!}(\Delta t)^2 - \ldots \right\}$$

$$= \lambda \Delta t + 0(\Delta t)$$

$$= \lambda \Delta t \qquad \text{(if } \Delta t \text{ is very small)}$$

i.e., the probability of one arrival in time $\Delta t = \lambda \Delta t$

Case (3): When, n = m > 1

$$P_m(\Delta t) = \frac{(\lambda \Delta t)^m . e^{-\lambda \Delta t}}{m!}$$

$$= \frac{\lambda^m (\Delta t)}{m!} \left\{ 1 - \lambda \Delta t + \frac{\lambda^2}{2!}(\Delta t)^2 - \ldots \right\}$$

$$= \frac{\lambda^m}{m!} \left\{ (\Delta t)^m - \lambda (\Delta t)^{m+1} ..... \right\}$$

$$= 0$$

i.e., the probability of more than one arrival in time $\Delta t = 0$.

From the above discussion we arrive to the following result which is sometimes is known as postulates of Poisson Process.

Postulates for the Poisson process:

Postulate (1):

The number of arrivals in non – overlapping intervals are statistically independent , i.e., the process has independent increments.

Postulate (2):

The probability that an arrival occurs between time t and time t + Δt is equal to $\lambda\ \Delta t + 0\ (\Delta t)$.

i.e., $P_1\ (\Delta t) = \lambda\ \Delta t + 0(\Delta t)$

Where λ is a constant independent of $P_n(t)$, Δt is small increment in t and $0\ (\Delta t)$ denotes the quantity which is of smaller order of magnitude then Δt s.t.,

$$\lim_{\Delta t \to 0} \frac{0(\Delta t)}{\Delta t} = 0$$

Postulate (3):

The probability that of more than one arrival between time t and t + Δt is $0\ (\Delta t)$, i.e., the probability of two or more arrivals during the small time interval Δt is negligible.

Thus $P_0\ (\Delta t) + P_1\ (\Delta t) + 0\ (\Delta t) = 1$

A queue model is generally represented by (A/B/C): (d/e)

Where,

A = Arrival pattern
B = Probability distribution of servicing time
C = The number of servicing channels
d = capacity of the system
e = the manner in which the customers are served

Solution of Queue Models:

Step 1: First of all we shall obtain the system of steady state equations governing the queue.
Step 2: In step two we solve the equations obtained in step one and find the probability distribution of queue length.
Step 3: Then we obtain various probability density functions and derive the formulae.

Model I : (M/ M/1) : (∞/FCFS)

This is a queuing model with Poisson arrival, Poisson service, single servicing channel, with infinite capacity. The service discipline is first come first serve. Here λ = mean arrival units and μ = mean service rate.

To find the steady state equations for the system:

There are n > 0 units in the system at time (t + Δ t) in the following ways
(i) (n – 1) units in the system at time t with probability $P_{n-1}(t)$
One arrival in time Δ t with the probability $P_1(\Delta t) = \lambda \Delta t$
No service in time Δ t with the probability $\phi_0(\Delta t) = 1 - \mu \Delta t$

The probability in this case is $P_{n-1}(t).\ \lambda \Delta t.\ 1 - \mu \Delta t$

(ii) 'n' units in the system at time t with probability $P_n(t)$
No arrival in time Δ t with the probability $P_0(\Delta t) = 1 - \lambda \Delta t$
No service in time Δ t with the probability $\phi_0(\Delta t) = 1 - \mu \Delta t$
Probability in this case is $P_n(t).\ (1 - \lambda \Delta t).\ (1 - \mu \Delta t)$

(iii) (n + 1) units in the system at time t with probability $P_{n+1}(t)$
No arrival in time Δ t with the probability $P_0(\Delta t) = 1 - \lambda \Delta t$
One service in time Δ t with the probability $\phi_1(\Delta t) = \mu \Delta t$
The probability in this case is $P_{n+1}(t).\ (1 - \lambda \Delta t).\ \mu \Delta t$

Since all the probabilities in the above case are mutually exclusive therefore
$P_n(t + \Delta t)$, the probability of n units in the system at time $(t + \Delta t)$ is obtained by adding the probabilities in the above three cases.
i.e.

$$P_n(t + \Delta t) = [P_{n-1}(t).\ \lambda \Delta t.\ 1 - \mu \Delta t + P_n(t).\ (1 - \lambda \Delta t).\ (1 - \mu \Delta t) + P_{n+1}(t).\ (1 - \lambda \Delta t).\ \mu \Delta t]$$

$$P_n(t + \Delta t) = [\lambda P_{n-1}(t) - (\lambda + \mu) P_n(t) + \mu P_{n+1}(t)] \Delta t + P_n(t) + 0(\Delta t)$$

$$\therefore \frac{P_n(t + \Delta t) - P_n(t)}{\Delta t} = \lambda P_{n-1}(t) - (\lambda + \mu) P_n(t) + \mu P_{n+1}(t) + \frac{0(\Delta t)}{\Delta t} \quad \text{.....(2)}$$

Similarly there is no unit in the system at time $(t + \Delta t)$ in the following two ways

(i) No unit in the system at time t, with the probability $P_0(t)$

No arrival in time Δt with the probability $P_0(\Delta t) = 1 - \lambda \Delta t$

The total probability in this case is given by $P_0(t).\ (1 - \lambda \Delta t)$

(ii) One unit in the system at time t, with probability $P_1(t)$

No arrival in time Δt with the probability $P_0(\Delta t) = 1 - \lambda \Delta t$

One service in time Δt with the probability $\phi_1(\Delta t) = \mu \Delta t$

The total probability in this case is given by $P_1(t).\ (1 - \lambda \Delta t).\ \mu \Delta t$

Adding the probabilities in above two cases, we get $P_0(t + \Delta t)$, the probability of no unit in the system in the time $(t + \Delta t)$

i.e. $P_0(t + \Delta t) = P_0(t).\ (1 - \lambda \Delta t) + P_1(t).\ (1 - \lambda \Delta t).\ \mu \Delta t$

Or, $P_0(t + \Delta t) = P_0(t).\ [- \lambda P_0(t) + \mu P_1(t)].\Delta t + 0\ (\Delta t)$

$$\therefore \quad \frac{P_0(t+\Delta t) - P_0(t)}{\Delta t} = -\lambda P_0(t) + \mu P_1(t) + \frac{0(\Delta t)}{\Delta t}$$

$as \ , \Delta t \rightarrow 0$

$$\frac{d}{dt} P_n(t) = \lambda P_{n-1}(t) - (\lambda + \mu) P_n(t) + \mu P_{n+1}(t) \text{.....} \quad \text{.....}(3)$$

$$\frac{d}{dt} P_0(t) = -\lambda P_0(t) + \mu P_1(t) \text{.....} \quad \text{.....} \quad .(4)$$

Now when the system is in steady state,

$$\lim_{n \to \infty} P_n(t) = P_n ,$$

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

Under steady state of the system, the equations (3) and 4 reduces to,

$$\lambda P_{n-1} - (\lambda + \mu) P_n + \mu P_{n+1} = 0 \text{.....}(5)$$

$and,$

$$-\lambda P_0 + \mu P_1 = 0 \text{.....}(6)$$

The equations (5) and (6) are the steady state equations of the system.

To solve the equations obtained in above step:
From (6),

$$P_1 = \frac{\lambda}{\mu} P_0 = \rho P_0 , \frac{\lambda}{\mu} = \rho < 1$$

Putting, n=1 in (5)

$$P_2 = -\frac{\lambda}{\mu} P_0 + \left(\frac{\lambda}{\mu} + 1 \right) P_1 = \left(\frac{\lambda}{\mu} \right) P_1 = \rho P_1 = \rho^2 P_0$$

Putting, n=2 in (5)

$$P_3 = -\frac{\lambda}{\mu} P_1 + \left(\frac{\lambda}{\mu} + 1 \right) P_3 = \rho^3 P_0$$

Proceeding in same way, we have

$$P_n = \rho^n P_0$$

$$for \ , n \geq 0$$

But,

$$\sum_{n=0}^{\infty} P_n = 1$$

or, $P_0 + P_1 + P_2 + P_3 ..... = 1$

or, $(1 + \rho + \rho^2 + \rho^3 + ...) P_0 = 1$

$$\frac{1}{1-\rho} . P_0 = 1$$

or, $P_0 = 1 - \rho$ ..... (7)

$\therefore$ $P_n = \rho^n (1 - \rho)$..... (8)

Which is the probability that these are n units in the system at any time.

Step 3:

A. To find the probability (queue size is ≥ N)

$$= \sum_{n=N}^{\infty} P_n = \sum_{n=0}^{\infty} P_n - \sum_{n=0}^{N-1} P_n$$

$$= 1 - \left[P_0 + P_1 + P_2 + ... + P_{n-1}\right]$$

$$= 1 - \left[1 + \rho + \rho^2 + ... + \rho^{n-1}\right]P_0$$

$$= 1 - \frac{1.\left(1 - \rho^N\right)}{\left(1 - \rho\right)}.\left(1 - \rho\right) = \rho^N = \left(\frac{\lambda}{\mu}\right)^N$$

B. To find E (Ls), expected line length (average number of customers) in the system

$$E(Ls) = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}$$

C. To find variance of queue length

$$Var(n) = \frac{\rho}{(1-\rho)^2}$$

D. To find E (Lq), expected queue length (average number of customers in the queue or average length of waiting line);

$$E\left(Lq\right) = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu\left(\mu-\lambda\right)}$$

E. To find E (L/L > 0), expected length of non – empty queue;

$$E\left(L/L>0\right) = \frac{1}{1-\rho} = \frac{\mu}{\mu-\lambda}$$

F. E (Wq), Expected waiting time in the queue (excluded service time), i.e., average waiting time of an arrival in the queue;

$$E(Wq) = \frac{\rho}{\mu(1-\rho)} = \frac{\lambda}{\mu(\mu-\lambda)}$$

G. E (Ws), Expected waiting time in the system (including service time), i.e., average time arrival spends in the system;

$$E(Ws) = \frac{1}{\mu(1-\rho)} = \frac{1}{(\mu-\lambda)}$$

H. E (W/W > 0), expected waiting time of a customer who has to wait, i.e., average waiting time of an arrival who waits;

$$E\left(W\ /W\ >0\right)=\frac{1}{\mu\left(1-\rho\right)}=\frac{1}{\left(\mu-\lambda\right)}$$

I. The probability of arrival during the service time of any given customer

$$=\left(\frac{\lambda}{\lambda+\mu}\right)^{m}\cdot\frac{\mu}{\left(\mu+\lambda\right)}$$

J. Relationship between E (Ls), E (Lq), E (Ws), and E (Wq)

$$E(Ls) = \lambda . E(Ws)$$

$$E(Lq) = \lambda . E(Wq)$$

$$E(Ws) = E(Wq) + \frac{1}{\mu}$$

Customer arrived at a sales counter manned by a single person according to a Poisson process with a mean rate of 20 per hour. The time required to serve the customer has an exponential distribution with a mean of 100 seconds. Find the average waiting time of a customer and queue length.

(M/M/I) : (∞/ FCFS)

λ= 20 per hour

μ = (60 × 60)/ 100 = 36 per hour

$$E(Wq) = \frac{\rho}{\mu(1-\rho)} = \frac{\lambda}{\mu(\mu-\lambda)} = 125 \text{ Seconds}$$

$$E(Ws) = \frac{1}{\mu(1-\rho)} = \frac{1}{(\mu-\lambda)} = 225 \text{ Sec}$$

Poisson Arrivals:

In general arrivals in a queuing system do not occur at regular intervals, but tend to be scattered in some fashion. The "Poisson assumption" specifies the behaviour of arrivals, by postulating the existence of constant λ, which is independent of time, queue length or any other random property of the queue, such that

P (an arrival occurs between time t and $t + \Delta t$) = $\lambda \Delta t$.....(1)

If the interval Δt is sufficiently small.

A waiting line for which arrivals occur in accordance with (1) is called a queue with Poisson arrivals.

The Birth and Death Process:

It is a special type of Markov's process, which is a powerful tool to analyze queues. Stated in terms of queuing birth and death process usually arise when there is a customer that is increased by birth or arrival and decreases by the death or departure of serviced customers .

We determine the probability distribution of number of customers in the queuing system at a particular instant of time. These probabilities will then be helpful in determination of operating characteristics for the various types of queuing models to be developed.

Let us define following

Δ t = A time interval so small that the probability of more than one arrival or service is negligible i.e., during a small time interval Δ t, only one arrival or departure can occur.

$\lambda \Delta$ t = Probability of an arrival during Δ t.
$\mu \Delta$ t = Probability of completion of one service during Δ t.
1 - $\lambda \Delta$ t = Probability of no arrival during Δ t.
1 - $\mu \Delta$ t = Probability of no service in time Δ t.

Let P_n (t + Δ t) is the probability of 'n' customers in the system at time t + Δ t . The system contains 'n' customers in the system in one of the following three ways:

1. The system contains ‘n’ customers in the system at time ‘t’ and there is no arrival and service in the time ‘Δ t’.
2. The system contains ‘n + 1’ customers in the system at time ‘t’ and there is no arrival and one service in the time ‘Δ t’.
3. The system contains ‘n – 1’ customers in the system at time ‘t’ and there is one arrival and no service in the time ‘Δ t’.

All the above three events are mutually exclusive and exhaustive, therefore

$$P_n(t+\Delta t) = P_n(t)(1-\lambda\Delta t)(1-\mu\Delta t) + P_{n+1}(t)(1-\lambda\Delta t)(\mu\Delta t) + P_{n-1}(t)(\lambda\Delta t)(1-\mu\Delta t)$$
$$= \left[P_n(t) - P_n(t)(\lambda+\mu) + \lambda P_{n-1}(t) + \mu P_{n+1}(t)\right]\Delta t$$

Terms of $(\Delta t)^2$ are dropped due to a very small amount.
At steady state $P_n(t)$ becomes independent of time t

So, $P_n(t + \Delta t) = P_n(t) = P_n$

Therefore, the above equation becomes

$$P_n = \left[P_n - P_n(\lambda+\mu) + \lambda P_{n-1} + \mu P_{n+1}\right]\Delta t$$
$$0 = \left[\lambda P_{n-1} - P_n(\lambda+\mu) + \mu P_{n+1}\right]\Delta t$$
$$0 = \left[\lambda P_{n-1} - P_n(\lambda+\mu) + \mu P_{n+1}\right]$$
$$\left[\lambda P_{n-1} - P_n(\lambda+\mu) + \mu P_{n+1}\right] = 0$$

Similarly, if the system is empty (no customer) at time (t + Δ t), then there will be no service completion during time Δ t. Thus we have only two possibilities instead of three.

$$P_0(t + \Delta t) = [P_0(t)(1 - \lambda \Delta t) + P_1(t)\mu]\Delta t$$

$$P_1\mu - P_0\lambda = 0; n = 0$$

Classification of Queues:

Generally queuing problem may be completely specified in the following symbolic form

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

The first and second symbols denote the type of distribution of inter – arrival times and of the inter service times, respectively. Third symbol specifies the number of servers, whereas fourth symbol stands for the capacity of the system and the last symbol denotes the queue discipline.

If we specify the following letters as

M = Poisson arrival or departure distributions,
E_k = Erlangian or Gamma inter – arrival or service time distribution.

Then $(M/ E_k /1) : (\infty/FCFS)$ defines a queuing system in which arrivals follow Poisson distribution, service time are Erlangian, single server, infinite population and FCFS queue discipline.