Unit 4: Queuing Theory 6 Hrs.

Basis of Queuing theory, elements of queuing theory, Kendall's Notation, Operating characteristics of a queuing system, Classification of Queuing models.

#Past Questions

1. Classify the queueing models with examples. -2021
2. Write Short notes on:
   a. Kendall's notation for queueing models.-2021/2023
   b. Queuing Theory -2024
3. Discuss the single-channel queueing model with suitable example. Give its assumptions . -2023
4. A radio machine on an average finds 5 customers coming to his shop every hour for repairing their radio sets. He disposes of each of them within 10 minutes on an average. The arrival and servicing times follows Poisson and exponential distribution respectively. In the light of the above facts determine: -2024
   a. Proportion of time during which shop remains empty.
   b. The average no. of customers in his system and queue.
   c. The average time spent by a customer n the queue and the service as well.
   d. The probability of finding at least five customers in his shop.

1. Basis of Queuing Theory

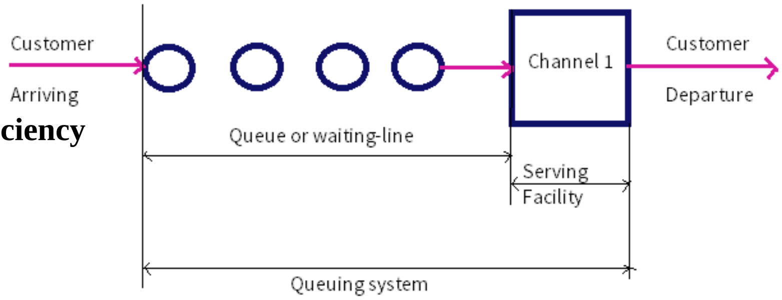
A queue is a line or list of customers who remain waiting for getting a service from a service center. A queue is formed if the arrival rate of customers is greater than the service rate.

In general, the queueing system consists of one or more queues and one or more servers and operates under a set of procedures. Depending upon the server status, the incoming customer either waits at the queue or gets the turn to be served. If the server is free at the time of arrival of a customer, the customer can directly enter into the counter for getting service and then leave the system. In this process, over a period of time, the system may experience " Customer waiting" and /or "Server idle time".

The time spent by the customers in waiting line is often expensive in terms of money, equipment, opportunities, etc. These costs related to waiting in line are called cost of waiting. The const of waiting can be decreased by adding additional service center; however, the addition service facilities would finally increase the cost of providing service because if there is no queue, the service center may be idle and the organization has to bear the cost of service center being idle. This cost is known as service idle cost. Similarly, if the numbers of service center are reduced, it minimizes the service cost but increase the waiting cost due to large waiting line. The main objective of **Queuing theory** is to mange balance between these two costs – cost of waiting and the service cost.

Queuing theory is the mathematical study of **waiting lines or queues**. It helps analyze:

- **Why waiting lines form**
- **How long people or items will wait**
- **How to reduce waiting time or improve service efficiency**

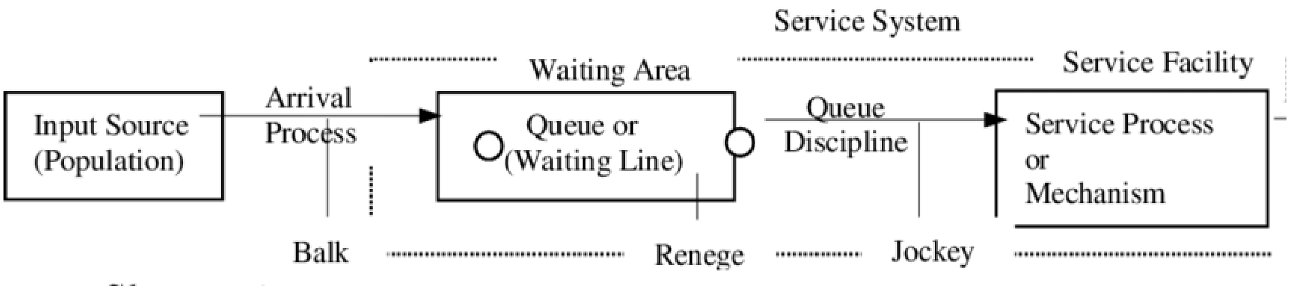


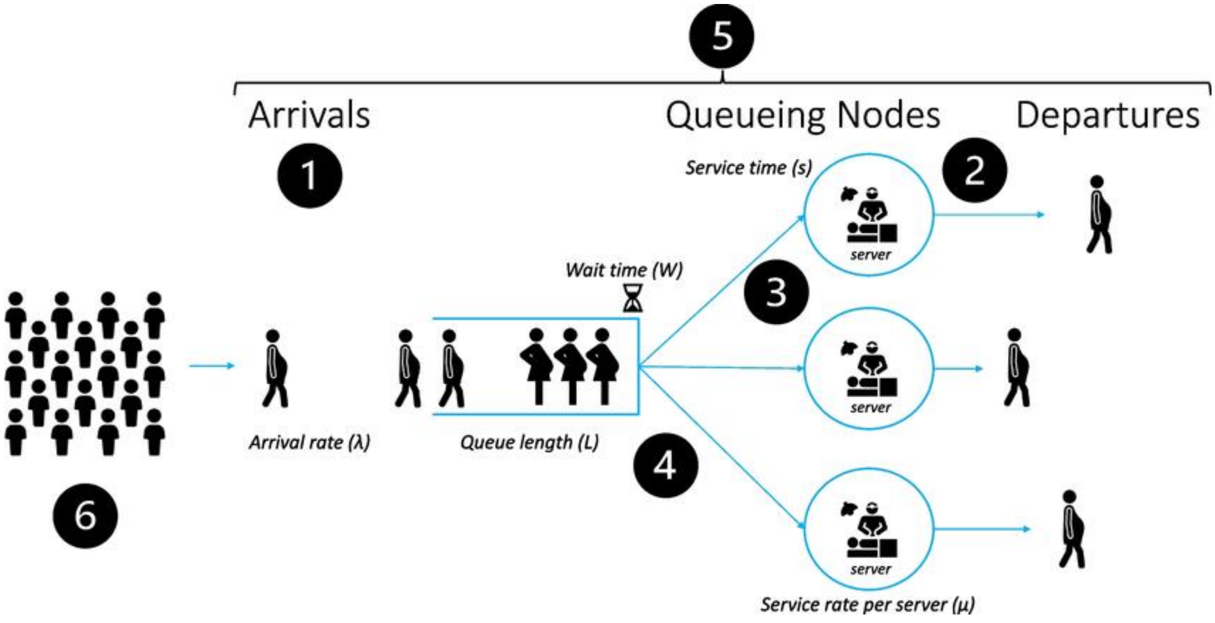
➤ Objectives:

- Balance service costs vs. customer waiting costs.
- Predict system performance (wait times, queue lengths).
- Optimize service facilities
- Minimize customer dissatisfaction and resource idle time

➤ Real-Life Applications:

- Telecom: Call center management.
- Transportation: Traffic light timing.
- Healthcare: Patient flow in hospitals.
- Manufacturing: Assembly line bottlenecks.
- Supermarkets
- Banks and ATMs





2. Elements of a Queuing System

A queueing system can be completely described by

(1) The input (arrival pattern)

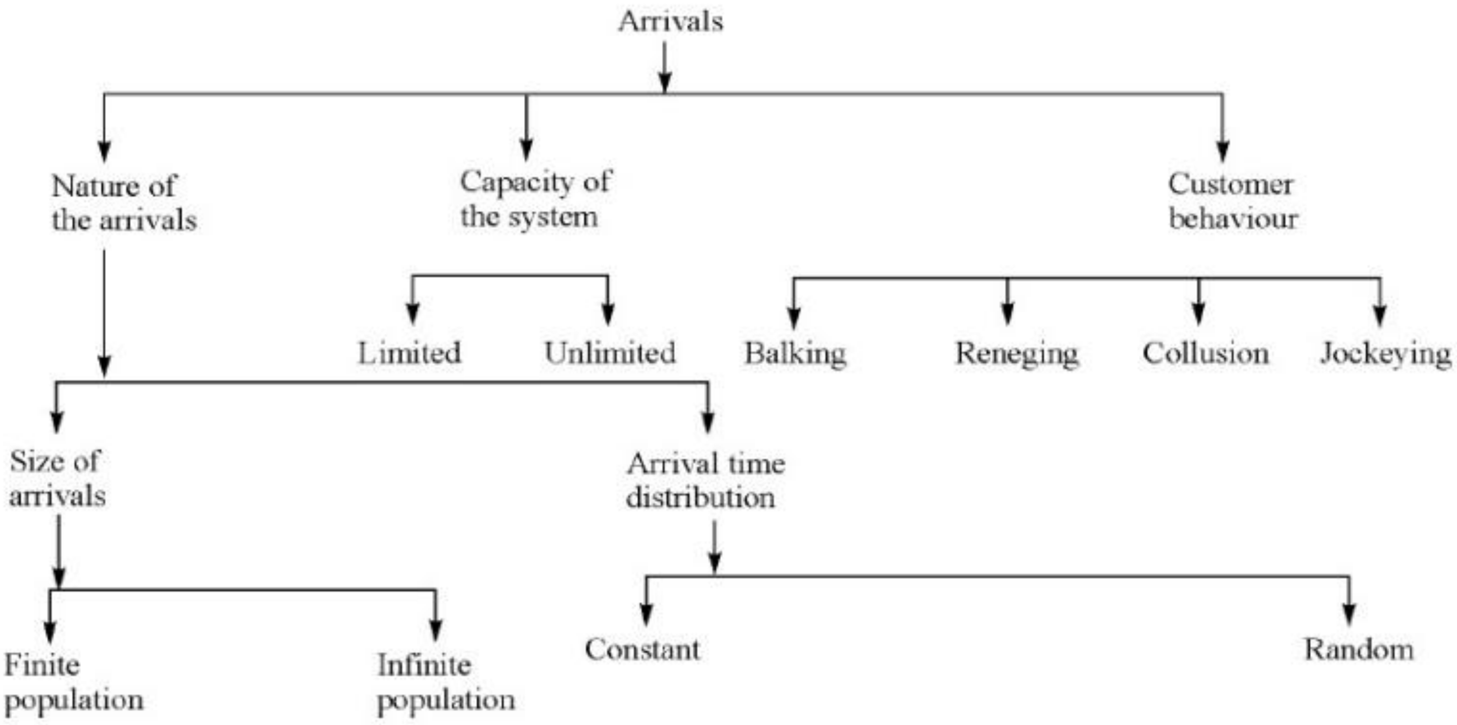
(2) The service mechanism (service pattern)

(3) The queue discipline and

(4) Customer's behavior

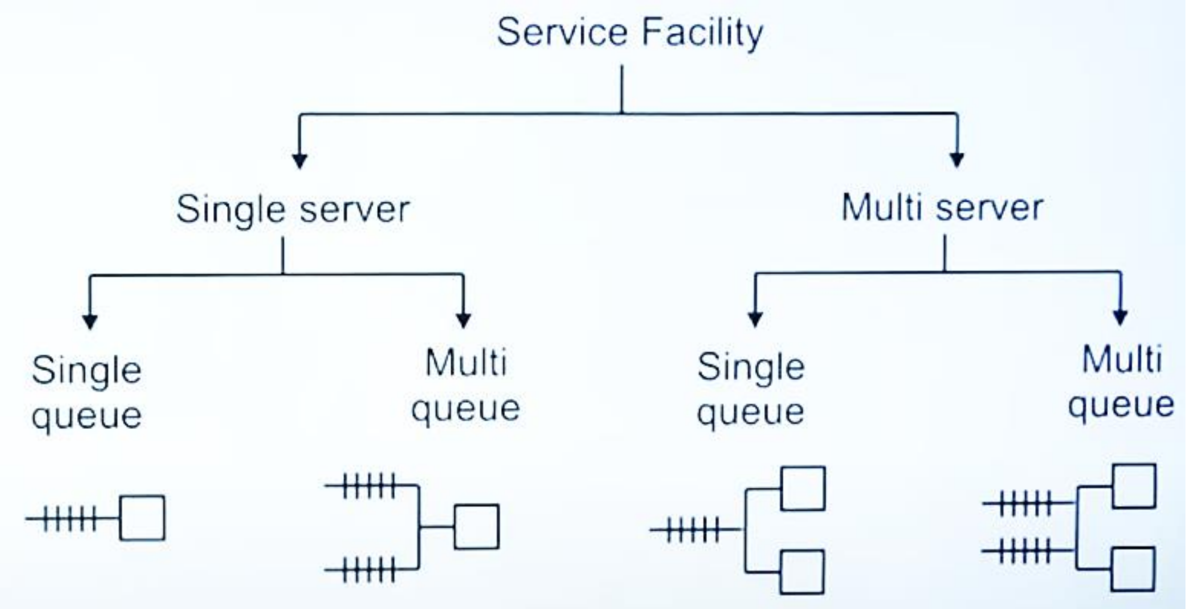
1. <u>The input (arrival pattern)</u>

The input described the way in which the customers arrive and join the system. Generally, customers arrive in a more or less random manner which is not possible for prediction. Thus, the arrival pattern can be described in terms of probabilities and consequently the probability distribution for **inter-arrival** times (the time between two successive arrivals) must be defined. We deal with those Queueing systems in which the customers arrive in Poisson process. The mean arrival rate is denoted by λ.



2. The Service Mechanism:-

This means the arrangement of service facility to serve customers. If there is infinite number of servers, then all the customers are served instantaneously or arrival and there will be no queue. If the number of servers is finite then the customers are served according to a specific order with service time a constant or a random variable. Distribution of service time follows '**Exponential distribution'**



3. Queueing Discipline:-

It is a rule according to which the customers are selected for service when a queue has been formed. The most common disciplines are

1. First come first served – (FCFS) / (FIFO)

2. Last in first out – (LIFO)
3. Selection for service in random order (SIRO)
4. Priority-based

4. Customer's behaviour

The arriving customers may have different attitudes and behaviors.

- Generally, it is assumed that the customers arrive into the system one by one. But in some cases, customers may arrive in groups. Such arrival is called **Bulk arrival.**
- If the queue length appears very large to a customer, he/she may not join the queue. This property is known as **Balking** of customers.
- Sometimes, a customer who is already in a queue will leave the queue in anticipation of longer waiting line. This kind of departure is known as **reneging.**
- If there is more than one queue, the customers from one queue may be tempted to join another queue because of its smaller size. This behavior of customers is known as **jockeying.**

---

3. Kendall's Notation

Kendall's shorthand classifies queues as:

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

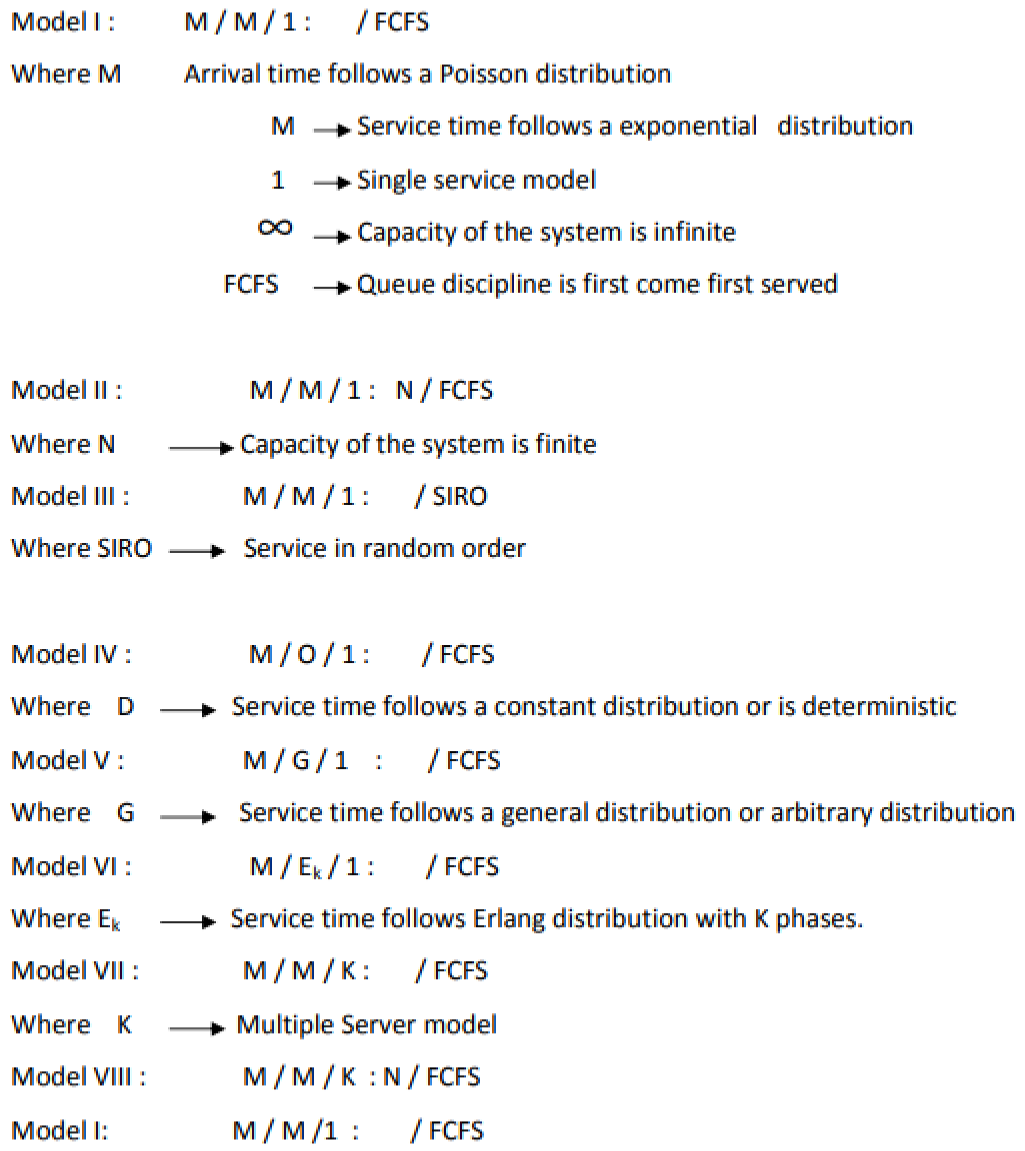
- **a**: Arrival process distribution (`M` = Markovian/Poisson, `D` = Deterministic, `G` = General).
- **b**: Service time distribution (`M`, `D`, `G`).
- **c**: Number of servers (`1`, `s`, `∞`).
- **d**: System capacity (max customers, e.g., `N` or `∞`).
- **e**: Queue discipline (FIFO, LIFO, SIRO).

Common symbols for **a** and **b** :

- M= Markovian (Arrival time follows Poisson distribution and service time follows an exponential distribution.)
- D= constant or deterministic inter-arrival-time or service-time.
- G = General service time (departures) distribution.
- GI = General independent arrival distribution
- Ek = Erlang-k distribution of inter-arrival or service time distribution with parameter k (i.e. if k = 1, Erlang is equivalent to exponential and if k = , Erlang is equivalent to deterministic).

Common symbols for **c** :

- 1 – single server
- S – multiple server
- ∞ - infinite server



Model I : M / M / 1 : / FCFS

Where M Arrival time follows a Poisson distribution

M → Service time follows a exponential distribution

1 → Single service model

∞ → Capacity of the system is infinite

FCFS → Queue discipline is first come first served

Model II : M / M / 1 : N / FCFS

Where N → Capacity of the system is finite

Model III : M / M / 1 : / SIRO

Where SIRO → Service in random order

Model IV : M / O / 1 : / FCFS

Where D → Service time follows a constant distribution or is deterministic

Model V : M / G / 1 : / FCFS

Where G → Service time follows a general distribution or arbitrary distribution

Model VI : M / E_k / 1 : / FCFS

Where E_k → Service time follows Erlang distribution with K phases.

Model VII : M / M / K : / FCFS

Where K → Multiple Server model

Model VIII : M / M / K : N / FCFS

Model I: M / M /1 : / FCFS

4. Operating Characteristics of a Queuing System

- **Mean Arrival rate (λ) =** number of customers arriving per unit time
- **Mean Service rate (μ) =** number of customers served per unit time.

Characteristic	**Formula**	**Meaning**
Server Utilization / Traffic intensity(ρ)	$\rho = \lambda / \mu$	% of time server is busy
System Length(Ls)	$Ls = \lambda / (\mu - \lambda)$	Avg. customers in queue + service
Queue Length (Lq)	$Lq = \lambda^2 / (\mu(\mu - \lambda))$	Avg. customers in queue only
Waiting time in system (Ws)	$W = 1 / (\mu - \lambda)$	Avg. time spent by customer in the system (queue + service)
Waiting time in queue (Wq)	$Wq = \lambda / (\mu(\mu - \lambda))$	Avg. time waiting before being served
P_0 (no customers)	$P_0 = 1 - \rho$	Probability that system is empty

5. Classification of Queuing Models

a. Single Server Queuing Model

Model I: (M/M/1): (∞/FCFS)

- **Meaning**:
  - **M/M/1**: Poisson arrivals, Exponential service times, Single server
  - **∞/FCFS**: Infinite queue capacity, First-Come-First-Served discipline
- **Characteristics**:
  - Customers never leave the queue (infinite waiting space)
  - Arrival rate (λ) must be less than service rate (μ) for system stability ($\rho = \lambda/\mu < 1$)
- **Example**:

🏧 **A single ATM in a 24-hour bank lobby**

  - Arrivals: Random customers (Poisson process)
  - Service: Exponential transaction times
  - Queue: No limit on waiting customers
  - Discipline: Customers served in arrival order

Model II: (M/M/1): (N/FCFS)

- **Meaning**:
  - **M/M/1**: Poisson arrivals, Exponential service times, Single server
  - **N/FCFS**: Finite capacity for **N total customers** (including the one being served), FCFS discipline
- **Characteristics**:
  - New arrivals are **blocked/rejected** if the system is full (queue holds N-1 customers)
  - Works even if $\rho = \lambda/\mu \geq 1$ (since finite size prevents infinite queues)
- **Example**:

👦🍳 **A small coffee shop with 5 seats**

  - Arrivals: Random customers
  - Service: Exponential time per order
  - Capacity: Max 5 customers total (1 ordering + 4 waiting)
  - New customers leave if all seats are occupied

---

b. Multi-Server Queuing Model

Model III: (M/M/S): (∞/FCFS)

- **Meaning**:
  - **M/M/S**: Poisson arrivals, Exponential service times, **Multiple servers (S)**
  - **∞/FCFS**: Infinite queue, FCFS discipline
- **Characteristics**:
  - Customers wait in a **common queue** and go to the **first available server**

    - System stable if $\rho = \lambda/(S\mu) < 1$
- **Example**:

🏢 **A bank with 4 tellers and a single queue**

    - Arrivals: Random customers
    - Service: Exponential transaction times
    - Queue: Infinite waiting space
    - Discipline: Next customer goes to whichever teller is free

---

Model IV: (M/M/S): (N/FCFS)

- **Meaning**:
    - **M/M/S**: Poisson arrivals, Exponential service times, **S servers**
    - **N/FCFS**: **Finite capacity for N customers** (N > S), FCFS discipline
- **Characteristics**:
    - Maximum **N customers total** (S being served + N-S waiting)
    - Arrivals are **rejected** when the system is full
- **Example**:

🚗 **A drive-through with 2 service windows and 5-car capacity**

    - Arrivals: Random vehicles
    - Service: Exponential service time per car
    - Capacity: Max 5 cars (2 at windows + 3 in queue)
    - New cars leave if 5 cars are already in the system

Notes:

- Arrival time often modeled using Poisson distribution (random arrivals)
- Service time often modeled using exponential distribution (random, memoryless service durations)

Key Property: Memoryless

The chance of an event happening in the future **does not depend** on how long you've already waited.

Example: If you're waiting in line and have already waited 5 minutes, the probability you'll wait 2 more minutes is the same as it was when you just arrived.

Real-Life Example:

- In a ticket counter, if customers arrive randomly at **10 per hour**, then the time between arrivals is **exponentially distributed** with rate $\lambda=10$.
- Most customers will arrive within a few minutes of each other, but occasionally, you may have to wait longer — this randomness is modeled by the exponential distribution.

Single server model

(M/ M/1) QUEUING MODEL

The M/M/1 queuing model is a queuing model where the arrivals follow a Poisson process, service times are exponentially distributed and there is one server.

The assumption of M/M/1 queuing model are as follows :

1. The number of customer arriving in a time interval t follows a poison process with parameter λ.
2. The interval between any two successive arrival is exponentially distributed with parameters λ.
3. The time taken to complete a single service is exponentially distributed with parameter μ.
4. The number of server is one.
5. Although not explicitly stated both the population and the queue size can be infinity.
6. The order of service is assumed to be FCFS.

Practical formulae involved in single server model I

• Arrival rate per hour	$= \lambda$
• Service rate per hour	$= \mu$
• Average utilization rate (or utilization factor), ρ	$= \lambda / \mu$
• Average waiting time in the system, (waiting and servicing time) W_s	$= 1 / (\mu-\lambda)$
• Average waiting time in the queue, Wq	$= \frac{\lambda}{\mu (\mu -\lambda)}$
• Average number of customers (including the one who is being served) in the system, L_s	$= \frac{\lambda}{(\mu - \lambda)}$
• Average number of customers(excluding the one who is being served) in the queue L_q	$= \frac{\lambda^2}{\mu(\mu-\lambda)}$
• Average number of customers in non-empty queue that forms time to time	$= \mu / (\mu-\lambda)$
• Probability of no customer in the system, or, system is idle or idle men in the factor P_0	$= 1 -(\lambda / \mu)$ or1 – ρ = 1- utilization factor
• Probability of no customer in queue and a customer is being served P1	$= (1 - \lambda / \mu) (\lambda / \mu)$
• Probability of having 'n' customers in the system	$=(1 - \lambda / \mu) (\lambda / \mu)^n$
• Probability of having 'n' customers in the queue	$= (1 - \lambda / \mu) (\lambda / \mu)^{n+1}$

Solved Example Problem 1

A television repairman finds that the time spent on his jobs has an exponential distribution with mean of 30 minutes. If he repairs sets in the order in which they came in, and if the arrival of sets follows a Poisson distribution approximately with an average rate of 10 per 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?

Solution:

From the data of the problem, we have
λ = 10/8 = 5/4 sets per hour; and μ= (1/30) 60 = 2 sets per hour

(a) Expected idle time of repairman each day = Number of hours for which the repairman remains busy in an 8-hour day (traffic intensity) is given by

(8) (λ /μ) = (8) (5/8) = 5 hours

Hence, the idle time for a repairman in an 8-hour day will be: (8 – 5) = 3 hours.

(b) Expected (or average) number of TV sets in the system

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{5/4}{2-(5/4)} = \frac{5}{3} = 2\text{(approx.) TV sets}$$

Solved Example Problem 2

On an average 96 patients per 24-hour day require the service of an emergency clinic. Also on an average, a patient requires 10 minutes of active attention. Assume that the facility can handle only one emergency at a time. Suppose that it costs the clinic Rs 100 per patient treated to obtain an average servicing time of 10 minutes, and that each minutes of decrease in this average time would cost Rs. 10 per patient treated. How much would have to be budgeted by the clinic to decrease the average size of the queue from one and one-third patients to half patient.

Solution:

From the data of the problem, we have
$\lambda = \frac{96}{24\times 60} = \frac{1}{15}$ and $\mu = \frac{1}{10}$ patients per minute; $p = \frac{\lambda}{\mu} = \frac{2}{3}$

1. Average number of patients in the queue

$$L_q = \frac{p^2}{1-p} = \frac{(2/3)^2}{1-2/3} = \frac{4}{3}$$

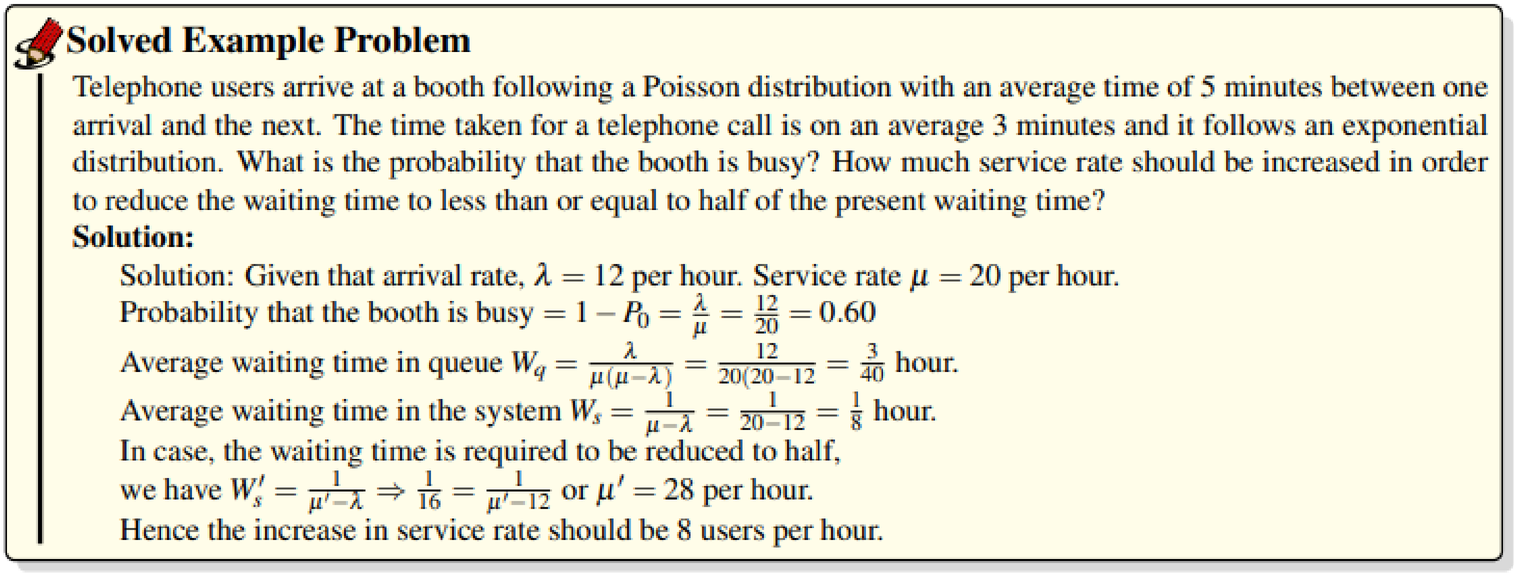
2. Fraction of the time for which there no patients, $P_0 = 1 - p = 1 - \frac{2}{3} = \frac{1}{3}$

3. When the average queue size is decreased from $4/3$ patient, the new service rate is determined as:

   $L_q = \frac{\lambda^2}{\mu(\mu-\lambda)}$ or $\frac{1}{2} = \frac{(1/15)^2}{\mu(\mu-1/15)^2}$, i.e $\mu = \frac{2}{15}$ patients per minute.

   Average rate of treatment required is: $\frac{1}{\mu} = \frac{15}{2} = 7.5$ minutes i.e. a decrease in the average rate of treatment is 2.5(= 10 – 7.5) minutes.

   Budget per patient = Rs (100 + 2.5 x 10) = Rs 125 per patient.



Solved Example Problem

Telephone users arrive at a booth following a Poisson distribution with an average time of 5 minutes between one arrival and the next. The time taken for a telephone call is on an average 3 minutes and it follows an exponential distribution. What is the probability that the booth is busy? How much service rate should be increased in order to reduce the waiting time to less than or equal to half of the present waiting time?

Solution:

Solution: Given that arrival rate, $\lambda = 12$ per hour. Service rate $\mu = 20$ per hour.

Probability that the booth is busy $= 1 - P_0 = \frac{\lambda}{\mu} = \frac{12}{20} = 0.60$

Average waiting time in queue $W_q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{12}{20(20-12)} = \frac{3}{40}$ hour.

Average waiting time in the system $W_s = \frac{1}{\mu-\lambda} = \frac{1}{20-12} = \frac{1}{8}$ hour.

In case, the waiting time is required to be reduced to half,

we have $W_s' = \frac{1}{\mu'-\lambda} \Rightarrow \frac{1}{16} = \frac{1}{\mu'-12}$ or $\mu' = 28$ per hour.

Hence the increase in service rate should be 8 users per hour.

5. A radio machine on an average finds 5 customers coming to his shop every hour for repairing their radio sets. He disposes of each of them within 10 minutes on an average. The arrival and servicing times follows Poisson and exponential distribution respectively. In the light of the above facts determine: -2024

a. Proportion of time during which shop remains empty.

b. The average no. of customers in his system and queue.

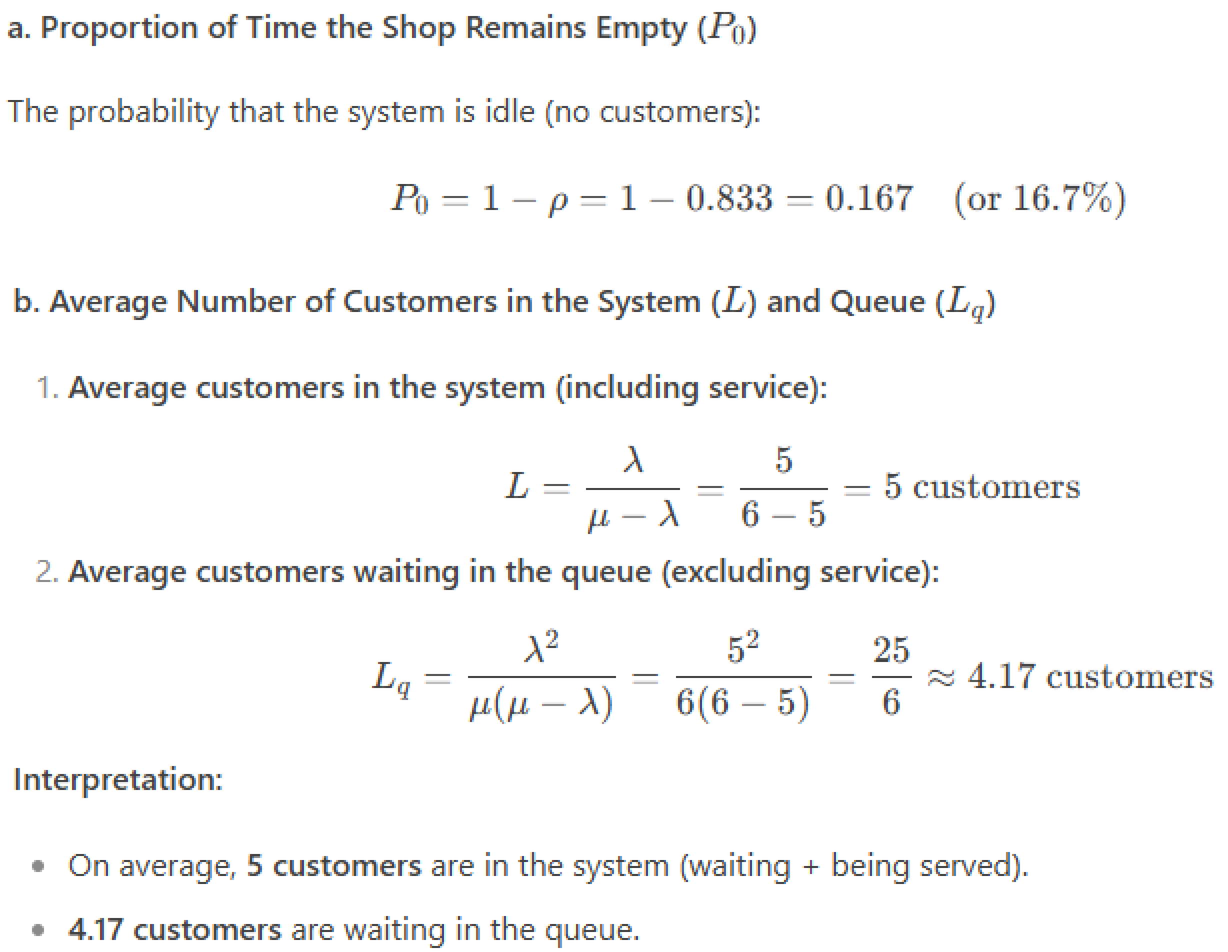
c. The average time spent by a customer n the queue and the service as well.

d. The probability of finding at least five customers in his shop.

Solutions:

Given:

- Arrival rate (λ) = **5 customers/hour**
- Service rate (μ) = **6 customers/hour** (since 10 minutes/customer = 6 customers/hour)



a. Proportion of Time the Shop Remains Empty (P_0)

The probability that the system is idle (no customers):

$$P_0 = 1 - \rho = 1 - 0.833 = 0.167 \quad \text{(or 16.7\%)}$$

b. Average Number of Customers in the System (L) and Queue (L_q)

1. **Average customers in the system (including service):**

$$L = \frac{\lambda}{\mu - \lambda} = \frac{5}{6-5} = 5 \text{ customers}$$

2. **Average customers waiting in the queue (excluding service):**

$$L_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{5^2}{6(6-5)} = \frac{25}{6} \approx 4.17 \text{ customers}$$

Interpretation:

- On average, **5 customers** are in the system (waiting + being served).
- **4.17 customers** are waiting in the queue.

\

c. Average Time Spent in the System (W) and Queue (W_q)

1. **Average time in the system (waiting + service):**

$$W = \frac{1}{\mu - \lambda} = \frac{1}{6 - 5} = 1 \text{ hour}$$

2. **Average waiting time in the queue:**

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{5}{6(6 - 5)} = \frac{5}{6} \approx 0.833 \text{ hours} = 50 \text{ minutes}$$

Interpretation:

- A customer spends **1 hour** on average in the shop.
- They wait **50 minutes** in the queue before service begins.

d. Probability of Finding at Least 5 Customers in the Shop ($P(n \geq 5)$)

The probability of **exactly** n customers in the system:

$$P_n = (1 - \rho)\rho^n = (0.167)(0.833)^n$$

For **at least 5 customers**, sum probabilities from $n = 5$ to ∞:

$$P(n \geq 5) = \rho^5 = (0.833)^5 \approx 0.402 \quad \text{(or } 40.2\%\text{)}$$

Interpretation: There's a **40.2% chance** that 5 or more customers are in the shop at any given time.