UNIT 2
QUEUING THEORY

LESSON 22

Learning Objective:

- **Explain standard queuing language and symbols.**
- **Explain the operating characteristics of a queue in a business model**
- **Apply formulae to find solution that will predict the behaviour of the model.**

Hello students,
In this lesson you are going to learn the various performance measures and their relevance in queuing theory.

PERFORMANCE MEASURES (OPERATING CHARACTERISTICS)

An analysis of a given queuing system involves a study of its different operating characteristics.

Important Notations

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

n = number of customers in the system (waiting and in service)

P_n = probability of n customers in the system

λ = average (expected) customer arrival rate or average number of arrivals per unit of time in the queuing system

μ = average (expected) service rate or average number of customers served per unit time at the place of service

$\frac{\lambda}{\mu} = \rho = \frac{\text{Average service completion time } (1/\mu)}{\text{Average inter-arrival time } (1/\lambda)}$

= traffic intensity or server utilization factor (the expected fraction of time for which server is busy)

s = number of service channels (service facilities or servers)

N = maximum number of customers allowed in the system.

L_s = average (expected) number of customers in the system (waiting and in service)

L_q = average (expected) number of customers in the queue (queue length)

L_b = average (expected) length of non-empty queue

W_s = average (expected) waiting time in the system (waiting and in service)

W_q = average (expected) waiting time in the queue

P_w = probability that an arriving customer has to wait

Some of the performance measures (operating characteristics of any queuing system that are of general interest for the evaluation of the performance of an existing queuing system, and to design a new system in terms of the level of service a customer receives as well as the proper utilization of the service facilities are listed:-

1. Time-related questions for the customers

a) What is the average (or expected) time an arriving customer has to wait in the queue (denoted by W_q) before being served.

b) What is the average (or expected) time an arriving customer spends in the system (denoted by W_s) including waiting and service. This data can be used to make economic comparison of alternative queuing systems.

2. Quantitative questions related to the number of customers

a) Expected number of customers who are in the queue (queue length) for service, and is denoted by L_q

b) Expected number of customers who are in the system either waiting in the queue or being serviced (denoted by L_s). The data can be used for finding the mean customer time spent in the system.

3. Questions involving value of time both for customers and servers

a) What is the probability that an arriving customer has to wait before being served (denoted by P_w)? It is also called blocking probability.
b) What is the probability that a server is busy at any particular point in time (denoted by ρ)? It is the proportion of the time that a server actually spends with the customer, i.e. the fraction of the time a server is busy.
c) What is the probability of n customers being in the queuing system when it is in steady state condition? It is denoted by $P_{n,}$ n = 0, 1……
d) What is the probability of service denial when an arriving customer cannot enter the system because the queue is full? It is denoted by P_d .

4. Cost-related questions

a) What is the average cost needed to operate the system per unit of time?
b) How many servers (service centres) are needed to achieve cost effectiveness?

To describe the distribution of these variables, we should specify its average value, standard deviation and the probability that the variable exceeds a certain value.

Transient-State and Steady-State

When a service system is started it progresses through a number of changes. However, it attains stability after some time. Before the start of the service operations it is very much influenced by the initial conditions (number of customers in the system) and the elapsed time. This period of transition is termed as transient-state. However, after sufficient time has passed, the system becomes independent of the initial conditions and of the elapsed time (except under very special conditions) and enters a steady- state condition.

In this chapter an analysis of the queuing system will be discussed under steady-state conditions.

Let $P_n(t)$ denote the probability that there are n customers in the system at time t. The rate of change in the value $P_n(t)$ with respect to time t is denoted by the derivative of $P_n(t)$ with respect to t, i.e. $P_n(t)$. In the case of steady - state, we have

$$\lim_{t\to\infty} P_n(t) = P_n \text{ (independent of t)}$$

or

$$\lim_{t\to\infty} \frac{d}{dt}\{P_n(t)\} = \frac{d}{dt}(P_n)$$

or

$$\lim_{t\to\infty} P'_n(t) = 0$$

In some cases when arrival rate of customers in the system is more than the service rate, then a steady - state cannot be reached regardless of the length of the elapsed time.

Relationships Among Performance Measures

By definition of various measures of performance (operating characteristic), we have

$$L_s = \sum_{n=0}^{\infty} n P_n \quad \text{and} \quad L_q = \sum_{n=s}^{\infty} (n-s) P_n$$

Some general relationships between the average system characteristics true for all queuing models are as follows:

(i) Expected number of customers in the system is equal to the expected number of customers in queue plus in service.

$$L_s = L_q + \text{Expected number of customers in service}$$

$$= L_q + \lambda/\mu$$

The value of expected number of customers in service, should not be confused with the number of service facilities but it is equal to ρ for all queuing models except finite queue case.

(ii) Expected waiting time of the customer in the system is equal to the average waiting time in queue plus the expected service time.

$$W_s = W_q + \frac{1}{\mu}$$

(iii) Expected number of customers served per busy period is given by

$$L_b = \frac{L_s}{P(n \geq s)} = \frac{\mu}{\mu - \lambda}$$

Where $P(n \geq s)$ = probability that the system being busy

(iv) Expected length of queue during busy period is given by

$$W_b = \frac{W_q}{P(n \geq s)} = \frac{1}{\mu - \lambda}$$

(v) Expected number of customers in the system is equal to the average number of arrivals per unit of time multiplied by the average time spent by the customer in the system.

$$L_s = \lambda W_s$$

$$\text{or } W_s = \frac{1}{\lambda} L_s$$

(vi) $$L_q = \lambda W_q$$

$$\text{or } W_q = \frac{1}{\lambda} L_q$$

$$= \frac{Ls}{\mu}$$

For applying formula (v) and (vi) for system with finite queue, instead of using λ, its effective value λ ($1-P_N$) must be used.

(vii) The probability, P_n of n customers in the queuing system at any time can be used to determine all the basic measures of performance in the following order.

$$L_s = \sum_{n=0}^{\infty} nP_n$$

$$\Rightarrow W_s = \frac{L_s}{\lambda}$$

$$\Rightarrow W_q = W_s - \frac{1}{\mu}$$

$$\Rightarrow L_q = \lambda W_q$$

PROBABILITY DISTRIBUTIONS IN QUEUING SYSTEMS

It is assumed that customers joining a queuing system arrive in random manner and follow a Poisson distribution or equivalently the inter-arrival times follow exponential distribution.

In most of the cases, service times are also assumed to be exponentially distributed. It implies that the probability of service completion in any short-time period is constant and independent of the length of time that the service has been in progress. The basic reason for assuming exponential service is that it helps in formulating simple mathematical models which ultimately help in analyzing a number of aspects of queuing problems.

The number of arrivals and departures (those served) during an interval of time in a queuing system is controlled by the following assumptions (also called axioms).

(i) The probability of an event (arrival or departure) occurring during the time interval (t, t+Δt) depends on the length of time interval Δt. That is, probability of the event does not depend either on number of events that occur unto time t or the specific value of t, meaning that the events that occur in non- overlapping time are statistically independent.

(ii) The probability of more than one event occurring during the time interval (t, t+ Δt) is negligible. It is denoted by 0(Δt).

(iii) Atmost one event (arrival or departure) can occur during a small time interval Δt. the probability of an arrival during the time interval (t, t + Δt) is given by

$$P_1 (\Delta t) = \lambda \Delta t + 0(\Delta t)$$

where λ is a constant and independent of the total number of arrivals up to time t; Δt is a small time interval and $0(\Delta t)$ represents the quantity that becomes negligible when compared to Δt as $\Delta t \rightarrow 0$, i.e.

$$\lim_{\Delta t \rightarrow 0} \{ 0 (\Delta t) / \Delta t \} = 0$$

DISTRIBUTION OF ARRIVALS (pure birth process)

The arrival process assumes that the customers arrive at the queuing at the queuing system and never leave it. Such a process is called pure birth process. The aim is to derive an expression for the probability P_n (t) if n arrivals during time interval (t, t+ Δt). the terms commonly used in the development of various queuing models are the following:

Δt = a time interval so small that the probability of more than one customer's arrival is negligible, i.e. during any given small interval of time Δt only one customer can arrive.

$\lambda \Delta t$ = probability that a customer will arrive in the system during time Δt.

1- $\lambda \Delta t$ = probability that no customer will arrive in the system during time Δt.

If the arrivals are completely random, then the probability distribution of a number of arrivals in a fixed time interval follows Poisson distribution.

DISTRIBUTION OF INTER- ARRIVAL TIMES (exponential process)

If the number of arrivals, n, in time t follows the Poisson distribution, then

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \quad n = 0,1,2,\ldots$$

is an associated random variable defined as the inter – arrival time T follows the exponential distribution

$$f(t) = \lambda e^{-\lambda t}$$ and vice – versa.

Markovian property of inter- arrival times

The markovian property of inter arrival times states that the probability that a customer currently in service is completed at some time t is independent of how long he has already been in service. That is

$$\text{Prob.} = \{T \geq t_1 \mid T \geq t_o\} = \text{prob.}\ \{0 \leq T \leq t_1 - t_o\}$$

where T is the time between successive arrivals.

DISTRIBUTION OF DEPARTURES (pure death process)

The departure process assumes that no customer joins the system while service is continued for those who are already in the system. Let, at time t =0 (starting time) there be $N \geq 1$ customers in the system. Since service is being provided at the rate of μ. Therefore, customers leave the system at the rate μ after being serviced. Such a process is called pure death process

Basic axioms

(i) probability of the departure during time Δt is $\mu \Delta t$.

(ii) probability of more than one departure between time t and $t + \Delta t$ is negligible.

(iii) The number o f departures in non- overlapping intervals are statistically independent.

The following terms are used in the development of various queuing models.

$\mu \Delta t$ = probability that a customer in-services at time t will complete service during time Δt.

$1 - \mu \Delta t$ = probability that the customer in-service at time t will not complete service during time Δt.

DISTRIBUTION OF SERVICE TIMES

The probability density function s(t) of service time is given by

$$S(t) = \begin{cases} \mu e^{-\mu t} & ; \ 0 \le t \le \infty \\ 0 & ; \ t < 0 \end{cases}$$

This shows that service times follows negative exponential distribution with mean $\mathbf{1 / \mu}$ and variance $\mathbf{1 / \mu^2}$ **.**

The area under the negative exponential distribution curve is determined as :

$$F(T) = \int_0^t \mu \, e^{-\mu t} \, dt = \left[-\mu \, e^{-\mu t} \right]_0^t$$

$$= -e^{-\mu t} + e^0 = 1 - e^{-\mu t}$$

It is also described as :

$$F(T) = f(t \le T) = 1 - e^{-\mu t}$$

Where F(T) is the area under the curve to the left of T. Thus

$$1 - F(T) = f(t \ge T) = e^{-\mu t}$$

is the area under the curve to the right of T.

For example, if mean service time ($1 / \mu$) at a service station is 2 minutes, then probability that service will take T or more minutes is shown below :

Service times of at least T :	0	1	2	3	4	5
Probability $f(t \ge T)$:	1	0.607	0.368	0.223	0.135	0.082

So, now let us summarize today's discussion:

Summary

We have discussed in details about

- Symbols used in queuing theory
- Operating characteristics of queuing.
- Relationships Among Performance Measures
- Probability distributions in queuing systems

Slide 1

Measuring System Performance

- The total time an "entity" spends in the system.
  - Denoted by W.
- The time an "entity spends in the queue.
  - Denoted by W_q.
- The number of "entities" in the system.
  - Denoted by L.
- The number of "entities" in the queue.
  - Denoted by L_q.
- The percentage of time the servers are busy.
- These quantities are variable over time.
- What might we be interested in knowing about these quantities?

Slide 2

An Example of Lack of Memory

- Suppose that the amount of time one spends in a bank is exponentially distributed with mean ten minutes.
  - What is λ?
  - What is the probability that a customer will spend more a quarter of an hour in the bank?
  - You have been waiting for ten minutes already. Now what is probability that you will spend more than a quarter of an hour in the bank?
  - What has a lack of memory, you or the distribution?

Slide 3

Point Processes

- A Point Process is a stochastic process that only changes at discrete points in time.
- For instance, the number of people in a fast food restaurant is a point process.
  - It is 9am and the doors to MacRonald's have just opened.
  - The first customer arrives after 3 minutes and is served immediately, taking a total of 4 minutes to fill the order.
  - The next customer arrives after 2 minutes and is served, taking 3 minutes.
  - The next customer arrives after another 1 minute and has to wait as only 2 servers are working.
  - The 3rd customers service takes 5 minutes.