# QUEUING MODEL (THEORY)

* 1. **Fundamental Concepts of Queuing Theory**

# Introduction

The first problem of queuing theory was raised by calls and Erlang was the first who treated congestion problems in the beginning of 20th century. His works inspired engineers,

Mathematicians to deal with queuing problems using probabilistic methods. Queuing theory became a field of applied probability and many of its results have been used in operations research, computer science, telecommunication, traffic engineering, reliability theory, and others. It analyze the shared facility needs to be accessed for service by a large number of jobs or customers. Examples: Waiting lines in cafeterias, hospitals, banks, theaters, airports etc.

# Definition of a queuing Model

A queuing model is a suitable model to represent a service oriented problem where customers arrive randomly to receive some service, the service time being also a random variable.

# Objective of a queuing model

The objective of a queuing model is to find out the optimum service rate and the number of servers so that the average cost of being in queuing system and the cost of service are minimized.

# Application of a queuing model

The queuing models are basically relevant to service oriented organizations and suggest ways and means to improve the efficiency of the service. This model can be applied in the field of business such as banks and booking counters, industries such as servicing of machines, government such as railway or post-office

counters, transportation such as airport and habour and in everyday life such as in

elevators, restaurants, hospitals among others

# Relationship between service and cost

An improvement of service level is always possible by increasing the number of employees. Apart from increasing the cost an immediate consequence of such a step is unutilized or idle time of the servers. In addition, it is unrealistic to assume that a large-scale increase in staff is possible in an organization. Queuing methodology indicates the optimal usage of existing manpower and other resources to improve the service. It can also indicate the cost implication if the existing service facility has to be improved by adding more servers.

# Arrival

The **arrival rate** is the rate at which customers arrive at the service facility during a specified period of time. For example, if 100 customers arrive at a store checkout counter during a 10-hour day, we could say the arrival rate averages 10 customers per hour.

The statistical pattern to the arrival can be indicated through

* + - 1. The probability distribution of the number of arrivals in a specific period of time
      2. The probability distribution of the time between two successive arrival (inter-arrival time)

The number of arrivals is a discrete variable whereas the inter-arrival times are continuous random and variable. A remarkable result in this context is that if the number of arrivals follows a Poisson Distribution, the corresponding inter-arrival time follows An Exponential Distribution. This property is frequently used to derive elegant results on queueing problems

# Service

The time taken by a server to complete service is known as a service time. The service time is a statistical variable and can be studied either as the number of services completed in a given period of time or the time taken to complete the

# February 16, 2016

service. The data on actual service time should be analyzed to find out the

probability distribution of service time. The number of services completed is a discrete random variable while the service time is a continuous random variable.

# Server

A server is a person or a mechanism through which service of offered. The service may be offered through a single server such as a ticket counter or through several channels such as a train arriving in a station with several platforms. Sometimes the service is to be carried out sequentially through several phrases known as multiphase service. In government, the papers move through a number of phase in terms of official hierarchy till they arrive at the appropriate level where a decision can be taken.

# Time spent in the queuing system

The time spent by a customer in a queuing system is the sum of waiting before service and service time. The waiting time of a customer is the time spent by a customer in a queuing system before the service starts. The probability distribution of waiting time depends upon the probability distribution of inter-arrival time and service time.

# Queue disciple

The queue discipline indicates the order in which members of the queue are selected for service. It is most frequently assumed that the customers are served on the first come first serve basis. T This is commonly referred to as FIFO (First In, First Out) system. Occasionally, a certain group of customers receive priority in service over the others even if they arrive late. This is commonly referred to a Priority Queue. The queue discipline does not always take into account the order of arrival. The server chooses on of the customers to offer service at random. Such a system is known as service in random order (SIRO). While allotting an item with high demand and limited supply such as a test match ticket or share of a public limited company. SIRO system is the only possible way of offering service when it sin not possible to identify the order of arrival.

# February 16, 2016

The service discipline determines t he rule according to the next customer is

selected. The most commonly used laws are

1. FIFO - First In First Out: who comes earlier leaves earlier
2. LIFO – Last Come First Out: who comes later leaves earlier
3. SIRO – Service In Random Order: the customer is selected randomly
4. PQ –Priority Queue : A category of customer are given precedence to be service earlier

# Calling population

The calling population is the source of the customers to the queuing system, and it can be either *infinite* or *finite.* An infinite calling population assumes such a large number of potential customers that it is always possible for one more customer to arrive to be served. For example, a grocery store, a bank, and a service station are assumed to have infinite calling populations; that is, the whole town or geographic area.

A finite calling population has a specific, countable number of potential customers. It is possible for all the customers to be served or waiting in line at the same time; that is, it may occur that there is not one more customer to be served. Examples of a finite calling population are a repair facility in a shop, where there is a fixed number of machines available to be worked on, a trucking terminal that services a fleet of a specific number of trucks, or a nurse assigned to attend to a specific number of patients.

# Kendall’s Notation

Kendall’s Notation is a system of notation according to which the various characteristics of a queuing model are identified. In 1951 Kendell introduced a set of notations which have become standard in the queuing models. A general queuing system is donated by: (a/b/c) : (d/e) where:

# February 16, 2016

a = probability distribution of the inter-

arrival time b = probability distribution of the service time c = number of servers in the system

d = maximum number of customers allowed in the system e = queue discipline

# Example 1

(M/M/1) : (∞/FIFO) – indicates a queuing system when the inter-arrival times and service times are exponentially distributed having one server in the system with first in first out discipline and number of customers allowed in the system can be infinite.

# Exercise 1

i. Hence M/M/1 denotes a system with Poisson arrivals, exponentially distributed service times and a single server.

ii .M/G/m denotes an m-server system with Poisson arrivals and generally distributed service times.

1. M/M/r/K/n stands for a system where the customers arrive from a finite- source with n elements where they stay for an exponentially distributed time, the service times are exponentially distributed, the service is carried out according to the request’s arrival by r severs, and the system capacity is K .
2. M/M/1/ / represents a single server that has unlimited queue capacity and infinite calling population, both arrivals and service are Poisson (or random) processes, meaning the statistical distribution of both the inter- arrival times and the service times follow the exponential distribution.
3. M/G/1/ / represents a single server that has unlimited queue capacity and infinite calling population, while the arrival is still Poisson process,

# February 16, 2016

meaning the statistical distribution of the inter-arrival times still follow the

exponential distribution, the distribution of the service time does not.

# Queuing Model or Queuing Theory Terminology

Queuing theory is the mathematical study of waiting lines (or *queues*) that enables mathematical analysis of several related processes, including arriving at the (back of the) queue, waiting in the queue, and being served by the Service Channels at the front of the queue.

**Traffic Intensity** - The ratio ג/µ is called the traffic intensity or the utilization factor and it determines the degree to which the capacity of service station is utilize

  Mean Rate of Arrival in the Queue (λ) Mean Service Rate (µ)

**Balking** - If a customer decides not to enter the queue since it is too long is called Balking **Reneging** - If a customer enters the queue but after sometimes loses patience and leaves it is called Reneging

**Jockeying** - When there are 2 or more parallel queues and the customers move from one queue to another is called Jockeying

**Waiting Time Cost** - The cost of waiting customers include either the indirect cost of lost business or direct cost of idle equipment and persons.

**Idle Time Cost -** The cost of idle service facilities is the payment to be made to the servers for the period for which they remain idle.

**Transient State of a system -** Queuing analysis involves the system’s behavior over time. If the Operating characteristics vary with time then it is said to be transient state of the system.

**Steady state of a system -** If the behavior becomes independent of its initial conditions (no. of customers in the system) and of the elapsed time is called Steady State condition of the system

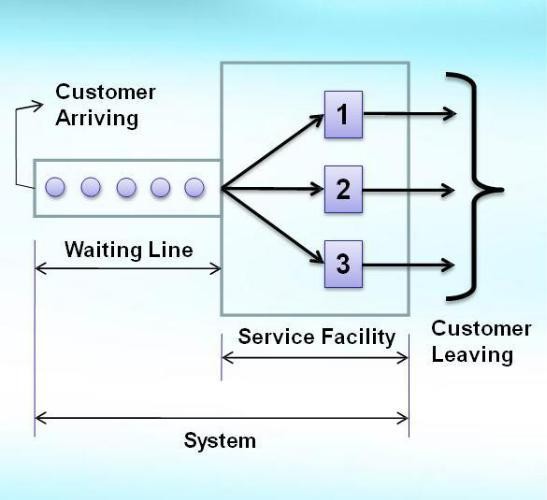
# Queuing Model Application

Queuing Model can be applied to various situations:

1. Where customers are involved such as restaurants, café, super market, airports etc.
2. Very useful in manufacturing units
3. Application for the problem of machine breakdown & repairs
4. Application for the scheduling of jobs in production control
5. Application for the minimization of traffic congestion at tollbooth
6. Provide solution of inventory control problems

# Major Constituents of Queuing System

1. Customer
2. Queue
3. Service Channel



# Figure 2: Major constituents of a queuing system

* + 1. **Assumption in queuing system**

1. The customers arrive for service at a single service facility at random according to Poisson distribution with mean arrival rate ג.

The service time has exponential distribution with mean service rate µ.

1. The service discipline followed is First Come First Served.
2. Customer Behavior is Normal
3. Service facility behavior is Normal
4. The calling source has infinite size
5. The mean arrival rate is less than the mean service rate
6. The waiting space available for customer in the queue is infinite

# Limitations of queuing model

1. The waiting space for the customer is usually limited
2. The arrival rate may be state dependent
3. The arrival process may not be stationary
4. The population of customers may not be infinite and the queuing discipline may not be First Come First Serve
5. Services may not be rendered continuously
6. The Queuing system may not have reached the steady state. It may be, instead, in transient state

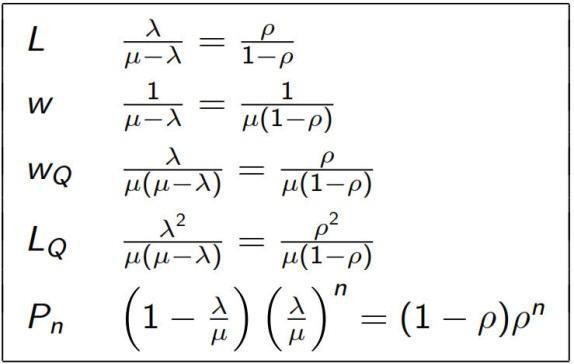
# Queuing Theory Equations

The following Measures of Performance of Queuing Systems are represented as follows:

1. Average number of customers in system (L)
2. Average number of customers in queue (LQ)
3. Average time spent in system per customer (w)
4. Average time spent in queue per customer (wQ)
5. Server utilization (ρ)

# Single-Server Queues (Poisson Arrivals & Infinite Capacity)

**Table 19: Steady-State Parameters of the M /G /1 Queue**



# Single-Server Queues (Poisson Arrivals & Infinite Capacity) Example 1

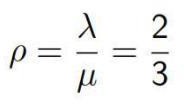
Assume that time between arrivals and service times at a single-chair unisex hair- styling shop have been shown to be exponentially distributed. The values of λ and µ are 2 per hour and 3 per hour, respectively. Computer the following

a ) The server utilization

1. The probabilities of having 0, 1, 2 and 3 or more customers in the shop
2. The number of customers in system, in queue and their waiting times

# Solution

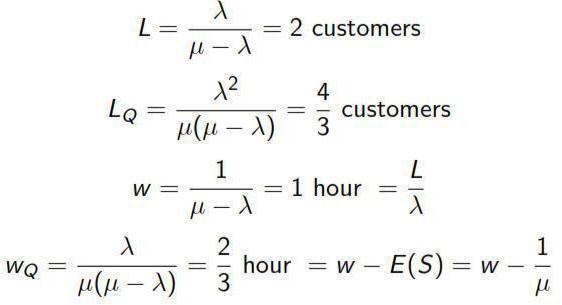
**The server utilization is**



# The probabilities of having 0, 1, 2 and 3 or more customers

−

# The number of customers in system, in queue and their waiting times

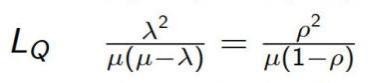


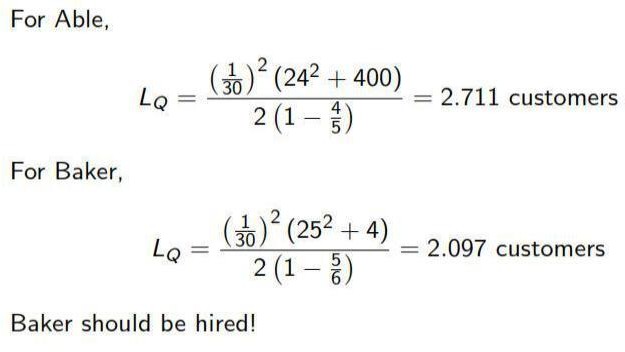
* + 1. **Single-Server Queues (Poisson Arrivals & Infinite Capacity) Example 2**

There are two workers competing for a job; Able and Joy. Able claims an average service time that is faster than Joy’s, but Jo

y claims to be more consistent, even if not as fast. The arrivals occur according to a Poisson process with a rate of λ = 2 per hour. Able’s statistics are an average service time of 24 minutes with a standard deviation of 20 minutes. Joy’s statistics are an average service time of 25 minutes with a standard deviation of 2 minutes. If the average queue length is the criterion for hiring, which worker should be hired?

# Solution



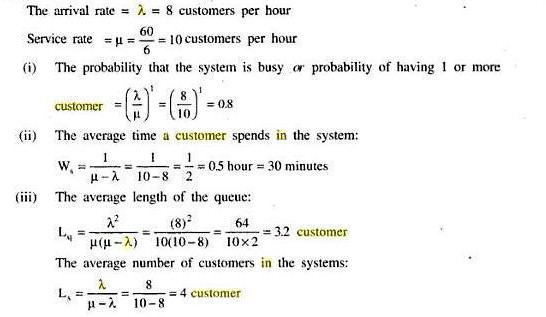


**Exercise 1**

Kenya Airways has one reservation clerk on duty in its Moi International Airport branch at any given time. The clerk handles information regarding passenger reservation and flight timings. Assume that the number of customers arriving during any given period is passion distribution with and arrival rate 8 per hour and that the clerk can service 1 customer in 6 minutes on an average, with an exponentially distributed service time.

* + - 1. What is the probability that the system is busy?
      2. What is the average time a customer spends in the system?
      3. What is the average length of the queue?
      4. What is the average number of the customers in the system?

# Solution



**Exercise 2**

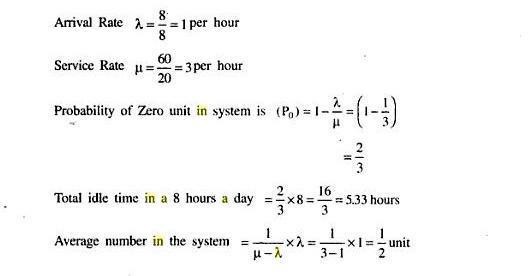
A Jua Kali mechanic finds that the time spent on his job has an exponential distribution with an average of 20 minutes. If repairs cars in the order in which they arrive in, and if the arrival is approximately Poisson with an average rate of 8 cars per 8 hours in a day.

1. What is the mechanic’s expected idle time each day?

**ICS 2307 SIMULATION AND MODELLING**

1. How many jobs are on an average in the system?

# Solution



**Exercise 3**

Cinema goers arrive at the Nyali Cinemax ticket’s counter at a rate of 12 per hour. There is one clerk serving the customers at a rate of 30 per hour.

1. What is the probability that there are no customers in the counter? (The system is idle)
2. What is the probability that there are more than 2 customers in the counter?
3. What is the probability that there are no customers waiting to be served?
4. What is the probability that a customer is being served and nobody waiting?

