INTRODUCTION

Queuing theory is a branch of mathematics that studies and models the act of waiting in lines. The origin of queuing theory dates back to 1909, when Argner Krarup Erlang (1878-1929) published his fundamental paper on congestion in telephone traffic in addition to formulating in analytic from several practical problems arising in telephony and solving them. Erlang laid solid foundation for queuing theory in terms of the nature assumptions and techniques of analysis, these are being routinely used to this day even in wider areas of modern communications and computer systems.In a way Erlang was pioneer in the applications of analytic methods to operational problems. His studies appear to mark the beginning of the study of operations research.

Kendall was the pioneer who viewed and developed queuing theory from the perspective of stochastic process. Queuing theory is generally considered as a branch of operations research because the results are often used when making business decisions about the resources needed to provide a service. The objective of queuing analysis is to offer a reasonably satisfactory service to waiting customers. It determines the measure of performance of waiting lines, such as the average waiting time in queue and the productivity of the service facility, which

can then be used to design the service installation.

Waiting for a service is a part of our daily life. We wait to eat in restaurants, we queue up at the check out counters in grocery sores and we line up for service in post offices. The waiting phenomenon is not an experienced limited to human beings only. Jobs wait to be processed on a machine, planes circle in a stack before given permission to land an airport and cars stop at traffic lights. Waiting cannot be eliminated completely without incurring in ordinate expenses and the goal is to reduce its adverse impact to tolerate levels.

This paper will take a brief look into the formulation of queuing theory along with examples of the models and applications of their use. A basic queuing system consists of an arrival process (how customers arrive at the queue, how many customers are present in total), the queue itself, the service process for attending to those customers, and departures from the system. Mathematical queuing models are often used in software and business to determine the best way of using limited resources.

Outline of the Project :Apart from the introductory chapter, we have described our work in four chapters

- Chapter 1 : Covers the necessary concepts of statistics, exponential and poisson distributaion and lack of memory property which are used to calculate the formulas for models.
- Chapter 2 : Introduces the queuing system, its components, its notations, transient and steady states, traffic intensity and its characteristics. Also we discusses the probability distributions in queuing theory, like arrivals,

inter, departure, service time ans its examples.

- Chapter 3 : Here, we introduce the Kendall's notation of queuing model, classifications of models and we define Model 1.

- Chapter 4 : This chapter, give some information on the contribution and applications of queuing theory in the field of healthcare.

Chapter 1

PRELIMINARIES

To begin understanding queues, we must first have some knowledge of probability theory. In particular, we will review the Exponential and Poisson probability distributions.

1.1 Poisson Distribution

A Poisson queue is a queuing model in which the number of arrivals per unit of time and the number of completions of service per unit of time, when there are customers waiting, both have the Poisson distribution. It is good to use if the arrivals are all random and independent of each other. For the Poisson distribution, the probability that there are exactly *x* arrivals during *t* amount of time is:

$$Prob(x) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}$$

Where t is a duration of time. Its units are, hours or days. λ is the expected (average) number of arrivals per hour, or day, or whatever units t is measured in. So, λt is therefore the expected number of arrivals during t amount of time. x is a possible number of arriving customers [1].

1.2 Exponential Distribution

In the most queuing situations, arrival of customers occurs in a totally random fashion. Random inter arrival and service time are described quantitatively in queuing models by the exponential distribution.

Exponential distribution is defined as,

$$f(t) = \lambda e^{-\lambda t}; t > 0$$

Then equivalently its probability density function is given by ,

$$f(t) = \lambda e^{-\lambda t}; t > 0$$

1.3 Lack of Memory Property

Another important property is forgetfulness or lack of memory. It suggests that the time until the next arrival will never depend on how much time has already passed.

Let the exponential distribution $f(t)$, represents the time t, between successive

events. If s is the interval. Then,

$$P\{t > T + s/t > s\} = P\{t > T\}$$

$$P\{t > T + s/t > s\} = \frac{P\{t > T + s, t > s\}}{P\{t > s\}}$$

$$= \frac{P\{t > T + s\}}{P\{t > s\}}$$

$$= \frac{e^{-\lambda(T+s)}}{e^{-\lambda s}}$$

$$= e^{-\lambda T}$$

$$= P\{t > T\}$$

There is a relation between exponential and poisson distributions. The poisson distribution is need to determine the probability of certain number of arrivals occurring in a given time period. The poisson distribution with parameter λ is given by ,

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

If we set $n = 0$, the poisson distribution gives us $e^{-\lambda t}$ which is equal to $P\{t > T\}$ form exponential distribution. With these distributions in mind, we can begin defining poisson process, waiting distribution, birth-death process from which we can develop the model [4].

Chapter 2

THE QUEUING SYSTEM

2.1 Characteristics of Queueing System

Key elements of queueing systems are,

- Customer: Refers to anything that arrives at a facility and requires service. Eg: people, machines, trucks, emails, packets, frames.
- Server: Refers to any resource that provides the requested service. Eg: repair persons, machines, runaways at airport, host, switch, router, disk drive, algorithm.

Systems	Customers	Server
Reception desk	People	Receptionist
Hospital	Patients	Doctors
Airport	Airplanes	Runway
Production line	Cases	Case-packer
Road network	Vehicles	Traffic light
Grocery	Shoppers	Checkout station
Computer	Jobs	CPU, disk, CD
Network	Packets	Router

Also, a system may have one or more than one servers. So a queuing system can be classified according to the number of servers as *"Single server queuing system"*, if there is only one server and *"Multi server queuing system"* if there are multiple servers. Generally, the behaviour of queue is charcterized by following parameters;

- Arrival process
- Service (departure) process
- Number of servers in the system
- Queuing discipline
- Capacity of the queue
- Size of queue

2.2 Components of Queuing System

A queuing system is composed of the following components (or parts)

(1) The Input(Arrival Pattern)

(2) The Service Mechanism

(3) The Queue Discipline

(4) Customers Behavior

(5) Capacity Of The System

2.2.1 The Input (Arrival Pattern)

Customers arriving to the system for a service will directly go to the service station without waiting in the queue if the server is free at that point of time. Otherwise he will wait in the queue till the server becomes free. Generally the customers arrival is unpredictatble. So the arrival pattern can be computed in terms of the probabilities.

2.2.2 The Service Mechanism

This includes the distribution of time of service to a customer, the number of servers and arrangement of servers (parallel or series, etc). If the number of servers is more than one, then this queue is an example of parallel counters for providing service. The system is said to be "*Tandem*" if the service to be

provided in multistage in sequential order. Service time is a random variable with the same distributions for all the arrivals.

2.2.3 The Queue Discipline

This is the manner in which customers form a queue and the manner in which they are chosen for service.

The simplest discipline is *"First Come First Served"* (FCFS), according to which the customers are served in the order of their arrivals. For example such type of queue discipline is observed at reservation counters, at bank counters etc.

If the last arrival get served first, we have *"Last Come First Served"* (LCFS) queue discipline. This is observed in government offices, where the file in which comes on the table last get cleared first. The other queue discipline are *"Random selection"* or *"Selection at Random Order"* (SIRO) and *"Priority Selection"* [2].

2.2.4 Customers Behaviour

Generally, it is assumed that the arrivals in the system are one by one. But in practice, customers may arrive in groups, such arrivals are called **Bulk arrivals**. The customers behave in the following ways;

- **Balking**

  On arrival a customer find the queue length very long and he may not join the queue. This phenomenon is known as Balking of customers.

- **Jockeying**

  If there is more than one queue, the customer from one queue may shift

to another queue because of its smaller size. This behaviour of customers known as Jockeying.

- **Reneging**

  A customer who is already in the queue leaves the queue due to long waiting line. This kind of departure from queue without receiving the service is called Reneging.

2.2.5 The Capacity Of The System

A system may have an infinity capacity, that is, the queue infront of the servers may grow to any length. Against this there may be limitations of space. So that when the space is filled to capacity, an arrival will not able to join the system and will be lost the system. The system is called "*delay system*" or a "*loss system*" according to whether the capacity if finite or infinite.

- **Queue Size**

  The total number of customers in the system who are actually waiting in the line and not being serviced.

- **Queue Length**

  Queue length may be defined as the line length plus number of customers being served.

- **Queuing Model**

  A queueing model is a mathematical description of a queuing system which makes some specific assumptions about the probabilistic nature of the ar-

rival and service processes, the number and type of servers, and the queue discipline and organization [5].

Some Notations Used In Queueing Theory

n = Number of customers (units) in the system

$P_n(t)$ = Transient state probability of exactly n units in the queuing system at time t

$p_n(t)$ = Steady state probability of exactly n units in the queuing system at time t

λ_n = Mean arrival rate per unit of time , when there are n units in the system

μ_n = Mean service rate per unit of time , when there are n units in the system

λ = Constant mean arrival rate for all n steady state

μ = Constant mean service rate

SorR= Number of parallel service places (parallel server)

$\rho = \frac{\lambda}{\mu}$ = Traffic intensity as utilization factor

Transient and Steady States

A system is said to be in "*Transient state*" when its operating characteristics are dependent on time. A "Steady state" system is the one in which the behavior of the system is independent of time. Let $P_n(t)$ denote the probability that there are n customers in the system, at time t. Then in steady state

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

$$\frac{dP_n(t)}{dt} = \frac{dP_n}{dt}$$

$$\lim_{t \to \infty} P_n(t) = 0$$

Traffic Intensity (Utilization factor)

A important measure of simple queue is its traffic intensity and is given by ,

$$\rho = \frac{\lambda}{\mu}$$

where ,

λ = the average customers arrival rate

μ = the average service rate:

The unit of traffic intensity is Erlang.

2.3 Examples

- Single server queuing system

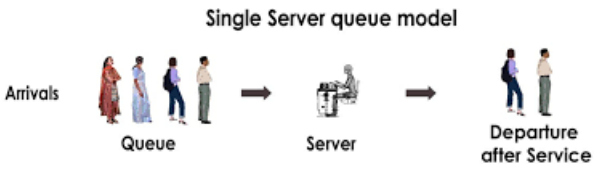

Figure 2.1: Eg:1

2.3. Examples

- Multi server queuing system

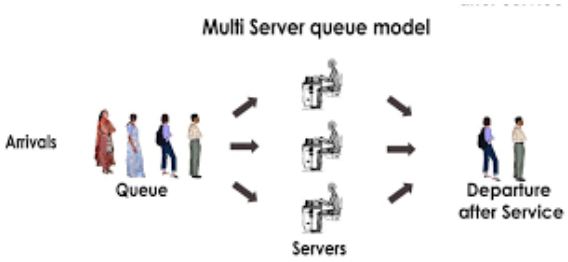

Figure 2.2: Eg:2

- Queuing system of finite population ; cars parked in a garage waiting to repair

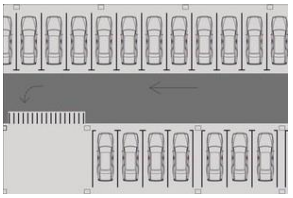

Figure 2.3: Eg:3

- Queuing system of infinite population ; boxes waiting to packed in a factory

Figure 2.4: Eg:4

2.4 Probability Distributions In Queueing Theory

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. It implies that the probability of service completion in any short term period is constant and independent of the length of the 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 the queuing problems.

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

- The probability of an event (arrival or departure) occurring during the time interval $(t, t + \Delta t)$ depends on the length of time interval Δt.
- The probability of more than one event occurring during the time interval $(t, t + \Delta t)$ is negligible. It is denoted by $O(\Delta t)$.
- Atmost one event can occur during a small time interval Δt. The probability of an arrival during the time interval $(t, t + \Delta t)$ given by

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

where λ is a constant and independent of the total number of arrivals up to time t ; Δt is a small time interval and $O(\Delta t)$ represents the quantity

that becomes negligible when compared to Δt as $\Delta t \to 0$

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

2.4.1 Distribution of Arrivals: Pure Birth Process

Even though the arrival pattern of the customers varies from one system to another and it is random too, mathematically, we show that the arrival have a poisson distribution. The model in which only arrivals are counted and no departure is takes place is called "*Pure birth model*".

We wish to derive the probability of n arrivals in time t. Denote it by $P_n(t)$, $(n \geq 0)$. The difference-differential equations governing the process in two different situations are as follows

Case 1:

For $n > 0$ there are two mutually exhaustive events of having n units at time $(t + \Delta t)$ in the system.

- There are n units in the system at time t and no arrival takes place during time interval Δt. So at time $(t + \Delta t)$ there will be n units in the system. Therefore, the probability of these two combined events will be

  = probability of number of units at time $t \times$ probability of number of arrivals during Δt

  $= P_n(t)(1 - \lambda \Delta t)$

- There are $(n - 1)$ units in the system at time t and one arrival takes place during time interval Δt. So at time $(t + \Delta t)$, there will be n units in the

system.

Therefore, probability of these two combined events will be

= probability of number of units at time t $\times$ probability of number of arrivals during Δt

$= P_{n-1}(t)\lambda\Delta t$

Adding the above two probability, we get probability of n arrivals at time $(t+\Delta t)$ as

$$P_n(t + \Delta t) = P_n(t)(1 - \lambda\Delta t) + P_{n-1}(t)\lambda\Delta t \tag{2.1}$$

Case 2:

When $n = 0$, i.e, there is no customers in the system.

Then,

$P_0(t + \Delta t)$ = probability of number of units at time $t\times$ probability of number of arrivals during Δt

$$P_0(t + \Delta t) = P_0(t)(1 - \lambda\Delta t) \tag{2.2}$$

Then,

$$\begin{aligned} P_0(t + \Delta t) &= P_0(t)(1 - \lambda\Delta t) \\ &= P_0(t) - \lambda P_0(t)\Delta t \\ P_0(t + \Delta t) - P_0(t) &= -\lambda P_0(t)\Delta t \end{aligned}$$

Dividing throughout by Δt and applying limit on both sides,

$$\lim_{\Delta t \to 0} \frac{P_0(t + \Delta t) - P_0(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{-\lambda P_0(t)\Delta t}{\Delta t}$$

$$P_0'(t) = \lim_{\Delta t \to 0} -\lambda P_0(t)$$

$$= -\lambda P_0(t)$$

ie,

$$\frac{P_0'(t)}{P_0(t)} = -\lambda \tag{2.3}$$

Integrating both sides with respect to t,

$$\log P_0(t) = -\lambda t + A \tag{2.4}$$

Where A is a constant of integration. Its value can be computed using the boundary conditions

$$P_n(t) = \begin{cases} 1 & \text{if } n = 0; \\ 0 & \text{if } n > 0. \end{cases} \tag{2.5}$$

Then $P_0(0) = 1$ and also put $t = 0$ in (2.4),

$$\Rightarrow \log 1 = 0 + A$$

$$\Rightarrow 0 = A$$

Therefore, from (2.4) we get,

$$\log P_0(t) = -\lambda t$$

$$\Rightarrow P_0(t) = e^{-\lambda t}$$

From (2.1),

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

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

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

Dividing throughout by Δt and applying limit on both sides,

$$\Rightarrow \lim_{\Delta t \to 0} \frac{P_n(t + \Delta t) - P_n(t)}{\Delta t} = \lim_{\Delta t \to 0} \{-\lambda P_n(t)\Delta t + P_{n-1}(t)\Delta t\}$$

$$\Rightarrow P_n'(t) = -\lambda P_n(t) + \lambda P_{n-1}(t)$$

Put $n = 1$, we get,

$$P_1'(t) = -\lambda P_1(t) + \lambda P_0(t)$$

$$= -\lambda P_1(t) + \lambda e^{-\lambda t}$$

Which is a linear differential equation of first order. Its solution is

$$e^{\lambda t} P_1(t) = \lambda t + B \qquad (2.6)$$

Using (2.5), we get $B = 0$.

ie,

$$e^{\lambda t}P_1(t) = \lambda t \tag{2.7}$$

Thus, (2.7) can be rewritten as,

$$P_1(t) = \lambda t e^{-\lambda t}$$

Arguing as above, we get,

$$P_2(t) = \frac{(\lambda t)^2}{2!}e^{-\lambda t}$$

Continuing this process we get,

$$P_n(t) = \frac{(\lambda t)^n}{n!}e^{-\lambda t} \tag{2.8}$$

which is a poisson distribution.

2.4.2 Distribution of Inter-Arrival Time

The time T between two consecutive arrival is called Inter arrival time. Here, mathematical development is given to show that T (Inter-arrival time) follows negative exponential law.

Proof

Let $f(T)$ be the probability density function of arrivals in time T. Then we show