QUEUING THEORY

Project report submitted to Christ College (Autonomous) in partial fulfilment of the requirement for the award of the M.Sc Degree programme in Mathematics

by

APARNA DEVAN

Register No. CCATMMS007

Department of Mathematics

Christ College (Autonomous)

Irinjalakuda

2021

CERTIFICATE

This is to certify that the project entitled **"QUEUING THEORY "** submitted to Department of Mathematics in partial fulfilment of the requirement for the award of the M.Sc Degree programme in Mathematics, is a bonafide record of original research work done by **Ms. APARNA DEVAN (CCATMMS007)** during the period of her study in the Department of Mathematics, Christ College (Autonomous), Irinjalakuda, under my supervision and guidance during the year 2019-2021.

Dr. JOJU K T
Associate Professor(Retired)
Department of Mathematics
Christ College(Autonomous)
Irinjalakuda

Dr. Seena V
HoD in Charge
Department of Mathematics
Christ College(Autonomous)
Irinjalakuda

External Examiner :

Place : Irinjalakuda
Date : 31 March 2021

DECLARATION

I hereby declare that the project work entitled "**QUEUING THEORY**" submitted to Christ College(Autonomous), Irinjalakuda in partial fulfilment of the requirement for the award of Master Degree of Science in Mathematics is a record of original project work done by me during the period of my study in the Department of Mathematics, Christ College(Autonomous), Irinjalakuda.

Place : Irinjalakuda **APARNA DEVAN**

Date : 31 March 2021

ACKNOWLEDGEMENT

First, there are no words to adequately acknowledge the wonderful grace that my Redeemer has given me. There are many individuals who have come together to make this project a reality. I greatly appreciate the inspiration; support and guidance of all those people who have been instrumental for making this project a success.

I express my deepest thanks to my guide Dr. JOJU K T, Associate Professor(Retired), Department of Mathematics, Christ College(Autonomous), Irinjalakuda, who guided me faithfully through this entire project. I have learned so much from him, both in the subject and otherwise. Without his advice, support and guidance, it find difficult to complete this work.

I take this opportunity to express my thanks to our beloved principal Fr. Dr. JOLLY ANDREWS CMI, who gave me the golden opportunity to do this wonderful project on the topic "**QUEUING THEORY**" .

I mark my word of gratitude to Dr. SEENA V, HoD in Charge and all other teachers of the department for providing me the necessary facilities to complete this project on time.

My project guide Dr. Seena V, deserves a special word of thanks for her invaluable and generous help in preparing this project in $LATEX$.

I want to especially thank all the faculty of the library for providing various

facilities for this project.

Words cannot express the love and support I have received from my parents, whose encouragement has buoyed me up from the beginning till the end of this work.

Aparna Devan

Contents

List of Figures

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,

$$\begin{aligned} P\{t > T + s/t > s\} &= P\{t > T\} \\ P\{t > T + s/t > s\} &= \frac{P\{t > T + s, t > s\}}{P\{t > s\}} \\ &= \frac{P\{t > T + s\}}{P\{t > s\}} \\ &= \frac{e^{-\lambda(T+s)}}{e^{-\lambda s}} \\ &= e^{-\lambda T} \\ &= P\{t > T\} \end{aligned}$$

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

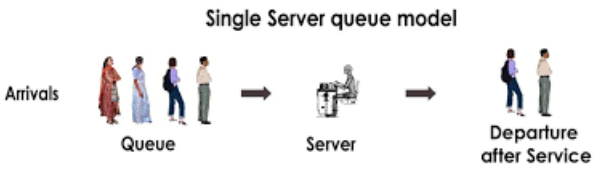

Figure 2.1: Eg:1

- Multi server queuing system

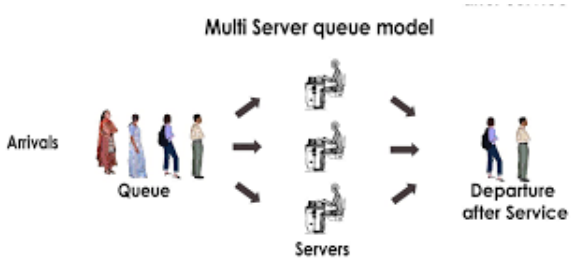

Figure 2.2: Eg:2

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

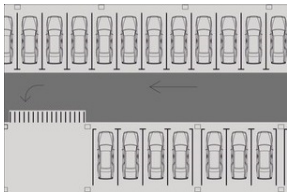

Figure 2.3: Eg:3

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

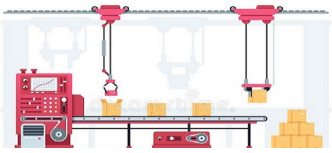

Figure 2.4: Eg:4

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 $\Delta t \to 0$

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

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,

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

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

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

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),

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

$$\implies 0 = A$$

Therefore, from (2.4) we get,

$$\begin{aligned} \log P_0(t) &= -\lambda t \\ \implies P_0(t) &= e^{-\lambda t} \end{aligned}$$

From (2.1),

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

Dividing throughout by Δt and applying limit on both sides,

$$\begin{aligned} \implies \lim_{\Delta t\to 0}\left\{\frac{P_n(t+\Delta t) - P_n(t)}{\Delta t}\right\} &= \lim_{\Delta t\to 0}\{-\lambda P_n(t)\Delta t + P_{n-1}(t)\Delta t\} \\ \implies P_n^{'}(t) &= -\lambda P_n(t) + \lambda P_{n-1}(t) \end{aligned}$$

Put $n = 1$, we get,

$$\begin{aligned} P_1^{'}(t) &= -\lambda P_1(t) + \lambda P_0(t) \\ &= -\lambda P_1(t) + \lambda e^{-\lambda t} \end{aligned}$$

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

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

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

that

$$f(T) = \lambda e^{\lambda t}$$

where λ is an arrival rate in time t.

Let t be the instant of an arrival. Since there is no arrival during $(t + \Delta t)$ and $(t + T, t + T + \Delta T)$ arrival be at $t + T + \Delta T$.

Putting $n = 0$ and $t = T + \Delta T$ in equation (2.8)

$$\begin{aligned} P_0(T + \Delta T) &= e^{-\lambda T + \Delta T} \\ &= e^{-\lambda T}.e^{\Delta T} \\ &= e^{-\lambda t}[1 - \lambda \Delta T + O(\Delta T)] \end{aligned}$$

But $P_0(T) = e^{-\lambda T}$. So, we get

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

Dividing both sides by ΔT and talking limit $\Delta T \to 0$, we get,

$$P_0'(T) = -\lambda P_0(T) \tag{2.9}$$

Clearly LHS of equation (2.9) is probability density function of T, say $f(T)$. Therefore ,

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

which is negative exponential law of probability for T.

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 independant of how long he has already been in service.

ie, $Prob(T \geq t_1/T \geq t_0) = Prob(0 \leq T \leq t_1 - t_0)$

where T is the time between successive arrivals.

2.4.3 Distribution of Departure (Pure death process)

The model in which only departures occur and no arrival takes place is called "*pure death process*". In this process, assume that there are N customers in the system at $t = 0$, no arrivals occur in the system, and departures occure at a rate μ per unit time. We will derive the distribution of departures from the system on the basis of following three axioms;

- The probability of exactly one departure during small interval Δt be given by $\mu\Delta t + O(\Delta t)$
- The term Δt is so small that the probability of more than one departure in time Δt is negligible.

- The number of departures in non-overlapping intervals is mutually independant.

The following three cases arises :

Case 1:

When $0 < n < N (1 \leq n \leq N-1)$, the probability will be

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

Case 2:

When $n = N$, that is there are N customers in the system.

Then

$$P_N(t+\Delta t) = P_N(t)(1-\mu\Delta t)$$

Case 3 :

When $n = 0$, that is there is no customer in the system,

$$P_0(t+\Delta t) = P_0(t) + P_1\mu\Delta t$$

Thus,

$$\begin{aligned}
P_0(t+\Delta t) &= P_0(t) + P_1(t)\mu\Delta t, \quad n = 0 \\
P_n(t+\Delta t) &= P_n(t)(1-\mu\Delta t) + P_{n+1}(t)\mu\Delta t, \quad 0 < n < N \\
P_N(t+\Delta t) &= P_N(t)(1-\mu\Delta t), \quad n = N
\end{aligned}$$

Rearranging the above equations, dividing by Δt and taking limit $\Delta t \to 0$, we get,

$$P_0'(t) = P_1(t)\mu, \;\; n = 0 \tag{2.10}$$

$$P_n'(t) = -P_n(t)\mu + P_{n+1}(t)\mu, \;\; 0 < n < N \tag{2.11}$$

$$P_N'(t) = -P_N(t)\mu, \;\; n = N \tag{2.12}$$

The above three equations are the required system of differential- difference for pure death process.

The solution of (2.12) can be written as,

$$logP_N(t) = \mu t + A$$

where A is constant of integration. Its value can be computed using the boundary conditions;

$$P_n(0) = \begin{cases} 1 & \text{if } n = N \neq 0 \\ 0 & \text{if } n \neq N. \end{cases}$$

Which gives A = 0. Therefore,

$$P_N(t) = e^{-\mu t} \tag{2.13}$$

Putting $n = N - 1$ in (2.11), we get,

$$\begin{aligned} P'_{N-1} &= -P_{N-1}(t)\mu + P_N(t)\mu \\ &= -P_{N-1}(t)\mu + e^{-\mu t}\mu \\ P'_{N-1}(t) + \mu P_{N-1}(t) &= \mu e^{-\mu t} \end{aligned}$$

which is linear differential equation of first order. Its solution is,

$$P_{N-1}(t) = \mu t e^{-\mu t}$$

Putting $n = N-2, N-3, \ldots, N-n$ in (2.11) we get,

$$P_{n-k}(t) = \frac{(\mu t)^k}{(k)!} e^{-\mu t}$$

In general ,

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

which is a poisson distribution.

2.4.4 Distribution of Service time

Let T be the random variable denoting the service time and t be the possible value of T. Let $S(t)$ and $s(t)$ be the cumulative distribution function and the probability density function of T respectively. To find $s(t)$ for the poisson departure case, it can be observed that the probability of no service during time $(0, t)$

which means the probability of having no departures during the same period. Thus,

$$\begin{aligned} P(\textit{Service time } T \geq t) &= P(\textit{no departure during } t) \\ &= P_N(t) \end{aligned}$$

where there are N units in the system and no arrival is allowed after N. Therefore,

$$P_N(t) = e^{-\mu t}$$

So,

$$\begin{aligned} S(t) &= P(\textit{Service time } T \leq t) \\ &= 1 - P(\textit{Service time } T \geq t) \\ &= 1 - e^{-\mu t} \end{aligned}$$

differentiating both sides with respect to t,we have

$$s(t) = \begin{cases} \mu e^{-\mu t} & \text{if } t \geq 0; \\ 0 & \text{if } t < 0. \end{cases}$$

which is an exponential distribution.

Chapter 3

QUEUING MODELS

3.1 Kendall's Notation For Representing Queuing Model

Using Kendall's notation,the queuing model can be defined by,

(a/b/c); (d/e/f)

a = Arrival distribution

b = Departure distribution (Server Time)

c = Number of parallel server

d = Queue discipline FCFS , LCFS , SIRO

e = Maximum number (finite or infinite)

f = Size of calling source (finite or infinite)

Standard notations for representing the arrivals or departure (a or b)

M = Markovian (poisson) for arrival or departure

D = Constant deterministic time

E_k= Erlang or Gamma distribution

G = General service time

GI = General distribution of inter arrival time

GS = General distribution of service time

3.2 Classification of Queuing Models

There are various deterministic and probabilistic queuing models. Some of the probabilistic models are given.

- Model 1 : $(M/M/1);(\infty/FCFS/1)$
- Model 2 : $(M/M/1);(N/FCFS/\infty)$
- Model 3 : $(M/M/C);(\infty/FCFS/\infty)$

3.3 Model 1 : $(M/M/1);(\infty/FCFS/\infty)$

The model is known as "*birth and death*" model, which deals with a queuing system having a single server, poisson arrival, exponential service and there is

no limit on the system capacity while the customers are served on a "first come first served" basis.

Formulation of difference- differential equations:

Let $P_n(t)$ denote the probability of n customers in the system at time t. Then the probability that the system has n customers at time $(t + \Delta t)$ may be expressed as the sum of combined probability of following four mutually exclusive and exhaustive events:

$$\begin{aligned}
P_n(t+\Delta t) &= P_n(t) \times P(\textit{no arrival in } \Delta t) \times P(\textit{no service completion in } \Delta t) \\
&\quad + P_n(t) \times P(\textit{one arrival in } \Delta t) \times P(\textit{one service completion in } \Delta t) \\
&\quad + P_{n+1}(t) \times P(\textit{no arrival in } \Delta t) \times P(\textit{one service completion in } \Delta t) \\
&\quad + P_{n-1}(t) \times P(\textit{one arrival in } \Delta t) \times P(\textit{noservicecompletionin}\Delta t) \\
&\quad = P_n(t)(1-\lambda\Delta t)(1-\mu\Delta t) + P_n(t)(\lambda\Delta t)(\mu\Delta t) \\
&\quad + P_{n+1}(t)(1-\lambda\Delta t)(\mu\Delta t) + P_{n-1}(t)(\lambda\Delta t)(1-\mu\Delta t) \\
&\quad = P_n(t)(1-\lambda\Delta t)(1-\mu\Delta t) + P_{n+1}(t)(1-\lambda\Delta t)(\mu\Delta t) + P_{n-1}(t)(\lambda\Delta t)(1-\mu\Delta t)
\end{aligned}$$

or,

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

ie,

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

Dividing by Δt and taking limit $\Delta t \to 0$, we get,

$$0 = -(\lambda+\mu)P_n(t) + \mu P_{n+1}(t) + \lambda P_{n-1}(t), \;\; n \geq 1 \tag{3.1}$$

(By Steady state condition LHS will be zero, since in steady state state behaviour of the system is independent of time)

If there is no customer in the system at time $(t + \Delta t)$, there will be no service during Δt. Then for $n = 0$,

$$\begin{aligned} P_0(t+\Delta t) &= P_0(t) \times P(\textit{no arrival in } \Delta t) \times P(\textit{no service completion in } \Delta t) \\ &\quad + P_1(t) \times P(\textit{no arrival in } \Delta t) \times P(\textit{one service completion in } \Delta t) \\ &= P_0(t)(1-\lambda\Delta t) + P_1(t)(1-\lambda\Delta t)(\mu\Delta t) \end{aligned}$$

(Since there is no customers, probability of service completion is one)

or,

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

Dividing by Δt and taking limit $\Delta t \to 0$, we get,

$$0 = -\lambda P_0(t) + \mu P_1(t), \;\; n = 0 \tag{3.2}$$

(By Steady state condition, LHS will be zero)

Therefore,

Equation (3.2) gives, $\quad P_1 = (\frac{\lambda}{\mu})P_0$

Put n=1 in equation (3.1), we get, $\quad P_2 = (\frac{\lambda}{\mu})^2 P_0$

In general, $\quad P_n = (\frac{\lambda}{\mu})^n P_0 = \rho^n P_0$

To obtain the value of P_0,we proceed as follows:

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

$$
\begin{aligned}
&= \sum_{n=0}^{\infty} \rho^n P_0 \\
&= P_0 \sum_{n=0}^{\infty} \rho^n \\
&= \frac{P_0}{1-\rho}
\end{aligned}
$$

or,

$$P_0 = 1 - \rho$$

and

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

Chapter 4

QUEUING THEORY APPLICATIONS IN HEALTHCARE

Healthcare is riddled with delays. Almost all of us have waited for days or weeks to get an appointment with a physician or schedule a procedure, and upon arrival we wait some more until being seen. In hospitals, it is not unusual to find patients waiting for beds in hallways, and delays for surgery or diagnostic tests are common.

Delays are the result of a disparity between demand for a service and the capacity available to meet that demand. Usually this mismatch is temporary and due to natural variability in the timing of demands and in the duration of time needed to provide service. A simple example would be a healthcare clinic where patients walk in without appointments in an unpredictable fashion and

require anything from a flu shot to the setting of a broken limb. This variability and the interaction between the arrival and service processes make the dynamics of service systems very complex. Consequently, it's impossible to predict levels of congestion or to determine how much capacity is needed to achieve some desired level of performance without the help of a queueing model.

Now, we discuss the overview of research using queuing theory as an analytical tool to predict how particular healthcare configurations affect delay in patient service and healthcare resource utilization. Also we look at applications to appointment scheduling where the main challenge is reducing patient waiting without greatly increasing server idleness.

Waiting Time and Utilization Analysis

In a queuing system, minimizing the time that customers (in healthcare, patients) have to wait and maximizing the utilization of the servers or resources (in healthcare, doctors, nurses, hospital beds, e.g.) are conflicting goals.

When a patient is waiting in a queue, he may decide to forgo the service because he does not wish to wait any longer. This is an important characteristic of many healthcare systems. The probability that a patient reneges usually increases with the queue length and the patient's estimate of how long he must wait to be served. It is possible to redesign a queuing system to reduce reneging. A common approach is to separate patients by the type of service required. Find the number of patients who leave an emergency department without being served is reduced by separating non-acute patients and treating them in dedi-

cated fast-track areas. Most of their waiting would be for tests or test results after having first seen a doctor. The paper also estimates the size of the waiting area for patients and those accompanying them.

Most analytical queuing models assume a constant customer arrival rate, many healthcare systems have a variable arrival rate. In some cases, the arrival rate may depend upon time but be independent of the system state. For instance, arrival rates change due to the time of day, the day of the week, or the season of the year. In other cases, the arrival rate depends upon the state of the system.

In most healthcare settings, unless an appointment system is in place, the queue discipline is either first-in-first-out or a set of patient classes that have different priorities. When arriving patients are placed in different queues, each of which has a different service priority, the queue discipline may be preemptive or non-preemptive. In the latter, low priority patients receive service only when no high priority patients are waiting, but the low priority patient who is receiving service is not interrupted if a high priority patient arrives and all servers are busy. In the preemptive queue discipline, however, the service to a low priority patient is interrupted in this event. So, here we models a single server queue and divides time into equally long slots (discretizing time). Periods of emergency interruptions are considered to have no server available from the point of view of the scheduled patients (vacation). The result is a discrete-time queuing model with exhaustive vacations.

Blocking occurs when a queuing system places a limit on queue length. For

example, an outpatient clinic may turn away walk-in patients when its waiting room is full. In a hospital, where in-patients can wait only in a bed, the limited number of beds may prevent a unit from accepting patients.

Appointment System

Compared to systems without appointments, systems with appointments reduce the arrival variability and waiting times at the facility. However, it is important to note that systems with appointments require patients to wait outside the facility. Of course, because it is not at the facility, this waiting can be productive time and therefore has lower cost to the patient. A key issue has been to reduce patient waiting times without causing a significant increase in doctor idle time, a significant cost for the healthcare facility. Many outpatient appointments allow booking appointments months in advance. So, the patient (if they are not able to visit at that time) without cancelling appointments could lead to waste of resources. They propose implementing short-notice appointment systems based on a queuing network analysis tailored to the realities of any particular outpatient clinic. Their approach assumes the availability of a certain number of staff who can be distributed amongst the different stations of the queuing network in several combinations. A combination is chosen based on its resulting utilization per station and expected patient length of stay in clinic. The implementations of these ideas did not improve the appointment system, a failure which they attribute to the clinic using many visiting doctors and the patients being unable to schedule visits with their primary care physician at short notice.

In a queuing network, there are several nodes at which services are dispensed. A patient may have to go through several nodes, and thus several queues in order to obtain the desired service. In the context of appointment systems, we can expect nodes where the ratio of demand to available service capacity is relatively high to become bottlenecks. Such bottlenecks would have high utilization and increase overall patient waiting times even though other nodes may have low utilization.

We can draw some conclusions from the above work. The variability in demand for healthcare services and service times mean that simplistic rules like mandating specific utilization levels or fixing patient to resource ratios would lead only to congestion and poor quality of service and are unlikely to be successful approaches to contain or reduce healthcare costs. Larger organizations with more patients are able to attain the same quality of service at higher utilizations than smaller organizations. Although appointment systems are often designed to avoid doctor idle time, it is possible to reduce patient wait time without significantly increasing doctor idle time. As long as increasing the productivity of healthcare organizations remains important, analysts will seek to apply relevant models to improve the performance of healthcare processes. This chapter shows that many models are available today. However, analysts will increasingly need to consider the ways in which distinct queuing systems within an organization interact [3].

CONCLUSION

This project consists of four chapters which discusses about queuing theory. Queuing theory is a major system in our day to day life. Every person has had to stand in queue at one point in our lives. So queuing theory can be used to help reduce waiting times and where waiting times are inevitable. Here we discussed about the queuing system and its components, probability distributions in queuing theory and classification of queuing models and discusses 2 queuing models with examples. Also we have included one chapter which discuss about the queuing theory application in healthcare. From these we can say that queuing theory plays an important role in our life by helping to make decisions, relieving human suffering of waiting and also to minimize costs/maximize profits. In short we can observe real-world systems and recognize potential problems, construct mathematical models representing these systems, analyze the models (performance analysis and decision making), use the analysis to provide strategies, heuristics and insights and thus solve real-world problems (connect theories and applications).

BIBLIOGRAPHY

[1] Samuel L. Baker. Queuing theory 1, 2006.

[2] RYAN Berry. Queuing theory. *Senior Project Archive*, pages 1–14, 2006.

[3] Samuel Fomundam and Jeffrey W Herrmann. A survey of queuing theory applications in healthcare. 2007.

[4] SC Gupta and VK Kapoor. *Fundamentals of mathematical statistics*. Sultan Chand & Sons, 2020.

[5] John F Shortle, James M Thompson, Donald Gross, and Carl M Harris. *Fundamentals of queueing theory*, volume 399. John Wiley & Sons, 2018.