Project

Report On

A BRIEF STUDY ON QUEUING THEORY

Submitted

in partial fulfilment of the requirements for the award of the degree of

BACHELOR OF SCIENCE

in

MATHEMATICS

by

AGGIE MATHEW. E(AB21AMAT037)

Under the Supervision of

Ms. PARVATHY T S

DEPARTMENT OF MATHEMATICS AND STATISTICS

ST. TERESA'S COLLEGE (AUTONOMOUS)

ERNAKULAM, KOCHI - 682011

MARCH 2024

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

CERTIFICATE

This is to certify that the dissertation entitled, **A BRIEF STUDY ON QUEUING THEORY** is a bonafide record of the work done by **Ms. AGGIE MATHEW.E.**under my guidance as partial fulfillment of the award of the degree of Bachelor of Science in Mathematics at St. Teresa's College (Autonomous), Ernakulam affiliated to Mahatma Gandhi University, Kottayam. No part of this work has been submitted for any other degree elsewhere.

Date:22/03/2024
Place: Ernakulam

Ms.PARVATHY T S
Assistant Professor,
Department of Mathematics,
St.Teresa's College(Autonomous),
Ernakulam.

Dr. Ursala Paul
Assistant Professor and Head ,
Department of Mathematics,
St. Teresa's College(Autonomous), Ernakulam.

External Examiners

1:.............................   2: ....

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of Ms.Parvathy T.S, Assistant Professor, Department of Mathematics, St. Teresa's College(Autonomous), Ernakulam and has not been included in any other project submitted previously for the award of any degree.

Ernakulam.

Date:22/03/2024 REGISTER NO: AB21AMAT037

ACKNOWLEDGEMENT

I want to express my sincere gratitude to Ms. Parvathy T S, Assistant Professor in the Department of Mathematics at St. Teresa's College, who served as our project's coordinator and made our work possible. I was able to complete all of my project's writing stages because of her direction and assistance.

Additionally, I want to express my gratitude to the members of my group for making our defence entertaining and for their insightful remarks and ideas.

I also want to express my gratitude to Dr. Ursala Paul and Dr. Elizabeth Reshma M. T, who collectively have indirectly led and assisted us in finishing this project.

Finally, I want to express my gratitude to God for guiding us through all of the challenges. Every day, I have felt your guiding.

Ernakulam.

Date: 22/03/2024. AGGIEMATHEW. E(AB21AMAT037)

CONTENTS

INTRODUCTION

Waiting in lines is a typical occurrence in daily life and can happen at bus stops, gas stations, eateries, ticket booths, bank counters, doctor's offices, and so on. In addition, waiting lines can be seen at tool cribs where mechanics operate and in workshops where machinery are waiting to be fixed. Calls that arrive wait to mature in the queue; in a warehouse where goods are kept until they are needed, aeroplanes waiting to take off or land, trucks waiting to be unloaded, phone exchanges, and so forth.

Anywhere that a customer (whether a person or a physical object) who needs service is forced to wait because there are more customers than service facilities or because service facilities are inefficient and take longer than expected to complete, a queue forms.

The goal of queuing theory is to create balanced systems that meet the needs of customers in a timely and efficient manner while remaining cost-effective for long-term sustainability. Queuing theory as a part of operations research can be used to inform business decision-making on how to create more efficient and low-cost workflow systems. By analysing factors such as arrival rates, service rates, and queue configurations, businesses and organizations can optimize their operations to minimize wait times, enhance customer satisfaction, and improve overall efficiency. Additionally, advancements in queuing theory continue to shape various industries, offering innovative solutions to complex queuing problems. As such, further research and application of queuing theory principles hold great promise for addressing the challenges of modern-day queuing systems.

Chapter 1

BASICS AND STRUCTURE OF A QUEUING SYSTEM

The mathematical study of waiting lines, or queues, is known as queuing theory. To estimate waiting periods and queue lengths, a queuing model is built. Because the findings are frequently applied to business choices regarding the resources required to offer a service, queuing theory is typically regarded as a subfield of operations research.

Queuing analysis is the probabilistic analysis of waiting lines. The various operating characteristics that these queuing models compute include the probability that n customers are in the queuing system, the average number of customers in the queuing system, the average number of customers in the waiting queue, the average time spent by a customer in the entire queuing system, the average time spent by a customer in the waiting queue, and finally the probability that the server is busy or idle.

Queuing theory is the study about queues. There are different types of queues. There is a customer and a server in all queues. The queue is mainly formed to get any services. We mainly discuss about Single server model of queuing theory in this project. In this single server model first come customer is served first. There are some quantities on which the queues depend.

Server: The one who serves clients is known as the server.

Queue: A queue of people is formed to receive service.

1.1 STRUCTURE OF QUEUING SYSTEM

A queuing system, a fundamental concept in operations research and computer science, comprises several key components that collectively ensure efficient and organized task processing. From arrival processes and queue disciplines to service mechanisms and system configurations, each element plays a crucial role in shaping the dynamics of a queuing system, influencing its performance and effectiveness.

Main Components of a Queuing System: A queuing system is characterised by these components:

-Calling population

- Queuing / Arrival process

- Service mechanism

- Queue discipline.

1.1.1 CALLING POPULATION

The term "calling population" describes the group of possible clients. They are also referred as customer (input) service.

There can be an infinite or finite calling population. We typically presume that the population in large-scale systems is unlimited. The way the arrival rate is expressed distinguishes the "finite" population model from the "infinite" population model.

The arrival rate in an infinite population model is independent of the total number of users in the system. The system is typically seen as being open, with clients entering from outside and exiting once the job is completed.

The population of the system has an impact on the arrival rate at a server in a finite population model. Customers (with a fixed population) simply move throughout the system from one server to another and from one queue to another; the system is typically perceived as being closed. Size of the calling population, behaviour of the arrivals, pattern of arrivals at the system are some of the main characteristics inputs to the system.

A few scenarios of calling population in different fields:

1.Customer Support Hotline: The calling population for a customer support hotline would be the total number of customers who may need assistance or have inquiries.

2.Hospital Emergency Room: In a hospital's emergency room, the calling population would be the number of individuals seeking urgent medical attention.

3.Online Ticket Sales: For an online ticketing platform, the calling population would be the users trying to purchase tickets for an event.

4.Public Transportation System: In a bus or train system, the calling population would be the passengers waiting at stops or stations.

1.1.2 QUEUING PROCESS

The term “queuing process” describes the quantity and duration of queues, whether they are single, multiple or priority queues. The structure of the service mechanism determines the type of queue, while the operational factors like available spaces, applicable laws and consumer behaviour determines a queues length or size.

A service system may occasionally be unable to handle more clients at once than in necessary. Until extra space is made available to accept new clients no more patrons are permitted entry. These kinds of circumstances are known as limited or finite source queues. Restaurants, movie theatres and other establishments have finite supply lines.

However, when a service system can handle an infinite (or limitless) number of clients at once, it is called an endless (or unlimited) source queue.

Even when there is more waiting room available clients frequently choose not to enter a service system if they find themselves in a big queue in front of the facility. In these situations, the duration of the queue is determined by the patron's disposition. For instance, most of the time a driver does not stop at a petrol station when he notices that numerous cases are waiting there.

1.1.3 QUEUE DISCIPLINE

The sequence or process in which clients are chosen for service from the queue is known as the queue discipline.

The customers in the queue are attended to in a variety of ways. Among them are:

1.Static queue discipline: Disciplines for static queues, these depend on each customers position in the queue. Few of these types of queue discipline include:

(a) First come first served that is FCFS service, where clients are served according to arrival order. An illustration of this discipline is the prepaid taxi queue found at airports where taxi are booked on a "first come first served" basis.

(b) Last come, first served is another discipline that is frequently used that is LCFS. In order to minimise handling and shipping costs this discipline is mostly used in cargo handling, where the last item loaded is withdrawn first and in production process, where goods that reach at the items in a workplace are piled one on top of the other, and things that reach last for service gets processed first.

2.Dynamic queue disciplines: These are determined by the unique characteristics of each customer in the queue. Such queue disciplines are as follows:

(a)Service in random order: Customers are chosen arbitrarily for service under this policy, regardless of when they enter the service system.

(b) Priority service: customers are categorized into priority classes under this policy based on several factors like urgency or service time. Every classes makes advantages of the FCFS rule to deliver services. Examples include phone or electric bills using cash or cheque.

(c)Pre-emptive priority: This policy states that even if a client with a lesser priority is already in service, a significant consumer may join the service right away after entering the system. That is client assistance that is halted (prior is emptied) in order to begin serving that client. Following the priority customises services, this interrupted service is restarted.

(d) Non-pre-emptive priority: In this instance, a valuable client is permitted to skip ahead in queues, but the service will begin as soon as the current one is finished.

1.1.4 SERVICE PROCESS

The way that clients are served and exit the service system is the focus of the service mechanism. It is distinguished by the following:

*The configuration of service facilities.

*The distribution of service hours.

*Behaviour of the server.

*Policies for management.

1.2 ARRANGEMENT OF SERVICE FACILITIES

The number of clients that can be successfully and concurrently served is how the service facility's capacity is determined. Service channels, sometimes referred to as service facilities or servers can be parallel, series. Mixed

(a).Series arrangement

The series arrangement consists of a number of service locations arranged in a specific order such that a customer must visit each location before the service is finished. Nonetheless, each service facility is free to operate freely and set it's own service standards. For instance, in college or university. Students go through each counter in the admission process one after the other to finish the necessary paper work.

(b).Parallel Arrangement

A customer can either join the queue of their choosing in front of service facilities or can be served by any service facility. Thanks to the parallel Arrangement, which places multiple service facilities is in close proximity to each other. A few instances of the pattern are the quantity of ticket counters at train stations and check in counters at airports. These are some examples of service facilities in parallel. Since client can obtain a ticket or check in at any counters, these are parallel.

(c).Mixed arrangement

The mixed arrangement comprises of service facilities grouped both parallel and in series. For instance, a new patient may need to travel to any of the OPD counters in a hospital, where he will be examined (served) one-by-one by different physicians (facilities in series).

Single Queue, Single Facility

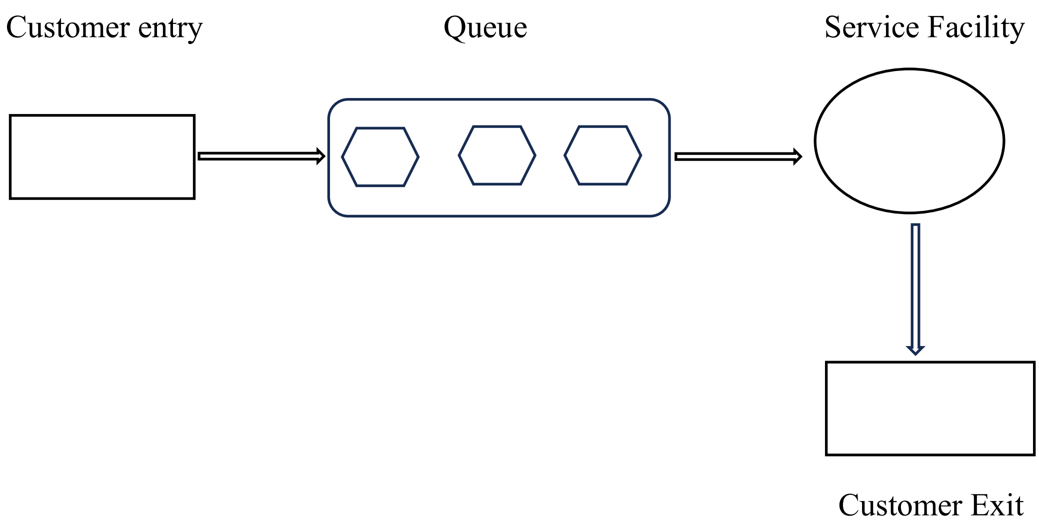

Fig.1.1

Single Queue, Multiple Facility in series

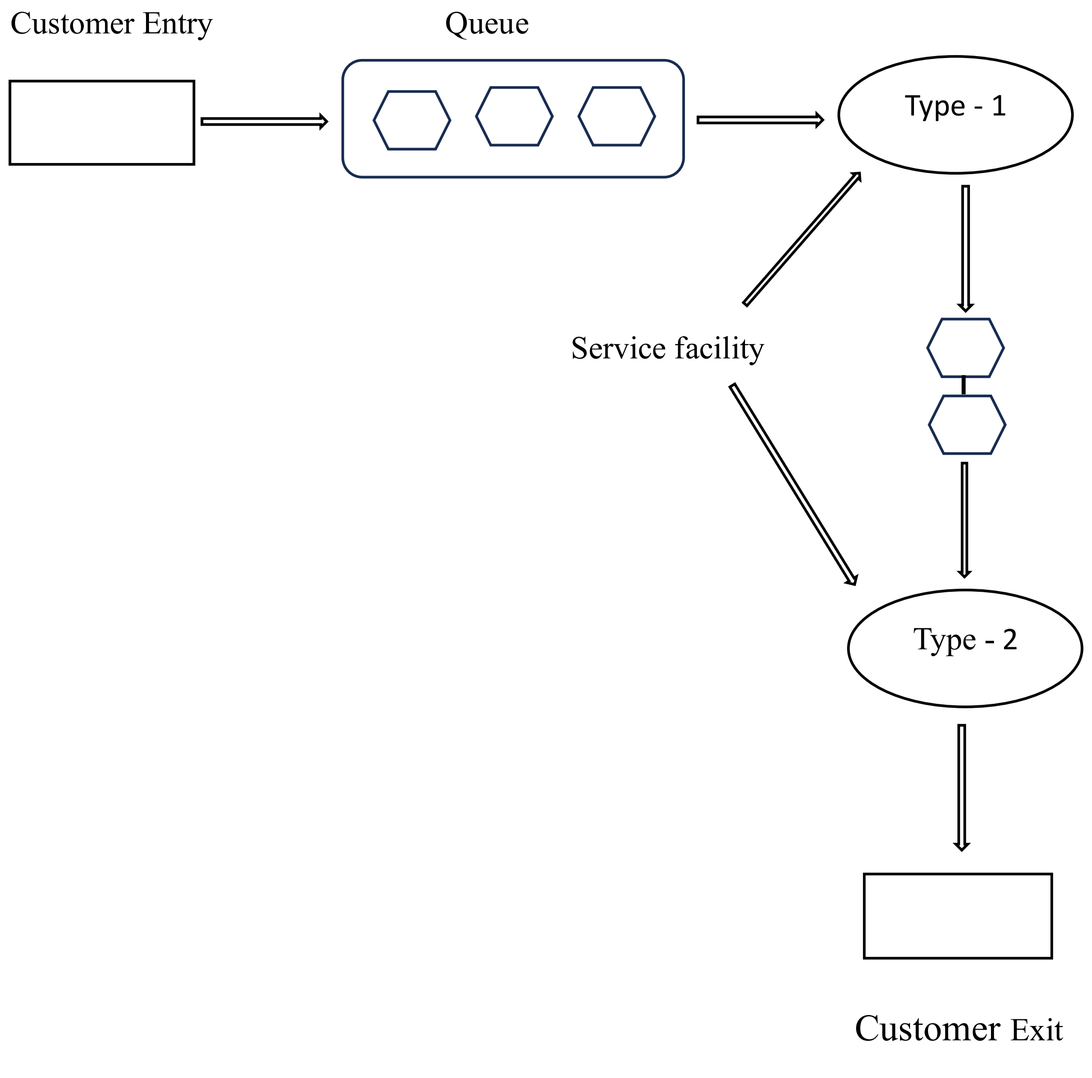

Fig.1.2.

Single queue, multiple service facilities in parallel

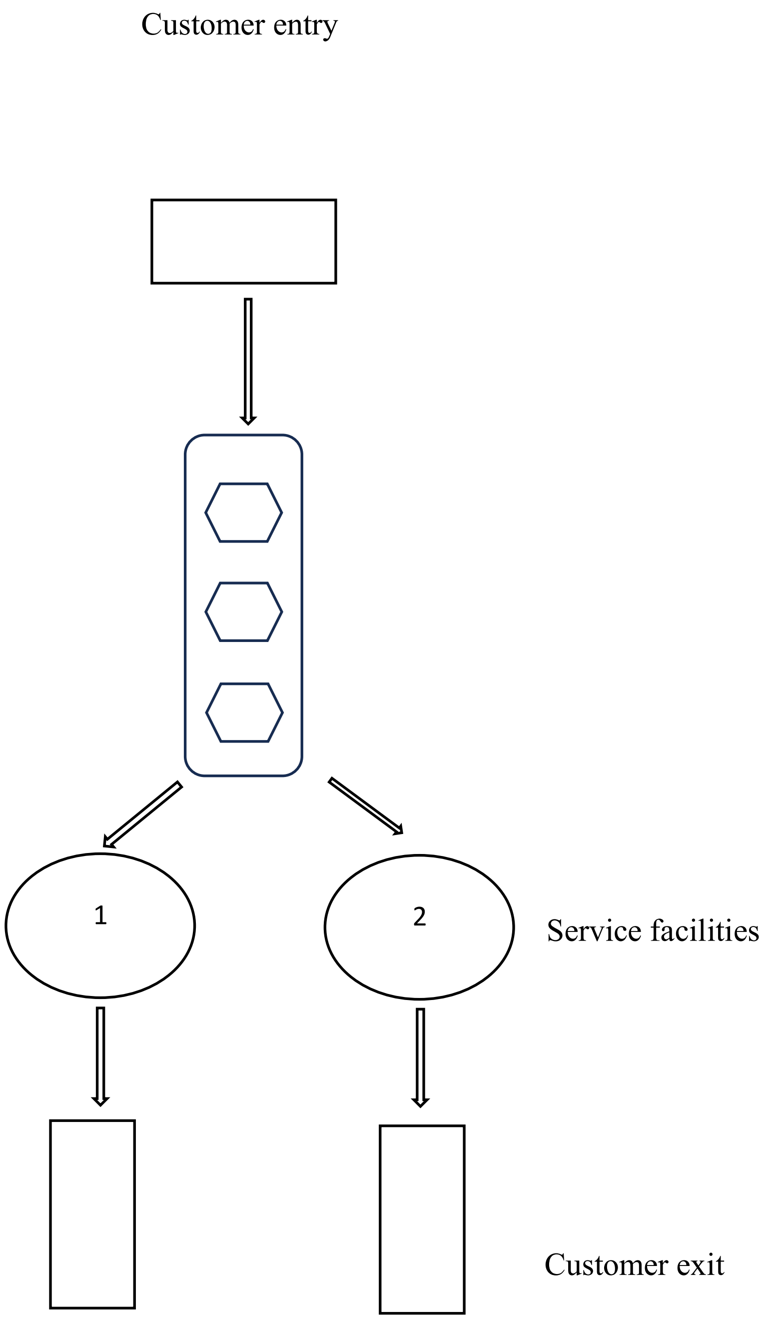

Fig. 1.3

Multiple queues, multiple service facilities in parallel

Customers Service facility Served customers

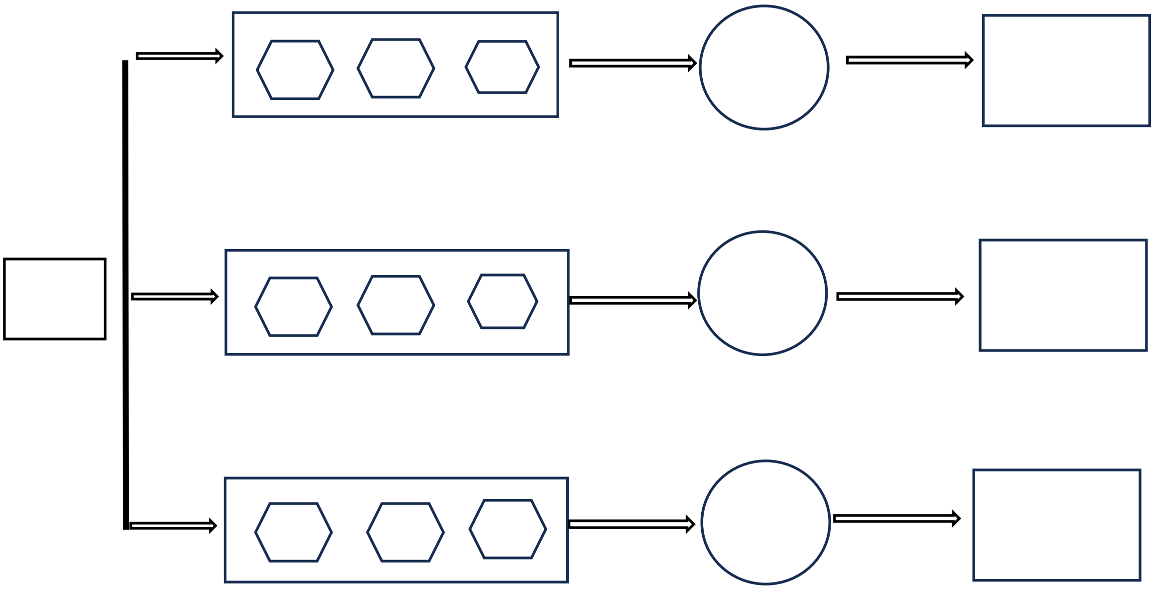

Fig. 1.4

Mixed arrangement

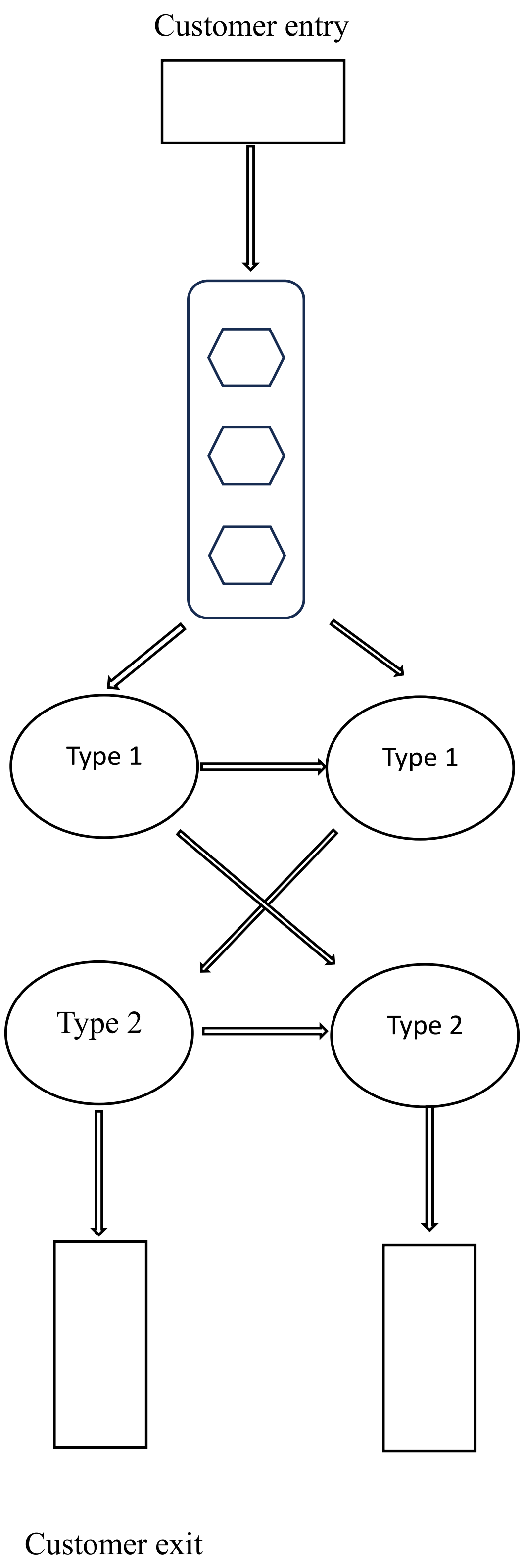

Fig. 1.5

Chapter 2

PERFORMANCE MEASURES OF A QUEUING SYSTEM

The performance measures, also known as operating characteristics, are used to evaluate the performance of the queuing system and to design a new system with regard to the quality of service that a customer receives as well as the correct use of service facilities. The performance measures include the following:

a) Average time that a customer spends in a queue and system;

W_q : Average time a customer is queued before being served.

W_s : Average time a customer spends in system after arrival, which includes waiting and service time.

b) Average number of customers queued up;

L_q : Average number of customers queuing up (queue length) for service.

L_s : Average number of people in the system (either in queue or waiting to be served)

c) Average cost to run the queuing system, includes average cost of running a system per unit of time and the number of servers needed to be cost effective.

d) The value of time for customers as well as servers;

P_w : Probability that an incoming customer will have to wait before the service is rendered (blocking probability).

P_n : Probability of n customers waiting in the queue for service.

P_d : Probability that an incoming customer is unable to join the queue, which means the capacity is full.

2.1.TRANSIENT AND STEADY STATES

At the start of a service operation, the queuing system is affected by initial conditions, like the number of people waiting in the queue and percentage of servers busy rendering service and so on. This period referred to as the transient state. After a certain period of busy service, the system is no longer affected by the initial conditions and the system is assumed to be in a steady state. Let $P_n(t)$ denote the probability of having customers in the queuing system at some time, say t, and let $P_n'(t)$ represent the change in $P_n(t)$ with regards to t.

Consider steady state, we can say that as $t \to \infty$, $P_n(t)$ tends to P_n, which means lim $P_n(t)$ = P_n , hence we can see that it is independent of time.

This would imply that lim $P_n'(t) = 0$. From this we understand that if the rate of arrival of customers is greater than service rate then the steady state will never be achieved.

Some Important Notations

Some important notations that are used for studying the queuing system are as follows:

n = number of customers in the system consisting of those who are waiting and in service

P_n = probability that there are n customers in the system

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

μ = average service rate or the average number of customers served per unit time

P_0 = probability that there are no customers in the system

s = Number of servers available

N = maximum number of customers allowed in the queuing system

W_s = average waiting time in the system for both waiting and service

W_q = average waiting time for customers in the queue

L_s = average number of customers in the queuing system both waiting and in service

L_q = average number of customers in the queue also known as queue length

P_w = probability that an arriving customer has to wait due to the system being busy:

$$1 - P_0 = (\lambda/\mu)$$

Utilization factor = the average fraction of time for which the servers are being utilized during customer service.

It is given by, $\lambda/\mu = \rho$, it is also known as the traffic intensity or the server utilization factor. In fact, for achieving a steady-state condition, it is necessary that, ρ is less than 1. This happens when the queue is of limited length.

2.2.RELATIONSHIPS AMONG PERFORMANCE MEASURES

Some relationships among the performance measures are as follows: (these hold for all infinite source queuing models)

1. Average waiting time for a customer waiting in the line:
   $Wq = Lq/\boldsymbol{\lambda}$
2. Average waiting time for a client in the system which includes service time:
   $Ws = Wq + 1/\mu$
3. Average number of clients in the system:
   $Ls = Lq + \boldsymbol{\lambda}/\mu$
4. Probability of a customer staying in system for more than time t:
   $P(T > t) = e^{-(\mu-\lambda)\lambda t}$
   and $P(T \leq t) = 1 - P(T > t)$
   where,
   T= time spent in the system
   t = specified time period, and
   e = 2.718
5. Probability that there are exactly n customers in the system is given by:
   $P_n = P_0\left(\frac{\lambda}{\mu}\right)^n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n$
6. Probability of waiting for service longer than time t is given by:
   $P(T > t) = \frac{\lambda}{\mu}e^{-(\mu-\lambda)t}$
7. Probability that the number of customers in the system, exceeds a given number, r is given by:
   $P(n > r) = \left(\frac{\lambda}{\mu}\right)^{r+1}$

2.3 PROBABILITY DISTRIBUTION

Customers are supposed to come at the queuing system in a random order, with the Poisson distribution accounting for their arrival and the exponential distribution accounting for the intervals between arrivals.

Customer service time is typically taken to be exponentially distributed. This suggests that the likelihood of a service being completed is fixed and unaffected by the amount of time it takes.

In a queuing system, the number of arrivals and departures (those served) over a given period of time is determined by the following presumptions, often known as axioms:

(i) The length of time interval Δt determines the chance of an event (arrival or departure) occurring during the time interval $(t, t + \Delta t)$. This means that the events that occur in non-overlapping time are statistically independent, and the chance that the occurrence will not depend on the number of events that occur up to time t or the particular value of t.

(ii) There is very little chance that more than one event will take place in the time span $(t, t + \Delta t)$. $0(\Delta t)$ is used to indicate this.

(iii) In a little period of time t, nearly one event (arrival or departure) may take place. $\lim \{0(\Delta t)/\Delta t\}=0$,$\Delta t\to 0$ is the quantity that becomes negligible when compared to Δt as $\Delta t \to 0$. This means that the probability of an arrival during the time interval $(t, t + \Delta t)$ is given by: $P_1(\Delta t) = \lambda\Delta t + 0(\Delta t)$ where λ is a constant and is independent of the total number of arrivals up to time t; Δt is a small time interval.

Distribution of arrivals (Pure birth process)

Customers are assumed to arrive at the queuing system and never depart, according to the arrival process. A pure birth process is what is known as such an arrival procedure. When creating different queuing models, the following phrases are frequently used:

In any given tiny span of time Δt, the probability of more than one customer arriving is minimal.

$\lambda\Delta t$ represents the likelihood that a client will enter the system at time Δt. $1 - \lambda\Delta t$ = likelihood that no client will enter the system at time Δt.

Distribution of interarrival times

The inter-arrival time in a fixed period follows the exponential distribution if the number of arrivals, r, in time t follows the Poisson distribution, and P (x = r) =$\frac{e^{-\lambda t}(\lambda t)^r}{r!}$, r = 0, 1, 2,..., P(x = t) = $\lambda e^{-\lambda t}$

Markovian property of Interarrival times

According to the Markovian property of interarrival times, a customer's likelihood of receiving service at a specific time, t, is independent of the service time. That is,

$P\{0 \leq T \leq t_1 - t_0\} = P\{T \geq t_1 | T \geq t_0\}$, where T is the interval of time between consecutive arrivals.

Distribution of departure

In this case no extra customers arrives into the queue but the service is still given to the customers already occupied in the queue. This process is known as pure death process.

Let N= number of customers

μ =rate of customers arriving

These are three axioms to explain distribution of departure for Poisson queuing system.

Axiom 1: In a time interval Δt probability of departure is $\mu \Delta t$

Axiom 2: Probability of more than one departure between time t and t+ Δt is negligible

Axiom 3: The number of departures in non-overlapping intervals are statistically independent.

Distribution of service time

Service time is the time that is required to give the service to a customer.

Let f(t)= probability density function of service time

$$f(t) = \begin{cases} \mu e^{-\mu t}; 0 \leq t < \infty \\ 0; t < 0 \end{cases}$$

From this function it is clear that service time follows negative exponential distribution.

2.4. ARRIVAL DISTRIBUTION THEOREM

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

Poof:

Δt = smallest time interval during which there is a minimal chance oof consumers arriving, that is just one customer can arrive in any given small time interval.

$\lambda\Delta t$ = The probability that a customer will enter the system at time Δt

$1-\lambda\Delta t$ = The probability that no customers will enter the system at the time Δt

Case 1: $n\geq1$, $t\geq0$

The probability of having n customers in the system at time $t+\Delta t$ is denoted as $P_n(t+\Delta t)$, which is the total of the joint probabilities of the 2 distinct of collectively exhaustive instances that follow :

(i)At time t, there are n consumers in the system, no arrivals occur within the time period Δt.

(ii)At the time t, there are (n -1) customers in the system and there is only one arrival and no departures at that period Δt.

By(i) probability = (probability of n units at time t)(Probability of no arrival during Δt)

$$=P_n(t)(1-\lambda\Delta t) \tag{1}$$

By (ii)Probability = (Probability of (n-1) units at time t)(Probability of one arrival timeΔt)

$$=P_{n-1}(t)(\lambda\Delta t) \tag{2}$$

Adding (1)and (2),we get probability of n arrivals at time $t+\Delta t$

$$\text{i.e}\quad P_n(t+\Delta t) = P_n(t)(1-\lambda\Delta t)+P_{n-1}(t)(\lambda\Delta t)+0(\Delta t) \tag{3}$$

Case 2: $n=0$, $t\geq0$

There won't be an arrival during Δt if there isn't a customer in the system at time $t+\Delta t$. Thus at time $t+\Delta t$, the probability that there would no customers in the system is given by,

$P_0(t+\Delta t)$= (Probability of no units at time t) × (Probability of no arrival in time Δt)

$$= P_0(t)(1- \lambda\Delta t) \tag{4}$$

Rewriting the equations (3) and (4):

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

$P_0(t+\Delta t) - P_0(t) = -P_0(t)\,\lambda\Delta t, \quad n=0$

Dividing the above equations by Δt,

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

$$\frac{P_0(t+\Delta t) - P_0(t)}{\Delta t} = -P_0(t)\lambda, \quad n=0$$

$\Delta t \to 0$, and taking limits on both sides, the differential difference equation system that follows can be obtained

By definition of first derivative $\lim_{\Delta t\to 0}\left[\frac{P_n(t+\Delta t)-P_n(t)}{\Delta t}\right] = \frac{d}{dt}P_n(t) = P_n'(t)$

$$P_n'(t) = -P_n(t)\lambda + P_{n-1}(t)\,\lambda, \quad n\geq 1 \tag{5}$$

$\frac{d}{dt}P_0(t) = P'_0(t) = -P_0(t)\,\lambda$

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

Integrating with respect to t,

$$\log P_0(t) = -\lambda t + C \tag{6}$$

C is the integration constant, when t=0,By boundary conditions, we have ;

$P_n(0)= \begin{cases} 1, & n=0 \\ 0, & n\geq 1 \end{cases}$

Substituting t=0 in (6) and $P_0(0)=1$ we get C=0

Now, equation (6) becomes,

$\log P_0(t) = -\lambda t$ or $P_0(t) = e^{-\lambda t}$, $t \geq 0$ (7)

Substituting n=1 in equation (5) and from equation (7)

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

$$P_1'(t) + P_1(t)\lambda = \lambda e^{-\lambda t} \tag{8}$$

A first order linear differential equation is found in equation (8). As a result, it can be resolved by multiplying by the integrating factor on both sides.

$$e^{\lambda t}[P_1'(t) + P_1(t)\lambda] = \lambda e^{-\lambda t} e^{\lambda t}$$

$$e^{\lambda t} P_1'(t) + e^{\lambda t} P_1(t)\lambda = \lambda$$

$$\frac{d}{dt}(e^{\lambda t} P_1(t)) = \lambda$$

Integrating w.r.t to t we get

$$e^{\lambda t} P_1(t) = \lambda t + B \tag{9}$$

Where B is the integration constant, and initial conditions can be used to determine its value once more. In other words, since $P_1(0) = 0$, we have $P_1(0) = B = 0$ when we set t = 0 in equation (9). Thus, the form of equation (9) is reduced to

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

Or . $$P_1(t) = \frac{\lambda t}{e^{\lambda t}} = \lambda t e^{-\lambda t} \tag{10}$$

Substituting n = 2 in equation (5) and the value of $P_1(t)$,

$$P_2'(t) + P_2(t)\lambda = P_1(t)\lambda$$

$$P_2'(t) + P_2(t)\lambda = \lambda^2 t e^{-\lambda t}$$

Multiplying with the integrating factor we get,

$$e^{\lambda t}P_2'(t) + e^{\lambda t}P_2(t)\lambda = \lambda(\lambda t)$$

Or. $$\frac{d}{dt}\left[P_2(t)e^{\lambda t}\right] = \lambda(\lambda t)$$

By integrating w.r.t t we get,

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

$$= \frac{(\lambda t)^2}{2!} + A$$

where A is the integrating constant with the initial conditions t = 0 and $P_2(0)$, then we have

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

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

In general,

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

The number of customers in the system at a time t prior to the start of the service facility is indicated by the general solution for $P_n(t)$, which has a Poisson distribution with a mean and variance equal to λt. This technique succeeds regardless of the type service given to the customers who are waiting in the system, as the average or projected number of customers is independent of the service.

Chapter 3

SINGLE SERVER QUEUING MODEL

These are the most basic kind of queuing systems, and you'll find them in many basic applications like fast-food restaurants or retail businesses. In these systems, customers arrive and wait in queue to be attended to by a single server.

A single server follows the first-come, first-served policy, serving each customer individually from the front of the queue. The consumer exits the queue when the service is finished, and there are now one fewer customer in the system.

The behaviour of the single server is critical in defining the system's overall performance, as is its service rate, or the speed at which it can process entities, as well as any potential capacity limitations.

A birth-death process, which defines the amount of jobs in the system at any given time as well as arrivals and departures from the queue, can be used to characterise the activity of a single queue, also known as a queueing node. An arrival raises k by 1 and a departure reduces k by 1, assuming that k represents the total number of jobs in the system (either being served or waiting, if the queue has a buffer of waiting jobs).

In order to help organisations make sensible choices about distributing resources and process optimisation, the main objective is to evaluate performance measures such as queue length, waiting time, and server utilisation.

The Poisson probability distribution characterises arrivals, while the exponential distribution defines service times and a single server or channel. Although the arrival of customers is independent, the average number of arrivals, or arrival rate, remains constant over time & compared to the average arrival rate, the average service rate is higher.

in a short interval of time, Δt, immediately before time t.
There are no arrivals or departures and the system is in state n (number of customers) at time t. There is one departure and no arrivals in state n + 1 (number of consumers) of the system.

There is only one arrival and no departure in state n-1 (number of consumers) of the system.

3.1. MODEL 1:{(M/M/1) : (∞/FCFS)} Exponential service-Unlimited queue

To obtain system of differential difference equations:

Let $P_n(t)$ be the probability of n customers at time t. The probability that the queuing system contains n customers at a time t+ Δt is the summation of joint probabilities of the three events. (probability of no arrival in Δt and no departure in Δt, probability of no arrival in Δt and one departure in Δt, probability of one arrival in Δt and no departure in Δt)

Now consider n≥ 1 and t≥0,

Let $P_n(t + \Delta t)$ = probability of n customers at time t+ Δt

P(one arrival of customer in at time Δt) $=\lambda\Delta t$

P(one departure at time Δt) $=\mu\Delta t$

P(no arrival in Δt) $=1-\lambda\Delta t$

P(no departure in Δt) $=1-\mu\Delta t$

Hence, $P_n(t + \Delta t) = P_n(t)$.P(no arrival in Δt and no departure in Δt) + Pn+1(t).P(no arrival in Δt and one departure in Δt) +Pn-1(t).(one arrival in Δt and no departure in Δt) =

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

$$= P_n(t)(1- \mu\Delta t - \lambda\Delta t + \lambda\mu\Delta t.\Delta t) + P_{n+1}(t)[\mu\Delta t - \lambda\mu\Delta t^2] + P_{n-1}(t)(\lambda\Delta t - \lambda\mu\Delta t^2)$$

As Δt is very small, hence the terms containing Δt^2 are negligible.

So the equation becomes,

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

Or. $P_n(t)[1-(\mu+\lambda)\Delta t] + P_{n+1}(t)[\mu\Delta t] + P_{n-1}(t)[\lambda\Delta t]$

Hence, $P_n(t+\Delta t) = P_n(t) - P_n(t)[\mu+\lambda]\Delta t + P_{n+1}(t)\mu\Delta t + P_{n-1}(t)\lambda\Delta t.$ (1)

Subtracting $P_n(t)$ and diving by Δt on both the sides, then the equation (1) becomes,

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

$$= P_{n+1}(t)\mu - (\mu+\lambda)P_n(t) + P_{n-1}(t)\lambda \qquad (2)$$

As Δt tends to 0, we get equation (2) as,

$$P_n{}'(t) = P_{n+1}(t)\mu - (\mu+\lambda)P_n(t) + P_{n-1}(t)\lambda \qquad (3)$$

Now for n = 0 and t ≥ 0

That is, at the time t+Δt there are no customers in the system and no service is completed within the time Δt. So we get the equation (3) as,

$$\frac{P_0(t+\Delta t) - P_0(t)}{\Delta t} = P_1(t)\mu - (\mu+\lambda)P_0(t) = P_1(t)\mu - \lambda P_0(t)$$

As Δt tends to 0, in this equation we get,

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

Or. $\frac{d}{dt}P_0(t) = P_1(t)\mu - \lambda P_0(t)$ (4)

To obtain the system of steady state equations:

$P_n(t)$ is not dependent on time in the case of steady state.

Then we have $\lim_{n\to\infty} P_n(t) = P_n$

And $\lim_{n\to\infty} \frac{d}{dt}[P_n(t)] = 0, \ for\ n = 0,1,2, \ldots.$

Substituting conditions on equation (3) and (4)

$$0 = P_{n+1}\mu - (\lambda + \mu)P_n + P_{n-1}\lambda, for \;\; n \geq 1 \tag{5}$$

If n= 0 then the equation (5) becomes,

$$0 = \mathrm{P}_1\mathrm{\mu} - \mathrm{\lambda P}_0 \tag{6}$$

The equations (5) and (6) are called the steady state differential equations.

To solve the system of difference equations:

By equation (6), $P_1\mu - \lambda P_0 = 0$

$\mathrm{P}_1\mathrm{\mu} = \mathrm{\lambda P}_0$, since n=0

This gives, $P_1 = \frac{\lambda}{\mu}P_0$

Substituting n=1 in equation (5),

$$P_2\mu - (\mu + \lambda)P_1 + \lambda P_0 = 0$$

$$P_2\mu = (\mu + \lambda)P_1 - \lambda P_0$$

$$P_2 = \frac{\mu + \lambda}{\mu}P_1 - \frac{\lambda}{\mu}P_0$$

Substituting, $P_2 = \left(\frac{\mu+\lambda}{\mu}\right)\frac{\lambda}{\mu}P_0 - \frac{\lambda}{\mu}P_0 \; = \; \frac{\lambda}{\mu}P_0\left[1 + \frac{\lambda}{\mu} - 1\right] = \left(\frac{\lambda}{\mu}\right)^2 P_0$

Substituting n=2 in equation (5), $P_3\mu - (\mu + \lambda)P_2 + \lambda P_1 = 0$

$$P_3\mu = (\mu + \lambda)P_2 - \lambda P_1$$

$$P_3 = \left(\frac{\mu + \lambda}{\mu}\right)P_2 - \frac{\lambda}{\mu}P_1$$

Substituting P_1 and P_2, $P_3 = \left(\frac{\mu+\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^2 P_0 - \frac{\lambda}{\mu}\frac{\lambda}{\mu}P_0 = \left(\frac{\lambda}{\mu}\right)^2 P_0\left[1 + \frac{\lambda}{\mu} - 1\right] = \left(\frac{\lambda}{\mu}\right)^3 P_0$

And so on. Hence in general, $P_n = \left(\frac{\lambda}{\mu}\right)^n P_0, n = 1,2,\cdots$ (7)

We know that sum of all probabilities is 1. Hence $\sum_{n=0}^{\infty} P_n = P_1 + P_2 + \cdots + P_n = 1$

So, $\sum_{n=0}^{\infty} P\left(\frac{\lambda}{\mu}\right)^n P_0 = 1$

$$P_0 = \frac{1}{\sum_{n=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^n}$$

We know that it is a geometric series with $\frac{\lambda}{\mu}$ as the common ratio. Here, $\sum_{n=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^n \textit{ is a geometric series.}$

Therefore, $\sum_{n=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^n = \frac{1}{\left(1-\frac{\lambda}{\mu}\right)}$

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

Therefore, $P_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho$ and $P_n = \left(\frac{\lambda}{\mu}\right)^n P_0$

Substituting P0, $P_n = \left(\frac{\lambda}{\mu}\right)^n (1-\rho), \rho < 1, n = 0,1,2, \ldots$

To obtain the waiting time probability density function, excluding the service time distribution:

In the steady state condition, each customer's waiting time distribution is the same. Let w be the amount of time the server needs in a steady-state condition to serve every customer in the system at a given moment. Let w's probability distribution function given by $\phi_\omega(t) = P(\omega \le t), 0 \le t \le \infty$

Let the service time that the server takes to serve each of the n customers be represented as $s_1, s_2, s_3, \ldots$

Then, $P(\omega \le t) = \begin{cases} P_0 = 1-\rho\,, n = 0 \\ P[\sum_{i=1}^{n} s_i \le t], n \ge 1 \end{cases}$

In a system the customers waiting time distribution function is provided by,

$$P(\omega \le t) = \begin{cases} P_0 = 1-\rho, n = 0\,, t = 0 \\ P\left(\sum_{i=1}^{n} s_i \le t\right), t \ge 0, n \ge 1 \end{cases}$$

Each customer's service time is independently & identically distributed, so, $\Psi_s(t) = \mu e^{-\mu t}$, Is the probability density function for each customer, where it is equal to the mean service rate.

Now, $\phi_n = \sum_{i=1}^{n}(P_n \times$probability that (n-1) customers were served at time t × probability that one customer is receiving service at time Δt) (8)

We know that $P_n = \left(1 - \frac{\lambda}{\mu}\right)$. Also the probability that (n-1) customers where served at time t follows poisson distribution. i.e $\frac{(\mu t)^{n-1} e^{-\mu t}}{(n-)!}$.

The probability that one customer is receiving service at Δt=$\mu \Delta t$

Hence the equation (8) becomes,

$$\phi_n = \sum_{i=1}^{n} \frac{[(\mu \cdot t)^{n-1} e^{-\mu t}]}{(n-1)!} \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^n \mu \Delta t$$

Hence $\phi_\omega(t)$ becomes,

$$\phi_\omega(t) = P(w \leq t) = \begin{cases} 1 - \rho, & t = 0 \\ \sum P_n \int_0^t \phi_h(t)\, dt \end{cases}$$

$$= \begin{cases} 1 - \rho, \quad t = 0 \\ \sum_{i=1}^{n} \quad \rho^n (1 - \rho) \int_0^t \quad \frac{[(\mu \cdot t)^{n-1} e^{-\mu t}]}{(n-1)!} \end{cases}$$

$$= \begin{cases} 1 - \rho, \quad t = 0 \\ \rho(1 - \rho) \int_0^t \mu e^{-\mu(1-\rho)t}\, dt \end{cases}$$

This shows that the waiting time distribution has a discontinuity at t= 0 and is continuous from 0 to infinity. Hence we get, $\phi_\omega(0) = 1 - \rho, t = 0$

And $\phi'_\omega(t) = \lambda(1 - \rho) e^{-\mu(1-\rho)t}\, dt\, , t > 0$

Calculating the busy period distribution

Consider w to be the variable representing the total time, both for service and while waiting in the system. Then we have the probability density function for the distribution is given by $\phi(w) = \frac{d}{dt}\phi_\omega(t) = \frac{\lambda\left(1-\frac{\mu}{\lambda}\right)e^{-(\mu-\lambda)t}}{\int_0^\infty \lambda\left(1-\frac{\lambda}{\mu}\right)e^{-(\mu-\lambda)t}\,dt} = \frac{\lambda\left(1-\frac{\lambda}{\mu}\right)e^{-(\mu-\lambda)t}}{\frac{\lambda}{\mu}}$

$$= (\lambda - \mu)e^{-\mu(1-\lambda)t}\,dt, t = 0$$

Thus the busy period distribution is given by,

$$\int_0^\infty \phi_\omega\, dt = \int_0^\infty (\mu - \lambda)e^{-(\mu-\lambda)t}\, dt = 1$$

3.2. MODEL 2:{(M/M/1):(∞/SIRO)}

Second model of single server is {(M/M/1):(∞/SIRO)}.It is also unlimited queue. SIRO means service in random order. The sole distinction between this model and model 1 is the queue discipline. Since there is no particular queue discipline that effects the derivation of P_n,even with this model we possess the following ;

$P_n = (1- \rho)\, \rho^n$ for n= 1,2,3,…………

For as long as P_n stays , unchanged other outcomes will likewise be unchanged.

3.3.Model 3:{(M/M/1):(N/FCFS)} Exponential service-finite queue

We are assuming same assumption in model 1 in model 3,with the exception that the system's capacity is limited to N customers .This suggests that no more customers will be permitted to join the system after the length of N users .

Physical limitations can create a limited queue, as in the case of an emergency department in a hospital or a one-man barber shop with a set of number of chairs waiting customers.

The differential equation for model 3 is same as the differential equation obtained in model 1 as there is only change that n<N

By equation 4 , we have ,$\frac{d}{dt}p(t) = \mu p(t) - \lambda P(t)$

By steady state for model 3 is also same as model 1

That is $\lambda P_0 = \mu P_1$ for n=0

By equation 5,we have

$\mu P_{n+1} - (\mu+\lambda)P_n + \lambda P_{n-1} = 0$

Here in this equation for model 3,

We put n=1,2,3..........N-1 and then equation becomes

$\lambda P_{N-1} = \mu P_N$ for n =N

To achieve steady state conditions in this scenario, the service rate need not surpass the arrival rate ,that is *($\mu > \lambda$).*

The probability that a customer will be in the system for n =0,1,2,……..N is determined using the standard process and the first two difference equations

By equation 7, we have***;***

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0 \qquad \text{for } n \leq N$$

We know that;

$$\sum_{n=0}^{N} P_n = 1 = \sum_{n=0}^{N} \left(\frac{\lambda}{\mu}\right)^n P_0$$

$$= \sum_{n=0}^{N} \rho^n p_0$$

$$1 = P_0 \sum_{n=0}^{N} \rho^n$$

$$= P_0 \left[1 + \rho + \rho^2 + \cdots + \rho^N\right]$$

$[1 + \rho + \rho^2 + \cdots + \rho^N$ is a geometric series. Therefore it's sum is $\frac{1-\rho^{N+1}}{1-\rho}$ *]*

Therefore $1 = P_0\left[\frac{1-\rho^{N+1}}{1-\rho}\right]$

$$P_0 = [\frac{1-\rho^{N+1}}{1-\rho}] \qquad \rho \neq 1 \text{ and } \rho = \frac{\lambda}{\mu} < 1$$

$$P_n = \begin{cases} (\frac{1-\rho}{1-\rho^{N+1}})\rho^n & \text{for} \quad n \geq N \quad \frac{\lambda}{\mu} \neq 1 \\ \frac{1}{N+1} & \text{for} \quad \frac{\lambda}{\mu} = 1 \quad \rho = 1 \end{cases}$$

In this case, the steady state solution also holds for a value greater than 1.This is because the system has a finite capacity.

$$p_n = (1-\rho)\,\rho^n \qquad \textit{for } \lambda < \mu \;\; \textit{and } N \to \infty$$

That is same as the model 1.

3.4.SOME OTHER QUEUING MODELS

In queuing systems, various types of servers exist, each with its own characteristics, other Common server types include:

Multiple-Server System:

A multi-server queuing system is a model used to analyse and understand the behaviour of queues where multiple servers are available to serve incoming requests or customers. In such systems, multiple servers operate concurrently, each capable of handling one customer at a time. This setup aims to improve efficiency and reduce waiting times compared to a single-server system.

One key advantage of a multi-server queuing system is its ability to distribute the workload among several servers, preventing bottlenecks and reducing overall service times. This is particularly beneficial in scenarios where there is variability in service times or arrival rates of customers. The distribution of work among multiple servers helps to balance the load, ensuring a more consistent and predictable performance.

Parallel Servers:

In queuing theory, a parallel server model involves multiple servers operating concurrently to serve incoming requests or customers. This configuration aims to improve system efficiency and reduce customer wait times by distributing the workload among several servers. Each server in the parallel system operates independently, handling its own queue of requests. This parallel structure enhances the overall throughput of the system, as multiple servers can work simultaneously to process incoming tasks.The advantages of parallel servers in queuing models include increased capacity, better resource utilization, and enhanced resilience against server failures.

Random Servers:

Models are employed to analyse and optimize systems involving waiting lines or queues. Queuing Model 2, typically known as the M/M/1 queue, represents a simple but widely used scenario where entities, such as customers or requests, arrive following a Poisson process, are served according to a memoryless exponential service time distribution, and there is a single server. The term "M/M/1" denotes the arrival process (Markovian/Markovian), service process (Markovian), and the number of servers (1). This model allows for the exploration of various performance metrics, including system utilization, average queue length, and response time, providing insights into the efficiency and capacity of the system.

Random servers in queuing models may refer to scenarios where servers are subject to variability in their service rates or where the number of servers fluctuates over time. This introduces an additional layer of complexity to the queuing model, as the service rates are no longer constant. Analysing such systems involves considering random variables, possibly following specific probability distributions, to model the stochastic nature of server performance.

Chapter 4

APPLICATIONS OF QUEUING THEORY

4.1 THE SIGNIFICANCE OF QUEING THEORY

The significance of queuing theory lies in its ability to explain queue attributes like average wait time and provide tools for queue optimization. For instance, without a way to queue up users or modify the speed at which requests are processed, a website with a high volume of traffic will eventually crash and slow down.Think about aircraft awaiting touchdown on a runway. There are real safety risks when multiple planes attempt to land simultaneously due to a lack of a queue.

Little's law provides valuable insights since it makes it possible to solve for significant variables, such as the number of customers in line or the average wait time, using only two more inputs. Without knowing any other details about the queue, Little's law establishes a connection between a queuing system's capacity, average time spent in the system, and average arrival rate. The equation is provided by $L=\lambda$ W, where lambda is the average rate of new consumers entering the system, W is the average amount of time spent in the system, and L is the average number of customers in the system.

Several industries have used queuing theory, including telecommunications, transportation, logistics, finance, emergency services, industrial engineering, computing, project management, and operation research.

Little's law is applied in a few different contexts, including in a cafe line Little's law, for instance, can predict how long it will take to receive your coffee if you are in line at Chicking.

Let's say there are two customers serviced every minute, one server, and fifteen people in queue. You would apply Little's law to calculate the estimated wait time, which is 15 persons in line divided by 2 people serviced each minute, or 7.5 minutes.

Optimizing military processes: The military had to figure out how long B-2 stealth bombers should be in repair in this real-world scenario. Because there are only twenty B-

2 aircraft, they must be prepared quickly. On the other hand, they need regular maintenance, which might take 18 to 45 days at a time. The balance between aircraft in use and aircraft undergoing maintenance can be found by applying Little's law.

Three B-2 bombers were expected to be undergoing maintenance at any given time, according to a flight schedule study. A estimated 7-day interval was also determined for the rate at which bombers entered maintenance. Thus:

L = number of maintenance items = 3, Arrival/departure rate (λ) = 1/7 day

What is the mean amount of time spent in maintenance?

This gives us, when we apply Little's law, W= 3/(1/7)=21 days.

Therefore, in order to satisfy the needs of both the available aircraft and the regular flight schedules, the target lead time for B-2 bomber maintenance required to be 21 days.

4.2 DAILY LIFE APPLICATIONS OF QUEING THEORY

In the field of computer science:

The role of queueing theory in computer science and in operations research. In computer science, queueing theory is used to evaluate the performance of computers, networks, and queues-related algorithms, including task scheduling, allocation of resources, packet switching, etc. It helps in the design of efficient systems, prediction of response times, and optimization of resource utilization.

ATMs for banks:

Customers come to ATMs at random, and so does the service time—that is, the amount of time it takes a consumer to complete a transaction. The arrival rate, service rate, utilization rate, waiting time in the line, and average number of clients in the line are all calculated using the queuing model. By anticipating how many people are in line, bank ATMs can improve the quality of their services through queuing.

Medical Facilities:

A queuing system contributes to reducing patient wait times and optimizing the use of hospital beds, nurses, and other staff. Although queuing is not new, hospitals have just started to use it efficiently.

System of traffic:

To cut down on delays on the roads, queuing theory could be used to minimize the flow of vehicular traffic. It is impossible to overstate the importance of transportation in human life. A simple queuing theory-based vehicle traffic model. In order to lessen traffic congestion on the roadways, it will assess when it is optimal to turn on or off the red, amber, and green lights. In addition to lowering fuel usage, queuing also frees up funds that the government can use to address issues in other economic sectors.

Railway stations:

It is usually challenging to get confirmed tickets for the trip, especially in a country like India where trains are among the most widely used and least expensive forms of transportation. The number of trains operating various routes, particularly those connecting the major cities, is not commensurate with the population of the nation. Indian Railway is attempting to supply more than a billion people's ever-increasing demand. The purpose of the queue system is to prevent passenger inconvenience, and it works well and produces useful outcomes.

4.3. SOME EXAMPLES BASED ON QUEUING THEORY:

1. In a shop that has self-service system, with one cashier, 8 customers arrive on an average in every 5 mins. The cashier can serve 10 customers within 5 mins. If it is given that the arrival and the service time are exponentially distributed, then calculate the following:

a) Average number of customers waiting in the queue for receiving service

b) Expected waiting time in the queue

c) What is the probability of having more than 6 customers in the system?

Solution

Mean arrival rate =λ =8 x 12 = 96 customers / hour

Mean service rate = μ = 10 x 12 =120 customers / hour.

a) Average number of customers waiting in the queue for receiving service is given by $L_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{96^2}{120(120-96)}$

= 3.2 customers = 3 customers(approximate)

b) Expected waiting time in the queue is given by

$$Wq = \frac{L_q}{\lambda} = \frac{3.2}{96} = 0.033$$

c) The probability of having more than 6 customers in the queuing system is given by

$$P_{6\ or\ more} = \left(\frac{\lambda}{\mu}\right)^{r+1} = \left(\frac{\lambda}{\mu}\right)^{6+1} = \left(\frac{96}{120}\right)^{7} = 0.209$$

2. According to a television repairman, the time spent on his jobs has an exponential distribution with a mean of thirty minutes. What happens when the arrival of sets follows a Poisson distribution with an approximate average rate of 10 per 8-hour day, and he repairs the sets in the order that they arrive?

What is the projected daily idle time for the repairman? The number of jobs that are ahead of normal has just brought inside?

Solution

$\lambda = \frac{10}{8} = \frac{5}{4}$ sets per hour and $\mu = (\frac{1}{30})\ 60 = 2$ sets per hour are what we got.

Estimated daily idle time for the repairman

Given that the amount of time the repairman spends working during an eight-hour day (the traffic intensity) is determined by:

$(8)\ (\frac{\lambda}{\mu}) = (8)\ (\frac{5}{8}) = 5$hours

As a result, a repairman's idle time during an 8-hour workday is equal to (8 – 5) = 3 hours.

Estimated (or mean) number of TV sets in the setup $L_s = \frac{\lambda}{1-\lambda} = \frac{5/4}{2-5/4} = \frac{5}{3} = 2(approx.)$

3. There is just one loading dock in a warehouse, and it is staffed by a workforce of three. Four trucks an hour on average arrive at the loading port; the arrival rate is Poisson distributed. An exponential distribution can be assumed for the average time required to load a truck, which is 10 minutes. The cost of operating a truck is Rs 20 per hour, and the

loading crew members get paid Rs 6 each hour. Would you suggest that the owner of the vehicle hire another crew of 3 members?

Solution

Based on the problem's data, we may determine that $\lambda = 4$ /hour and $\mu = 6$ / hours respectively.

Regarding Current Crew

Total hourly cost=Loading crew costs +waiting time costs

= {(Total loaders) × (Hours of wage)} + {(Averaged waiting time per truck, Ws) (Averaged arrival rate per hour, λ) × (Cost of waiting an hour)}

$=6 \times 3 + \frac{1}{6-4} \times 4 \times 20 = 58$ Rs. per hour

Considering the Proposed Crew

Total hourly cost $=6 \times 6 + \frac{1}{12-4} \times 4 \times 20$

= 46 Rs per hour

The owner of the truck must hire three more loaders since their combined hourly wage will be less than the current rate after the new crew is added.

4. At a one man barber shop customers arrive according to probability distribution with a mean arrival rate of 5/hr. The hair cutting time is exponential distribution with a hair cut taking 10 min on an average assuming that the customers are always willing to wait find:

a) Average number of customer in the shop

b) Average waiting time of a customer

c) The percent of time an arrival Can walk right with out having to wait

d) The probability of a customer waiting more than 5mins

Solution

λ=5/hr, μ= $(\frac{1}{10}) \times 60$, Mean service rate=6/hr

(a) Average number of customers in the shop.

$L_s = L_q + \frac{\lambda}{\mu} = \frac{\lambda^2}{\mu(\mu-\lambda)} + \frac{\lambda}{\mu} = \frac{5^2}{6(6-5)} + \frac{5}{6}$

L_s=5 customers

(b) Average waiting time of a customer.

$W_q = \frac{L_q}{\lambda} = \frac{\lambda^2}{\mu(\mu-\lambda)\lambda} = \frac{5^2}{6(6-5)5} = 0.833$ hr

(c) Percent of time arrival can walk right without having to wait.

$P = [\,1 - \frac{\lambda}{\mu}] \times 100 = 1 - \frac{5}{6} \times 100 = 16.66\%$

d) Probability of a customer waiting more than 5mins.

$t = \frac{5}{60} = \frac{1}{12}$

$\therefore P = \frac{\lambda}{\mu} e^{-(\mu-\lambda)t} = \left(\frac{5}{6}\right) e^{\left(\frac{5-6}{12}\right)} = 0.766$

REFERENCES

[1] CINLAR E ,"*Introduction to stochastic process*", prentice hall, New Jersey(1975)

[2] FELLER W, "*Introduction to probability Theory and its application*", Academic Press, New York.(1968)

[3] J K SHARMA, "*Operation Research, theory and application,6 th edition*"Trinity

[4] ROBERT R COOPER, "*An Introduction to queuing theory*" , North Holland (1981)