

**An-Najah National University
Faculty of Graduate Studies**

**An Inventory Control Model with (M/M/1) Queueing
System and Lost Sales**

By

Imad Ramzi Mohammed Jomah

Supervisor

Dr. Mohammed N. Asad

**This Thesis is Submitted in Partial Fulfilment of the Requirements for
the Degree of Master of Science in Computational Mathematics,
Faculty of Graduate Studies, at An-Najah National University, Nablus,
Palestine.**

2011

Dedications

بسم الله الرحمن الرحيم

**An Inventory Control Model with (M/M/1) Queueing System
and Lost Sales**

To my parents, my wife and my children
To every one whom I love and respect.

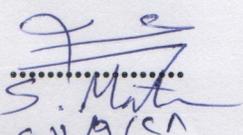
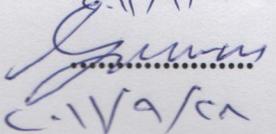
By

Imad Ramzi Mohammed Jomah

This Thesis was defended successfully on 14/09/2011 and approved by:

Defense Committee Members

Signature

- | | | |
|-------------------------|---------------------|---|
| 1- Dr. Mohammed N. Asad | (Supervisor) | 
.....
28.9.2011 |
| 2- Dr. Saed Mallak | (External Examiner) | 
.....
28.9.2011 |
| 3- Dr. Samir Matar | (Internal Examiner) | 
.....
S. Matar
8.11.2011 |
| 4- Dr. Loai Malhees | (Member) | 
.....
C.11/2011 |

Dedications

بسم الله الرحمن الرحيم

To my parents, my wife and my kids.

To every one whom I love and respect.

ACKNOWLEDGEMENT

I would like to thank my teacher and supervisor Dr. Mohammed N. Asad for his support, guidance and for his help with all mathematical models and operation research which based on his wide experience in teaching and research. Thanks and appreciation to the defence committee members Dr. Saed Mallak, Dr. Samir Matar and Dr.loai Malhees for their time and patience. Also I would like to thank all the Mathematics Department members who taught me and guide me to the knowledge. Also I would like to thank my colleague Haythem Joma for his help and support.

Special thanks and special appreciation to the one who encouraged during my research work wife Nada, and to my kids Jomana and Salah Al-dein for their support, and patience.

Finally, I would like to thank my parents, my brothers, and my sister for their support.

إقرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

An Inventory Control Model with (M/M/1) Queueing System and Lost Sales

أقر بان ما اشتملت عليه هذه الرسالة إنما هي نتاج جهدي الخاص ، باستثناء ما تمت الاشارة إليه
حيثما ورد ، وأن هذه الرسالة ككل ، أو أي جزء منها لم يقدم من قبل لنيل أية درجة أو لقب علمي
أو بحثي لدى أي مؤسسه تعليمية أو بحثيه أخرى.

Declaration

The work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted elsewhere for any other degree or qualification.

Student's name:

اسم الطالب :

Signature:

التوقيع :

Date:

التاريخ:

Table of contents

<i>Section</i>	<i>Subject</i>	<i>Page</i>
CH 1	Introduction	1
CH 2	Basic Concepts from Probability Theory	6
2.1	Random Variables	6
2.2	Probability, Conditional Probability and Independence	6
2.3	Some Probability Distributions as Models	7
2.3.1	Poisson Distribution	7
2.3.2	Geometric Distribution	8
2.3.3	Exponential Distribution	8
2.3.4	Binomial Distribution	9
2.3.5	Discrete Uniform Distribution	9
2.4	Memoryless Property of the Exponential Random Variable	9
2.5	Stochastic Process (Some Definitions)	10
2.6	Markov property	16
2.7	Markov Chain	16
2.8	Continuous-time Markov Process	17
2.9	Transition Probability Matrix	17
2.10	Stationary Distribution	18
CH 3	Basic Queueing Models & Inventory Systems	20
3.1	Terminology	20
3.2	Basic Queueing Model	22
3.3	Basic components of a queueing model	23
3.3.1	3.3.1 Input source (calling population):	23
3.3.2	3.3.2 Queue(waiting line):	27

<i>Section</i>	<i>Subject</i>	<i>Page</i>
3.3.3	Service Facility	29
3.4	Kendall's Notation	34
3.5	Steady State Solutions	36
3.6	Utilization	37
3.7	Performance Measures	39
3.8	Little's Law	42
3.9	Relationships Among Performance Measures	43
3.10	PASTA Property	44
3.11	Inventory Control System	45
3.12	General Inventory Model	45
3.13	Lost Sales and A backorder	47
3.14	Quality of Service (QoS)	49
3.15	Lead Time	49
CH 4	Single Server System with Inventory and Lost Sales	50
4.1	Introduction	50
4.2	Single Server System with Inventory and Lost Sales	50
4.2.1	Definition (The General Queueing-Inventory System)	52
4.2.2	Definition (Assumption on the Random Behavior of the System)	52
4.2.3	Definition (Reorder policy)	53
4.2.4	Definition ($M / M / 1 / \infty$ Queueing System with Inventory)	54
4.2.5	Theorem (Joint Queueing-Inventory Process)	54
4.2.6	Measures of System Performance	56
4.2.7	Cost Structure and Overall Profit	59
4.3	Examples for the Replenishment Order Size Distribution	60
4.3.1	Deterministic Order Size	60
4.3.2	Uniformly Distributed Order Size	62

<i>Section</i>	<i>Subject</i>	<i>Page</i>
4.3.3	Binomial Distributed Order Size	64
4.3.4	Geometry Distributed Order Size	66
4.4	Numerical Results	68
4.4.1	Total Cost Algorithm for all Performance Measures	68
4.4.1.1	Example	69
4.4.2	Impact Q, P Parameters on TC of Uniform, Binomial and Geometric	72
4.4.3	Results for our Research	83
4.5	Single Server System with Inventory and Backordering	83
CH 5	Conclusions	84
	References	86
	Appendixes	91

LIST OF TABLES

Number	Name	Page
Table(4.1)	Total cost respect to the Q parameter as $Q = 1:50$, $0 < p \leq 0.5$	73
Table(4.2)	Total cost respect to the Q parameter as $Q = 1:100$, $0 < p \leq 0.5$	74
Table(4.3)	Total cost respect to the Q parameter as $Q = 1:1000$, $0 < p \leq 0.5$	75
Table(4.4)	Total cost respect to the Q parameter as $Q = 1:50$, $1 \geq p > 0.5$	78
Table(4.5)	Total cost respect to the Q parameter as $Q = 1:100$, $1 \geq p > 0.5$	79
Table(4.6)	Total cost respect to the Q parameter as $Q = 1:1000$, $1 \geq p > 0.5$	80

LIST OF FIGURES

Number	Name	page
Fig(2.1)	A sample function of a counting process	11
Fig(2.2)	Counting process with time	12
Fig(3.1)	Components of a basic queueing process	23
Fig(3.2)	Single Server – Single Queue Model	29
Fig(3.3)	Single Server – Several Queue Model	30
Fig(3.4)	Several, Parallel Servers – Single Queue Model	31
Fig(3.5)	Several, Parallel servers – Several Queues Model	31
Fig(3.6)	Multiple Servers in a Series	32
Fig(4.1)	Interface for VB.NET program	69
Fig(4.2)	Deterministic in VB.NET	70
Fig(4.3)	Uniform in VB.NET	71
Fig(4.4)	Binomial in VB.NET	71
Fig(4.5)	Geometric in VB.NET	72
Fig(4.6)	Total cost respect to the Q parameter as $Q = 1:50, 0 < p \leq 0.5$	73
Fig(4.7)	Total cost respect to the Q parameter as $Q = 1:100, 0 < p \leq 0.5$	74
Fig(4.8)	Total cost respect to the Q parameter as $Q = 1:1000, 0 < p \leq 0.5$	75
Fig(4.9)	Total cost respect to the Q parameter as $Q = 1:1000, 0 < p \leq 0.5$, increasing lemnda	76
Fig(4.10)	Total cost respect to the Q parameter as $Q = 1:1000, 0 < p \leq 0.5$, increasing V	77
Fig(4.11)	Total cost respect to the Q parameter as $Q = 1:1000, 0 < p \leq 0.5$, decreasing lemnda, decreasing V	77
Fig(4.12)	Total cost respect to the Q parameter as $Q = 1:50, p > 0.5$	78
Fig(4.13)	Total cost respect to the Q parameter as $Q = 1:100, p > 0.5$	79
Fig(4.14)	Total cost respect to the Q parameter as $Q = 1:1000, p > 0.5$	80
Fig(4.15)	Total cost respect to the P parameter as $Q = 50$ fixed, $0 < p \leq 1$	81
Fig(4.16)	Total cost respect to the P parameter as $Q = 100$ fixed, $0 < p \leq 1$	82

**An Inventory Control Model with (M/M/1) Queueing System and
Lost Sales**

By

Imad Ramzi Mohammed Jomah

Supervised by

Dr. Mohammed N. Asad

Abstract

In this thesis We investigate M/M/1/ ∞ - queueing systems with inventory management, continuous review, and lost sales. Demand is Poisson, service times and lead times are exponentially distributed. These distributions are used to calculate performance measures of the respective system.

In case of infinite waiting room the key result is that the limiting distributions of the queue length processes are the same as in the classical M/M/1/ ∞ -system.

We compute performance measures and derive optimality conditions under different order distributions. Although we can completely determine analytically the steady state probabilities for the system.

We are able to derive functional relations for replenishment order size distributions that is in single server system with inventory. A computer programs were developed in this thesis to obtain the optimal policy.

Chapter One

Introduction

Everyone during his daily activities, has to stand in queues, whether in banks, government departments, petrol stations, or registry offices in universities and so forth. Queues have become common phenomena in contemporary societies.

In general we do not like to wait. But reduction of the waiting time usually requires extra investments. To decide whether or not to invest, it is important to know the effect of the investment on the waiting time. So we need models and techniques to analyse such situations[1].

Queueing Theory is one of the methods of operations research concerned with mathematical analysis of the positions that make up the lines waiting to find a suitable solution on them, often associated queueing theory with inventory models.

Queueing Theory is mainly seen as a branch of applied probability theory. Its applications are in different fields. For this area there exists a huge body of publications, a list of introductory or more advanced texts on queueing theory can be found in the bibliography[38]. Some good introductory books are [20], [21], [30].

The importance of inventory management for the quality of service (QoS) of today's service systems is generally accepted and optimization of systems in order to maximize QoS systems is therefore an important topic[28].

There are many different classical definitions of quality that connected with inventory availability ([35] p. 232). QoS characteristics are well established. But evaluation of these characteristics usually is done in models either from inventory theory or from queueing theory.

Berman and Sapna ([4-6]) investigate the behavior of service systems with an attached inventory. Their approach can be characterized as follows. Define a Markovian system process and then use standard optimization methods to find the optimal control strategy of the inventory or at least structural properties of the optimal policies[29].

All these models assume that the demand, which arrives during the time the inventory is zero, is backordered. The models differ with respect to the lead time, service time, and arrival distributions, waiting room size, order size and reorder policy. In all these models a continuous review for the inventory is assumed.

Mohebbi searched on a continuous-review inventory system with lost sales and variable lead time, and Turning to the impact of growth use of computer systems[23].

This thesis is devoted to present explicit performance measures for service facilities where demand for single item from a single server queueing system of M/M/1-type with an attached inventory under continuous review and lost sales.

We analyze single server queueing systems of M/M/1-type with an attached inventory. Customers arrive according to a Poisson process with intensity λ and each customer, who is served, needs exactly one item from the inventory and has an exponentially distributed service time with parameter μ . Consequently, the demand rate of the inventory is equal to λ if there are no customers waiting in queue otherwise the demand rate is equal to the service rate μ . The variable replenishment lead time, which is the time span between ordering of materials and receipt of the goods, is exponentially distributed with parameter ν .

The entire order is received into stock at the same time. The type of inventory system is defined to be a continuous review system where the inventory state is inspected after every single demand event and orders are placed every time the inventory on hand reaches a reorder point r . The on-hand stock is the stock that is physically on the shelf. The systems under investigation differ with respect to the size of replenishment orders and the reorder policy. Every system under consideration has the property that no customers are allowed to join the queue as long as the inventory is empty. This corresponds to the lost sales case of inventory management. However, if inventory is at hand, customers are still admitted to enter the waiting

room even if the number of customers in the system exceeds the inventory on hand [28].

The strategy of our investigation in this thesis is as follows:

We start from the basic concepts of probability theory and will discuss a number of important distributions which have been found useful for our study. Then we will cover basic concepts of queueing theory and inventory theory in chapter 3, we describe the basic queueing model. Then we discuss some important fundamental relations for queueing and inventory systems. In chapter 4 after we start from the observation of Berman and Kim [2],[3] whom proved that in an exponential system with zero lead times an optimal policy does not place an order unless the inventory is empty and a certain number of customers are waiting, we discuss and report our results. In chapter 5 we give the summary of our main results and conclusions. .

The objectives of this study are to:

- 1- discuss and activate Single server system ($M/M/1-\infty$) with inventory and lost sales(definitions and theorems),
- 2- find the steady state probability distribution and calculate the most important performance measures to our system,
- 3- Investigate four examples for the replenishment order size distribution for $M/M/1$ -type with an attached inventory,

4- compute the optimal total cost and the optimal total profit for each replenishment order size distribution,

5- use the computer with some program that was developed to find the analytic value for performance measures of any distribution,

6- find the optimal policy of our study by comparing between the four examples, using the computer with some program , graphs and tables.

Chapter Two

Basic Concepts from Probability Theory

This chapter was devoted to some basic concepts from probability theory and discussed a number of important distributions which have been found useful for describing random variables in many applications.

2.1 Random Variable

A random variable is a real valued function defined on the sample space. Random variables are denoted by capitals, X, Y, etc. The expected value or mean of X is denoted by $E(X)$ and its variance by $\sigma^2(X)$ where $\sigma(X)$ is the standard deviation of X [1].

2.2 Probability, Conditional Probability and Independence

Let x be a random variable that can assume only a finite number m of different values in the set $X = \{v_1, v_2, \dots, v_m\}$. We denote p_i as the probability that x assumes the value v_i [11]:

$$p_i = p_r \{ x = v_i \}, \quad i = 1, \dots, m.$$

Then the probabilities p_i must satisfy the following two conditions:

$$p_i \geq 0 \quad \text{and} \quad \sum_{i=1}^m p_i = 1.$$

When the random variable x can take values in the continuum, then the probability that $x \in (a, b)$: $p_r \{ x \in (a, b) \} = \int_a^b p(x) dx$

And so must satisfy the following two conditions:

$$p(x) \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} p(x) dx = 1.$$

Consider a sample space Ω . Let A be a set in Ω , the probability of A is the function on Ω , denoted $P(A)$, We use the notation $P(A/B)$ for the conditional probability of A given B, which is the probability of the event A given that we know that event B has occurred [11].

If events A and B are independent, which means that if one of them occurs, the probability of the other to occur is not affected, then [37]:

$$P(A \cap B) = P(A) P(B).$$

2.3 Some Probability Distributions as Models

There are many well-known probability distributions which are of great importance in the world of probability and all the applications on them, such as Poisson, exponential, binomial, uniform, Gaussian and geometric distribution. However we will discuss a number of important distributions which have been found useful for our study.

2.3.1 Poisson Distribution

The probability distribution of a Poisson random variable X with parameter μ representing the number of successes occurring in a given time interval or a specified region of space is given by the formula [1]:

$$P(X = n) = \frac{\mu^n e^{-\mu}}{n!}, \quad n = 0, 1, 2, \dots$$

For the Poisson distribution we have that : $E(X) = \sigma^2(X) = \mu$.

2.3.2 Geometric Distribution

The probability distribution of a geometric random variable X with parameter p is given by :

$$P(X = n) = (1-p) p^{(n-1)}, \quad n = 1, 2, \dots$$

For this distribution we have :

$$E(X) = \frac{p}{1-p}, \quad \sigma^2(X) = \frac{p}{(1-p)^2}.$$

2.3.3 Exponential Distribution

The probability density function $f(t; \mu)$ of an exponential distribution with parameter μ is given by[37]:

$$f(t; \mu) = \begin{cases} \mu e^{-\mu t} & t \geq 0 \\ 0 & , t < 0 \end{cases}$$

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

$$F(T) = \int_0^t \mu e^{-\mu t} dt = [-e^{-\mu t}]_0^t = -e^{-\mu t} + e^0 = 1 - e^{-\mu t} .$$

It is also described as : $F(T) = f(t \leq T) = 1 - e^{-\mu t}$, Where $F(T)$ is the area under the curve to the left of T . Thus $1 - F(T) = f(t \geq T) = e^{-\mu t}$.

If the area under the curve to the right of T . Thus we have the cumulative distribution function which given by :

$$F(t; \mu) = \begin{cases} 1 - e^{-\mu t} & , t \geq 0 \\ 0 & , t < 0 \end{cases}$$

For this distribution we have : $E(X) = 1 / \mu$, $\sigma^2(X) = 1 / \mu^2$.

2.3.4 Binomial Distribution

The binomial distribution is a discrete distribution described by the relationship as [30]:

$$P(x) = C_x^n p^x q^{n-x},$$

where $E(X) = np$, $\sigma^2(X) = npq$, $C_x^n = \binom{n}{x} p^x (1-p)^{n-x}$.

2.3.5 Discrete Uniform Distribution

If a random variable has any of n possible values k_1, k_2, \dots, k_n that are equally spaced and equally probable, then it has a discrete uniform distribution [16]. The probability of any outcome k_i is $1 / n$.

2.4 Memoryless Property of the Exponential Random Variable

An important property of an exponential random variable X with parameter μ is the memoryless property.

Memoryless Property states that " the future is independent of the past " i.e. the fact that it hasn't happened yet, tell us nothing about how much longer it will take before it does happen. In mathematical terms :

A variable X is memoryless with respect to t if for all $s \geq 0$ and $t \geq 0$;

$$P(X > s + t | X > t) = P(X > s) = e^{-\mu s}$$

Also $P(X > s + t) = P(X > s) P(X > t) = e^{-\mu s} e^{-\mu t} = e^{-\mu (s+t)}$.So the remaining lifetime of X , given that X is still alive at time t , is again exponentially distributed with the same mean $1/\mu$ [1].

We often use the memoryless property in the form : $P(X < t + \Delta t | X > t) = 1 - e^{-\mu \Delta t}$ [1].

2.5 Stochastic Process (Some Definitions)

A **Stochastic Process** (SP) is a family of random variables $\{X(t) | t \in T\}$ defined on a given probability space, indexed by the time variable t , where t varies over an index set T .

Just as a random variable assigns a number to each outcome s in a sample space S , a stochastic process assigns a sample function $x(t, s)$ to each outcome s , where a sample function $x(t, s)$ is the time function associated with outcome s of an experiment .

A stochastic process $\{X(t), t \geq 0\}$ is said to be a **Counting Process** if $X(t)$ represents the total number of "events " that have occurred up to time t .

A counting process $X(t)$ must satisfy the following conditions:

1. $X(t) \geq 0$ and $X(0) = 0$,
2. $X(t)$ is integer valued,
3. If $s < t$, $X(s) \leq X(t)$,
4. For $s < t$, $X(t) - X(s)$ equals the number of events that have occurred on the interval $(s, t]$ [8].

This figure shows a sample function of a counting process

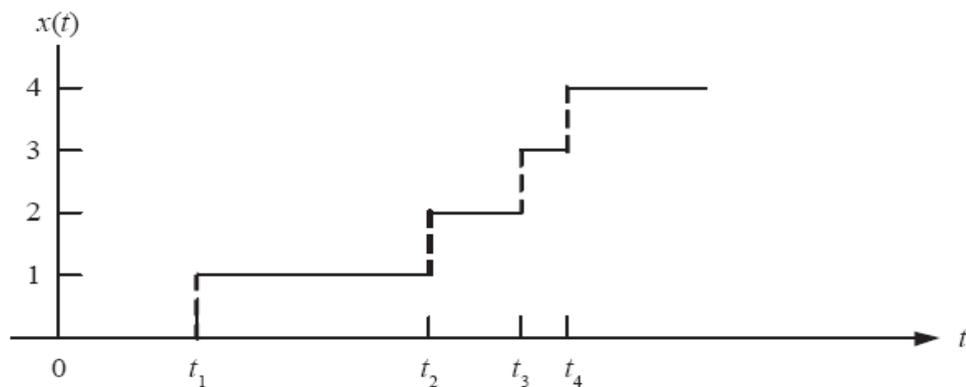


Figure 2.1 A sample function of a counting process

One of the most important types of counting processes is the renewal process, to clarify this definition consider a sequence of events which happen first at time $T_0 = 0$, then keep happening at random intervals, The events occur at times T_i ($i = 1, 2, \dots$), as figure below:

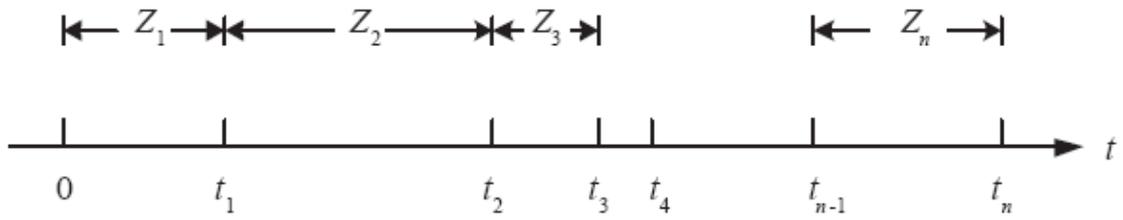


Figure 2.2 Counting process with time

The random variable T_i denotes the time at which the i th event occurs, and the values t_i of T_i are called **points of occurrence**. Suppose $Z_n = T_n - T_{n-1}$, then Z_n denotes the time between the $(n - 1)$ st and the n th events, often called renewal periods. The sequence of ordered random variables $\{Z_n, n \geq 1\}$ is sometimes called an **interarrival process**. If all random variables Z_n are independent and identically distributed, then $\{Z_n, n \geq 1\}$ is called a **Renewal process**. Renewal processes are useful for modelling streams of packets on a wire, jobs to be processed, etc. If a collection of random variables X_i all have the same distribution, and are independent of each other, then we say that the X_i are **independent and identically distributed** random variables. This is often expressed as *iid*.

A counting process $X(t)$ is said to possess **independent increment** if the number of events which occur in disjoint time intervals are independent. That is, for any $s > t > u > v > 0$, the random variable $X(s) - X(t)$, and the random variable $X(u) - X(v)$ are independent.

A counting process $X(t)$ is said to be a **Poisson process** with rate λ (> 0) if

1. $X(0) = 0$.
2. $X(t)$ has independent increments.
3. The number of events in any interval of length t " $[0, t]$ " has Poisson distribution with mean λt [41].

That is, for all $s, t > 0$,

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

This formula is the Poisson distribution with parameter λt , for which

$$E(X(t)) = \lambda t \quad ; \quad \sigma^2(X(t)) = \lambda t \quad .$$

These three conditions will be henceforth called the Three Poisson process conditions.

By definition, the Poisson process has what is called **stationary increments**. that is the number of events in the interval $(s + h, t + h)$ has the same distribution as the number of events in the interval (s, t) for all $t > s$ and $h > 0$. In both cases, the distribution is Poisson with parameter $\lambda(t - s)$. i.e. "the random variable $X(t + h) - X(s + h)$, has the same distribution for the random variable $X(t) - X(s)$ ".

Intuitively, if we choose the time interval $\Delta = t - s$ to be arbitrarily small (almost a "point" in time), then the probability of having an occurrence

there is the same regardless of where the "point" is . Loosely speaking, every point in time has the same chance of having an occurrence . Therefore, occurrences are equally likely to happen at all times . This property is also called **time-homogeneity** .

Clarify by symbols, it is easily verified that

$$P(\text{arrival in } (t, t + \Delta t]) = \lambda \Delta t + o(\Delta t), \quad (\Delta t \rightarrow 0)$$

Hence , for small Δt ,

$$P(\text{arrival in } (t, t + \Delta t]) \approx \lambda \Delta t .$$

So in each small time interval of length Δt the occurrence of an arrival is equally likely . In other words , Poisson arrivals occur completely random in time .

Another important property of the Poisson process is that the inter-arrival times of occurrences is exponentially distributed with parameter λ . This is shown by considering s to be an occurrence and T the time until the next occurrence , noticing that $P(T > t) = P(X(t) = 0) = e^{-\lambda t}$, and recalling the properties of independent and stationary increments. as a result, the mean interarrival time is given by $E[T] = 1 / \lambda$.

By the memoryless property of the exponential distribution , the time until the next occurrence is always exponentially distributed and therefore , at any point in time, not necessarily at points of occurrences , the future evolution of the Poisson process is independent of the past " The process

forgets its past ", and is always probabilistically the same. The Poisson process is therefore memoryless.

$$P(S \leq t + \Delta t | S > t) = P(S \leq t) \quad ; \quad s, t \geq 0 \quad .$$

Actually, the independence of the past can be explained also by the Poisson process property of independent increments, and the fact that the future evolution is probabilistically the same can also be explained by the stationary increments property .

We have shown that in the Poisson process, the interval between successive events are independent and identically distributed exponential random variables ' *iid* ' and we also identify the Poisson process as a renewal process with exponentially distributed intervals .

The Poisson process is an extremely useful process for modeling purposes in many practical applications, such as, e.g. to model arrival processes for queueing models or demand processes for inventory systems. It is empirically found that in many circumstances the arising stochastic processes can be well approximated by a Poisson process .

If a counting process $X(t)$ is a Poisson process then , for a small interval Δt , we have:

1. $P(X(\Delta t) = 0) = 1 - \lambda\Delta t + o(\Delta t)$
2. $P(X(\Delta t) = 1) = \lambda\Delta t + o(\Delta t)$
3. $P(X(\Delta t) \geq 2) = o(\Delta t) \quad .$

The above three conditions will henceforth be small interval conditions called one at a time.

There is another very important properties of a Poisson process, such as Merging property and Splitting property.

2.6 Markov Property

In probability theory and statistics, the terms Markov property refers to a property of a stochastic process. A stochastic process has the Markov property if the conditional probability distribution of future states of the process depends only upon the present state; that is, given the present, the future does not depend on the past. A process with this property is called Markov process [21].

2.7 Markov Chain

A Markov chain is a random process with the property that the next state depends only on the current state. That describe by a sequence of random variables X_1, X_2, X_3, \dots with the Markov property, namely that, given the present state, the future and past are independent. Formally,

$$P(X_{n+1} = x \mid X_1, X_2, \dots, X_n) = P(X_{n+1} = x \mid X_n)$$

The possible values of X_j from a countable set S called the state space.

2.8 Continuous-time Markov Process

In probability theory, a continuous-time Markov process is a stochastic process $\{ X(t) : t \geq 0 \}$ that satisfies the Markov property and takes values from set called the state space; it is the continuous-time version of a Markov chain. The Markov property states that at any times $s > t > 0$, the conditional probability distribution of the process at time s given the whole history of the process up to and including time t , depends only on the state of the process at time t . In effect, the state of the process at time s is conditionally independent of the history of the process before time t , given the state of the process at time t [25].

2.9 Transition Probability Matrix

In mathematics, a stochastic matrix, probability matrix, or transition matrix is used to describe the transitions of a Markov chain. It has found use in probability theory, statistics and linear algebra, as well as computer science.

There are several different definitions and types of stochastic matrices;

- A right stochastic matrix is a square matrix each of whose rows consists of nonnegative real numbers, with each row summing to 1.
- A left stochastic matrix is a square matrix whose columns consist of nonnegative real numbers whose sum is 1.

- A doubly stochastic matrix where all entries are nonnegative and all rows and all columns sum to 1.

If the state space discrete, the transition probability distribution can be represented by a matrix, called the transition matrix, with the (i, j) th element of P equal to:

$$p_{ij} = \Pr(X_{n+1} = j \mid X_n = i).$$

Since each row of P sums to one and all elements are non-negative, P is a right stochastic matrix. If the Markov chain is time-homogeneous, then the transition matrix P is the same after each step, so the k -step transition probability can be computed as the k -th power of the transition matrix P^k .

2.10 Stationary Distribution

Let X_n describe a Markov Chain having state space $\{1, 2, \dots, N\}$ and transition probability $P(i, j)$ from state i to state j . Vector π is called a stationary distribution, if elements of the vector π (possibly of infinite dimension) are non-negative numbers summing up to one, and if they satisfy the following equation:

$$\pi(i) = \sum_{j=1}^N \pi(j)P(j, i).$$

In matrix notation, this can be written as

$$\pi = \pi P.$$

Note that matrix notation is most useful for the finite case [17].

In other words, the stationary distribution π is a normalized (meaning that the sum of its entries is 1) left eigenvector of the transition matrix associated with the eigenvalue 1.

- Throughout the thesis we will assume that unless otherwise specified an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is given where all random variables are defined on.

Chapter Three

Basic Queueing Models & Inventory Systems

In this chapter we will cover basic queueing theory concepts and inventory systems. We describe the basic queueing model. Then we discuss some important fundamental relations for this model. Although we discuss the essential concepts of inventory systems. For this area there exist many papers and publications, a list of these or advanced texts on queueing theory is found in the references. Some good books on this topic are [20], [33].

3.1 Terminology

Customers - independent inputs or entities that arrive at random times to a server and wait for some kind of service, then leave. Since queueing theory is applied in different fields, also the terms job and task are often used instead customer.

Server - can only service one customer at a time, length of time to provide service depends on type of service, customers are served in Particular discipline.

Service Channel - mechanism by which to provide the required service, this may be a person (as a bank teller), or a machine, or a space (airport runway) or a team, or other.

Service Facility - also called the service center usually consists of a single service channel or multiple channels. The service center is often named processor or machine.

Service Capacity - there may be a single server or a group of servers helping the customers.

Queue - customers that have arrived at server but are waiting for their service to start are in the queue.

Queue Length at time t - number of customers in the queue at time t .

Queue Discipline - is the order or manner in which customers from the queue are selected for service.

Service Time - the time required to provide the service from the arrival of the customer to complete the requested service.

Waiting Time - for a given customer, how long that customer has to wait between arriving at the server and when the server actually starts the service.

Patient Customer - If a customer, on arriving at the service system stays in the system until served, no matter how much he has to wait for service.

3.2 Basic Queueing Model

The subject of queueing theory can be described as follows: consider a service centre and a population of customers, which at some times enter the service centre in order to obtain service. It is often the case that the service centre can only serve a limited number of customers. If a new customer arrives and the service is exhausted, he enters a waiting line and waits until the service facility becomes available. So we can identify three main elements of a service centre : a population of customers, the service facility and the waiting line. Also within the scope of queueing theory is the case where several service centres are arranged in a network and a single customer can walk through this network at a specific path, visiting several service centres. As a simple example of a service centre consider an airline counter: passengers are expected to check in, before they can enter the plane. The check-in is usually done by a single employee, however, there are often multiple passengers. A newly arriving and friendly passenger proceeds directly to the end of the queue, if the service facility (the employee) is busy[38].

The basic queueing model is shown in figure 3.1.

Components of a Basic Queuing Process

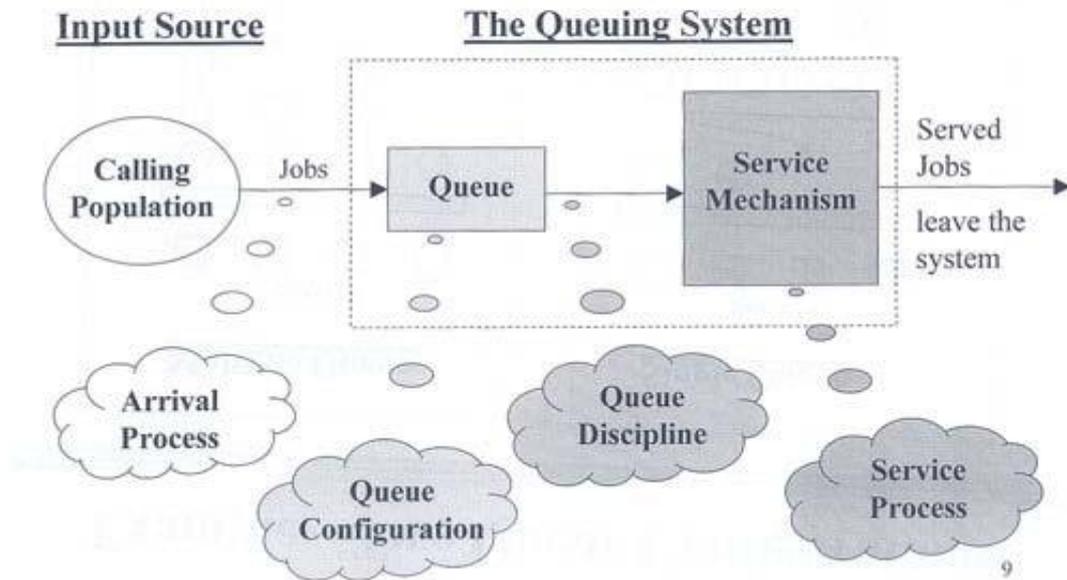


Figure 3.1: Components of a basic queuing process

Throughout this chapter there are some basic components in every queuing system which they are discussed next.

3.3 Basic Components of a Queueing Model

3.3.1 Input Source (calling population): These are potential customers of the system. The rate at which customers arrive at the service facility is determined by the arrival process. An input source is characterized by[12]:

a) Size of the Calling Population

The size represents the total number of potential customers who will require service.

According to source

The source of customers can be finite or infinite. For example, all people of a city or state (and others) could be the potential customers at a supermarket. The number of people being very large, it can be taken to be infinite. Whereas there are many situations in business and industrial conditions where we cannot consider the population to be infinite—it is finite.

According to numbers

The customers may arrive for service individually or in groups (in batches). Single arrivals are illustrated by patients visiting a doctor, students reaching at a library counter etc. On the other hand, families visiting restaurants, ships discharging cargo at a dock are examples of bulk, or batch arrivals.

According to time

Customers arrive in the system at a service facility according to some known schedule (for example one patient every 15 minutes or a candidate for interview every half hour) or else they arrive randomly. Arrivals are considered random when they are independent of one another and their occurrence cannot be predicted exactly. The queuing models wherein customers' arrival times are known with certainty are categorized as deterministic models. (insofar as this characteristic is concerned) and are easier to handle. On the other hand, a substantial majority of the queuing

models are based on the premise that the customers enter the system stochastically, at random points in time.

Because the calculations are far easier for the infinite case, this assumption often is made even when the actual size is some relatively large finite number, and it should be taken to be the implicit assumption for any queueing model that does not state otherwise. The finite case is more difficult analytically because the number of customers in the queueing system affects the number of potential customers outside the system at any time. However, the finite assumption must be made if the rate at which the input source generates new customers is significantly affected by the number of customers in the queueing system[20].

Often most queueing models assume that the customers population is of infinite size.

b) Arrival Pattern

Means the mechanism of the arrival of customers to the system. In view of the random nature of this process, making it difficult to predict accurately, it is resorting to probability distributions for this purpose. So we usually assume that the inter-arrival times are independent and have a common distribution (i.e. iid random variables). In many practical situations customers arrive according to a Poisson stream (i.e. exponential inter-arrival times).

c) **The Behaviour of Customers**

Customers may be patient and willing to wait (for a long time). Or customers may be impatient and leave after a while. For example, in call centres, customers will hang up when they have to wait too long before an operator is available, and they possibly try again after a while[1].

Now, Let us see some interesting observations of human behaviour in queues :

- **Balking** - Some customers even before joining the queue get discouraged by seeing the number of customers already in service system or estimating the excessive waiting time for desired service, decide to return for service at a later time. In queuing theory this is known as balking.
- **Reneging** - customers after joining the queue, wait for sometime and leave the service system due to intolerable delay, so they renege. For example, a customer who has just arrived at a grocery store and finds that the salesmen are busy in serving the customers already in the system, will either wait for service till his patience is exhausted or estimates that his waiting time may be excessive and so leaves immediately to seek service elsewhere.
- **Jockeying** - Customers who switch from one queue to another hoping to receive service more quickly are said to be jockeying [12].

Often most queueing models assume that the customers will join to waiting lines and do not leave until they are given the required service.

3.3.2 Queue(waiting line): a queue is characterized by the maximum permissible number that it can contain to stand. Thus there may be a single queue or multiple queues according to whether this number of input is infinite or finite.

In the waiting room there can be limitations with respect to the number of customers in the system. This is determined by length (or size) of the queue which depends upon the operational situation such as: waiting room (physical) space, legal restrictions and attitude of the customers.

The assumption of an infinite queue is the standard one for most queueing models, even for situations where there actually is a (relatively large) finite upper bound on the permissible number of customers, because dealing with such an upper bound would be a complicating factor in the analysis. However, for queueing systems where this upper bound is small enough that it actually would be reached with some frequency, it becomes necessary to assume a finite queue [20].

There are a number of ways in which customers in the queue are served (queue disciplines). Some of these are [30], [38]:

FCFS: customers are serviced on the first-come, first-served, that means a customer that finds the service center busy goes to the end of the queue.

LIFO: (Last in, First out): a customer that finds the service center busy proceeds immediately to the head of the queue. She will be served next, given that no further customers arrive.

Random Service: the customers in the queue are served in random order.

Round Robin: every customer gets a time slice. If her service is not completed, she will re-enter the queue.

Priority Disciplines: every customer has a (static or dynamic) priority, the server selects always the customers with the highest priority. This scheme can use preemption or not.

Some books and papers give the symbol (SIRO) for the order discipline: service in random order.

There exists another important queue discipline which is processor sharing (in computers that equally divide their processing power over all jobs in the system).

In our thesis for the queuing models that we shall consider, the assumption would be that the customers are serviced on the first-come, first-served basis (FCFS).

3.3.3 Service Facility: the service mechanism consists of one or more service channels. Service systems are usually classified in terms of their number of channels, or numbers of servers. And can be a channel service of a single stage or multistage. A service facility may include one person or several people operating as a team. Most elementary models assume one service facility with either one or a finite number of servers.

There are three aspects of a service facility: the configuration of the service facility, the service rate and the service time. Here we spell out;

a) Configuration of the service system : In the Configuration of the service system the first stage is the simplest, while the last stage is more complicated [12].

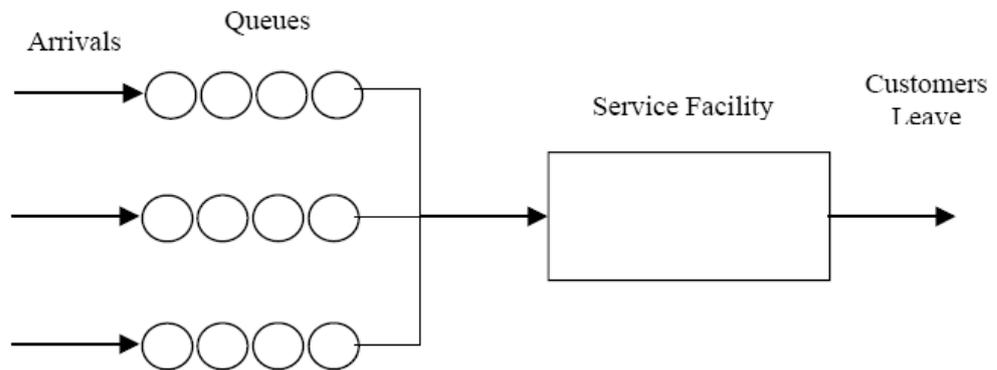
i) **Single Server – Single Queue** - The models that involve one queue – one service station facility are called single server models where customer waits till the service point is ready to take him for servicing. Students arriving at a library counter is an example of a single server facility [10].



Single Server – Single Queue Model

Figure 3.2: Single Server – Single Queue Model

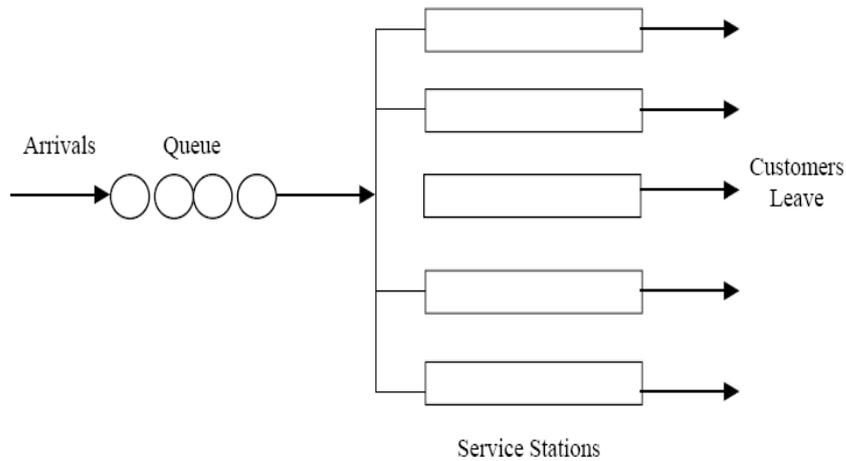
ii) **Single Server – Several Queues** – In this type of facility there are several queues and the customer may join any one of these but there is only one service channel.



Single Server – Single Queue Model

Figure 3.3: Single Server – Several Queue Model

iii) **Several (Parallel) Servers – Single Queue** – In this type of model there is more than one server and each server provides the same type of facility. The customers wait in a single queue until one of the service channels is ready to take them in for servicing.



Several, Parallel Servers – Single Queue Model

Figure 3.4: Several, Parallel Servers – Single Queue Model

iv) **Several Servers – Several Queues** – This type of model consists of several servers where each of the servers has a different queue. Different cash counters in an electricity office where the customers can make payment in respect of their electricity bills provide an example of this type of model.

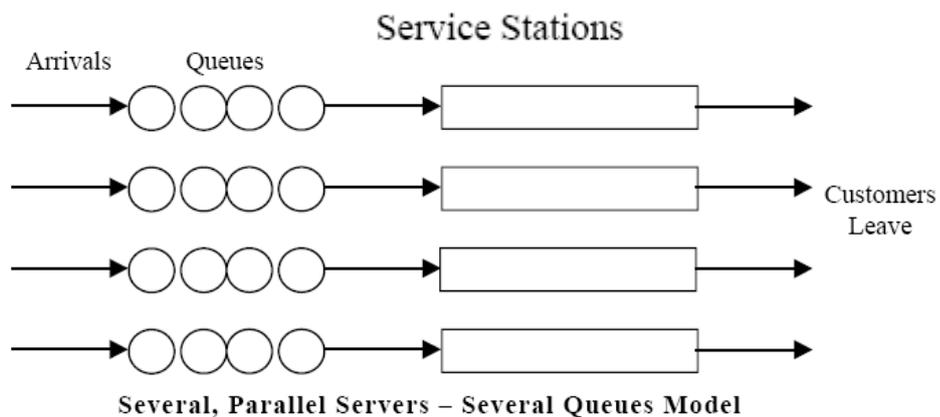
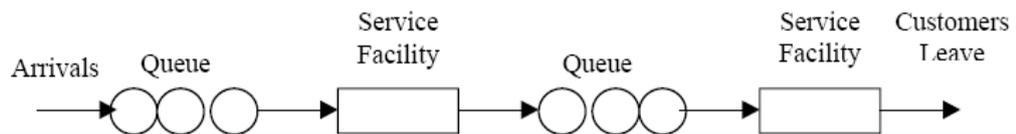


Figure 3.5: Several, Parallel servers – Several Queues Model

v) **Service Facilities in a Series** – In this, a customer enters the first station and gets a portion of service and then moves on to the next station, gets some service and then again moves on to the next station. and so on, and finally leaves the system, having received the complete service. For example, machining of a certain steel item may consist of cutting, turning, knurling, drilling, grinding, and packaging operations, each of which is performed by a single server in a series.



Multiple Servers in a Series

Figure 3.6: Multiple Servers in a Series

b) The service rate

The service rate describes the number of customers serviced during a particular time period. For example, In the clinic, providing the service on an average 4 customers in an hour, the service rate would be expressed as 4 customers/hour and service time would be equal to 15 minutes/customer.

c) The service time

Generally, we consider the service time only. The service time indicates the time required to provide the service from the arrival of the customer to complete the requested service.

It is clear that in many cases we are having difficulty to determine this time for sure, here also show the random nature to determine the time of service, This requires resorting to the use of probability distributions to estimate the times of service, whether customers or channels that provide the service. Often assume that all models of queues having the same distribution for all channels of service.

Usually we assume that the service times are independent and identically distributed, and that they are independent of the inter-arrival times. For example, the service times can be deterministic or exponentially distributed. It can also occur that service times are dependent of the queue length. For example, the processing rates of the machines in a production system can be increased once the number of jobs waiting to be processed becomes too large [1].

In practice, the most widely used distribution of service time is exponential distribution. In our thesis, the model (Single server) will be used distribution of service time is exponential distribution.

Remark: we obtain that the Characteristics of queuing systems :

Arrival Process: the distribution that determines how the tasks arrives in the system.

Service Process: the distribution that determines the task processing time

Number of Servers: total number of servers available to process the tasks

3.4 Kendall's Notation

A commonly used shorthand notation, called Kendall notation, for such single queue models describes the arrival process, service distribution, the number of servers and the buffer size (waiting room) as follows[41]:

arrival process / service distribution / number of servers / waiting room

Based on the above Characteristics, queuing systems can be classified by the following convention (by symbols separated by slashes):

$$A / B / m / N - S$$

where A denotes the distribution of the inter-arrival time, B denotes the distribution of the service times, m denotes the number of servers, N denotes the maximum size of the waiting line in the finite, case (if $N = \infty$ then this letter is omitted) and the optional S denotes the service discipline used (FCFS, LIFO and so forth). If S is omitted the service discipline is always FCFS. For A and B the following abbreviations are very common[38]:

- M (Markov): this denotes the exponential distribution with $A(t) = 1 - e^{-\lambda t}$

and $a(t) = \lambda e^{-\lambda t}$, where $\lambda > 0$ is a parameter. The name M stems from the fact that the exponential distribution is the only continuous distribution with the Marko property, i.e. it is memoryless.

- D (Deterministic): all values from a deterministic “distribution” are constant,

i.e. have the same value.

- Ek (Erlanger-k): Erlangen distribution.
- Hk (Hyper-k): hyper exponential distribution.
- G (General): general distribution, not further specified[33].

As an example applied to Kendall's notation, a system with exponential inter-arrival and service times, one server and having waiting room only for N customers (including the one in service) is abbreviated by the four letter code:

M/M/1 - N.

M/M/1 is the most simple queueing system (with FCFS service) which can be described as follows: we have a single server, an infinite waiting line, the customer inter-arrival times are iid and exponentially (Poisson arrival) distributed with some parameter λ and the customer service times are also

iid and exponentially distributed with some parameter μ . This is of our thesis model.

3.5 Steady State Solutions

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

We are mainly interested in steady state solutions, i.e. where the system after a long running time tends to reach a stable state, e.g. where the distribution of customers in the system does not change (limiting or stationary distribution). This is well to be distinguished from transient solutions, where the short-term system response to different events is investigated (e.g. a batch arrival) [38].

3.6 Utilization

An important measure for queueing systems performance is the utilization, denoted ρ . It is the proportion of time that a server is busy on average. In many systems, the server is paid for its time regardless if it is busy or not.

If you have two identical servers and one is busy 0.4 of the time and the other 0.6. Then the utilization is 0.5. We always have that $0 \leq \rho \leq 1$. If we consider an $M/M/\infty$ queue (Poisson arrivals, exponentially distributed service times and infinite servers) and the arrival rate is finite, the utilization is zero because the mean number of busy servers is finite and the mean number of idle servers is infinite [41].

Consider a $G/G/1$ queue (that is, a single-server queue with arbitrary arrival process and arbitrary service time distribution, with infinite buffer). Let S be a random variable representing the service time and let the mean service time $E[S] = 1/\mu$, i.e., μ denotes the service rate. Further, let λ be the mean arrival rate ($1/\lambda$ is the expected inter-arrival time and $1/\mu$ is the expected service time). Assume that $\mu > \lambda$ so that the queue is stable, namely that it will not keep growing forever, and that whenever it is busy, eventually it will reach the state where the system is empty.

For a stable $G/G/1$ queue, we have that that $\rho = \lambda / \mu$. To show the latter let L be a very long period of time. The average number of customers (amount of work) arrived within time period L is: λL . The average number of customers (amount of work) that has been served during time period L is

equal to $\mu \rho L$ (ρ is utilization factor). Since L is large and the queue is stable, these two values are equal. Thus, $\mu \rho L = \lambda L$. Hence, $\rho = \lambda / \mu$ [24].

The amount of work arriving per unit time equals $\rho = \lambda E[S]$. The server can handle 1 unit work per unit time. To avoid that the queue eventually grows to infinity, we have to require that $\rho < 1$. We note that the mean queue length also explodes when $\rho = 1$, except in the D/D/1 system, i.e., the system with no randomness at all. If $\rho < 1$, then ρ is called the occupation rate or server utilization or traffic intensity, because it is the fraction of time the server is working [1].

Often, we are interested in the distribution of the number (of customers, jobs or packets) in the system. Consider a G/G/1 queue and let p_n be the probability that there are n in the system. Having the utilization, we can readily obtain p_0 the probability that the G/G/1 queue is empty. Specifically,

$$p_0 = 1 - \rho = 1 - \lambda / \mu$$

If we have a multi-server queue, e.g. G/G/ c , then the utilization will be defined as the overall average utilization of the individual servers. That is, each server will have its own utilization defined by the proportion of time it is busy, and the utilization of the entire multi-server system will be the average of the individual server utilization [41].

When λ_n (mean arrival rate of new customers when n customers are in system) is a constant for all n , this constant is denoted by λ . When the mean service rate per busy server μ_n (also represents combined rate at which all busy servers achieve service completions) is a constant for all $n \geq 1$, this constant is denoted by μ . In this case, $\mu_n = c\mu$ when $n \geq c$, i.e., when all c servers are busy.

In $G/G/c$, the queue will grow to infinity if $\lambda \geq c\mu$, except if it's a $D/D/n$ queue. And we have to require that $\rho < c$. Here the occupation rate (utilization) per server $\rho = \lambda / c\mu$. where ρ is the expected fraction of time the individual servers are busy.

3.7 Performance Measures

Relevant performance measures in the analysis of queueing models are:

- The distribution of the waiting time and the sojourn time of a customer. The sojourn time is the waiting time plus the service time.
- The distribution of the number of customers in the system (including or excluding the one or those in service).
- The distribution of the amount of work in the system. That is the sum of service times of the waiting customers and the residual service time of the customer in service.
- The distribution of the busy period of the server. This is a period of time during which the server is working continuously.

In particular, we are interested in mean performance measures, such as the mean waiting time and the mean sojourn time [1].

Some of the performance measures (operating characteristics of any queuing system) in general interest for the evaluation of the performance of an existing queuing system, and to design a new system in terms of the level of service.

Important Notations and Terminology:

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

\mathbf{n} = number of customers in the system (waiting and in service). Also called

state of the system.

Queue length = \mathbf{n} - number of customers being served.

$P_{\mathbf{n}}$ = probability of \mathbf{n} customers in the system.

$P_{\mathbf{n}}(t)$ = probability that exactly \mathbf{n} customers are in the system at time t .

Given number at time 0.

$N(t)$ = number of customers in the system at time t ($t \geq 0$).

λ = expected customer arrival rate of new customers when \mathbf{n} customers are in system or average number of arrivals per unit of time in the queuing system.

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

ρ = server utilization factor (the expected fraction of time for which server is busy).

s = number of service facilities (parallel channels) in the system.

N = maximum number of customers allowed in the queueing system.

L_s = expected number of customers in queueing system (waiting and in service).

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

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

W_q = expected waiting time in the queue.

departure rate: the mean number of customers whose processing is completed in

a single unit of time.

Response Time T : also known as the sojourn time, is the total time that a customer spends in the queueing system.

Response time = waiting time + sojourn time .

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

In some cases when arrival rate of customers in the system is more than the service rate, then a steady - state cannot be reached regardless of the length of the elapsed time [13].

3.8 Little's Law

Little's law is a general result holding even for G/G/1-Queues; it also holds with other service disciplines than FIFO. It establishes a relationship between the average number of customers in the system, the mean arrival rate and the mean customer response time (time between entering and leaving the system after getting service) in the steady state. The following derivation is from.

Assume that λ_n is a constant λ for all n. It has been proven that in a steady-state queueing process[22],

$$L = \lambda W.$$

Furthermore, the same proof also shows that $L_q = \lambda W_q$. Also $L_s = \lambda W_s$.

If λ_n are not equal, then λ can be replaced in these equations by λ' , the average arrival rate over the long run. The proof found in [32].

3.9 Relationships Among Performance Measures

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

$$E(n) = L_s = \sum_{n=0}^{\infty} np_n, \quad E(n-s) = L_q = \sum_{n=s}^{\infty} (n-s)p_n.$$

Some general relationships between the average system characteristics true for all queuing models are as follows[13]:

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

$$\begin{aligned} L_s &= L_q + \text{Expected number of customers in service} \\ &= L_q + \lambda / \mu \end{aligned}$$

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

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

$$W_s = W_q + 1 / \mu$$

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

$$L_s = \lambda W_s \text{ Or } W_s = L_s / \lambda .$$

(iv) similarly, $L_q = \lambda W_q \text{ Or } W_q = L_q / \lambda .$

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

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

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

$$W_s = L_s / \lambda , \quad W_q = W_s - 1 / \mu , \quad L_q = \lambda W_q .$$

3.10 PASTA Property

For queueing systems with Poisson arrivals, so for M//. systems, the very special property holds that arriving customers find on average the same situation in the queueing system as an outside observer looking at the system at an arbitrary point in time. More precisely, the fraction of customers finding on arrival the system in some state A is exactly the same

as the fraction of time the system is in state A. This property is only true for Poisson arrivals[1].

In general this property is not true. For instance, in a D/D/1 system which is empty at time 0, and with arrivals at 1, 3, 5, ... and service times 1, every arriving customer finds an empty system, whereas the fraction of time the system is empty is 1/2.

This property of Poisson arrivals is called PASTA property, which is the acronym for Poisson Arrivals See Time Averages. Intuitively, this property can be explained by the fact that Poisson arrivals occur completely random in time. A rigorous proof of the PASTA property can be found in[39].

3.11 Inventory Control System

An inventory control system is a process for managing and locating objects or materials. In common usage, the term may also refer to just the software components [14].

3.12 General Inventory Model

The inventory problem involves placing and receiving orders of given sizes periodically. From this standpoint, an inventory policy answers two questions:

1. How much to order?
2. When to order?

The basis for answering these questions is the minimization of the following inventory cost function:

$$\begin{aligned} (\text{Total inventory cost}) = & (\text{Purchasing cost}) + (\text{setup cost}) + (\text{Holding cost}) \\ & + (\text{Shortage cost}) \end{aligned}$$

1. Purchasing cost is the price per unit of an inventory item. At times the item is offered at a discount if the order size exceeds a certain amount, which is a factor in deciding how much to order.
2. Setup cost represents the fixed charge incurred when an order is placed regardless of its size. Increasing the order quantity reduces the setup cost associated with a given demand, but will increase the average inventory level and hence the cost of tied capital. On the other hand, reducing the order size increases the frequency of ordering and the associated setup cost. An inventory cost model balances the two costs.
3. Holding cost represents the cost of maintaining inventory in stock. It includes the interest on capital and the cost of storage, maintenance, and handling.
4. Shortage cost is the penalty incurred when we run out of stock. It includes potential loss of income and the more subjective cost of loss in customer's goodwill.

An inventory system may be based on periodic review (e.g., ordering every week or every month), in which new orders are placed at the start of each period. Alternatively, the system may be based on continuous review, where a new order is placed when the inventory level drops to a certain level, called the reorder point. An example of periodic review can occur in a gas station where new deliveries arrive at the start of each week. Continuous review occurs in retail stores where items (such as cosmetics) are replenished only when their level on the shelf drops to a certain level [33].

3.13 Lost Sales And A Backorder

The typical way service level is measured in industry is the demand filled over total demand. Unfilled demand becomes a backorder or lost sales. Lost sales demand is often not known or measured by the management[24].

When a demand occurs and the item is out of stock, often the customer will not wait for the stock to be replenished and thereby the demand is a lost sale (and not a backorder) [24].

The lost sales situation arises e.g. in many retail establishments [9], where the intense competition allows customers to choose another brand or to go to another store. This can be considered as a typical situation for being described by a pure inventory model. But there are other areas of applications, where lost sales models are appropriate as well. E.g. these

models apply to cases such as essential spare parts where one must go to the outside of the normal ordering system when a stockout occurs [7].

The essential spare part problem is central for many repair procedures, where broken down units arrive at a repair station, queue for repair, and are repaired by substituting a failed part by a spare part from the inventory. A similar problem arises in production processes where rough material items are needed to let the production process run. Both of these latter problems are usually modeled using pure service systems, but these queueing theoretical models neglect the inventory management. Lost sales are in these contexts known as losses of customers. There is a huge amount of literature on loss systems, especially in connection with teletraffic and communication systems, where losses usually occur due to limited server capacity or finite buffer space. But there is another occurrence of losses due to balking or renegeing of impatient customers. However, only in the essential spare part problem of repair facilities a sort of inventory at hand is considered [28].

Lost sales in inventory theory and losses of customers in queueing theory are technical terms for similar, even often the same, events in real systems. The difference is set by the appropriate model selection done by the investigators: Either emphasizing the inventory management point of view or emphasizing the service system's point of view, both cases mostly neglect the alternative aspect [28].

3.14 Quality of Service (Q o S)

In the field of computer networking and other packet-switched telecommunication networks, the traffic engineering term quality of service (Q o S) refers to resource reservation control mechanisms rather than the achieved service quality. Quality of service is the ability to provide different priority to different applications, users, or data flows, or to guarantee a certain level of performance to a data flow [15].

3.15 Lead Time

Lead time is the time interval between the initiation and the completion of a production process. Or the time that it would take a supplier to delivery goods after receipt of order.

Chapter Four

Single Server System with Inventory and Lost Sales

4.1 Introduction

This chapter is devoted to present explicit performance measures for service facilities where demand for single items from single server queueing systems of M/M/1-type with an attached inventory under continuous review and lost sales. We will rely on recently research and papers written on this subject as [28], [29]. We investigate four examples for the replenishment order size distribution for M/M/1-type with an attached inventory. For each of our examples we compute the steady state probability distribution and calculate the most important performance measures. Then we discuss the impact of its parameters on cost structure as well as availability measures and service grades for the inventory to directly enable cost optimization in an integrated model. We will explain this by tables and graphs to support our results through computerized programs.

4.2 Single Server System with Inventory and Lost Sales

We analyze single server queueing systems of M/M/1-type with an attached inventory. Customers arrive according to a Poisson process with intensity λ and each customer, who is served, needs exactly one item from the inventory and has an exponentially distributed service time with parameter μ . Consequently, the demand rate of the inventory is equal to λ if

there are no customers waiting in queue otherwise the demand rate is equal to the service rate μ . The variable replenishment lead time, which is the time span between ordering of materials and receipt of the goods, is exponentially distributed with parameter ν . The entire order is received into stock at the same time. The type of inventory system is defined to be a continuous review system where the inventory state is inspected after every single demand event and orders are placed every time the inventory on hand reaches a reorder point r . The on-hand stock is the stock that is physically on the shelf. The systems under investigation differ with respect to the size of replenishment orders and the reorder policy. Every system under consideration has the property that no customers are allowed to join the queue as long as the inventory is empty. This corresponds to the lost sales case of inventory management. However, if inventory is at hand, customers are still admitted to enter the waiting room even if the number of customers in the system exceeds the inventory on hand [28].

Let $Z = ((G(t), I(t)), t \geq 0)$ denote the joint queue length and inventory process.

Where $G(t)$ denote the number of customers present at the server at time $t \geq 0$, either waiting or in service, and $I(t)$ denote the on-hand inventory at time $t \geq 0$. The state space of Z is $E_Z = \{(n, k) : n \in \mathbb{N}_0, k \in \{0, \dots, M\}\}$, where M is the maximal size of the inventory, which depends on the order policy, see Definition 4.2.3. We shall henceforth refer to Z as the queueing-inventory process.

4.2.1 Definition (The General Queueing-Inventory System)

At a service system with an attached inventory undistinguishable customers arrive one by one and require service. There is a single server with unlimited waiting room under first come, first served (FCFS) regime and an inventory with maximal capacity of M (identical) items. Each customer needs exactly one item from the inventory for service, and the on-hand inventory decreases by one at the moment of service completion. If the server is ready to serve a customer which is at the head of the line and there is no item of inventory this service starts only at the time instant (and then immediately) when the next replenishment arrives at the inventory. Customers arriving during a period when the server waits for the replenishment order are rejected and lost to the system (“lost sales”) [28].

A served customer departs from the system at once and the associated item is removed from the inventory at this time instant as well. If there is another customer in the line and at least one further item in the inventory, the next service starts immediately [29].

4.2.2 Definition (Assumption on the Random Behavior of the System)

For the service system with inventory management from Definition 4.2.1 we assume:

Customers are of stochastically identical behavior. To the server there is a Poisson- λ -arrival stream, $\lambda > 0$. Customers request an amount of service time which is exponentially distributed with mean 1. Service is provided

with intensity $\mu > 0$ [28]. The replenishment lead time is exponentially distributed with parameter $\nu > 0$.

4.2.3 Definition (Reorder Policy)

From Definition 4.2.1 we consider the following policy for Single server system with inventory and lost sales :

If the inventory is depleted after the service of a customer is completed, then immediately a replenishment order is triggered.

The decision of the order size may be randomized according to a discrete probability density function p on the integers $\{1, 2, \dots, M\}$, where M is the maximal capacity of the inventory. So the size of a replenishment order is k with probability p_k , where p_k is a discrete probability function . Let F_p denoted to the discrete distribution function, then $\bar{F}_p := 1 - F_p$ its tail distribution function F_p . We abbreviate the probability that the size of a replenishment order is at least k units by q_k , i.e. $q_k = \bar{F}_p(k-) = \sum_{h=k}^M p_h$.

The mean order size is denoted by:

$$\bar{P} = \sum_{k=1}^M k \cdot p_k .$$

Where service times and inter-arrival times are independent and independent of the order size and lead times. All constitute an independent family of random variables.

Service system which has described by the previous definitions is the lost sale case of classical inventory management where customer demand is not

backordered but lost in case there is no inventory on hand [35]. We recall the backordered case in periphrastic discussion at the end of this chapter.

4.2.4 Definition (M / M / 1 / ∞ Queueing System with Inventory)

A service system with inventory according to Definition 4.2.1, with the stochastic assumptions of Definition 4.2.2, and under some prescribed policy from Definition 4.2.3, is called an M/M/1/∞ system with inventory management under that policy.

4.2.5 Theorem (Joint Queueing-Inventory Process)

For the M/M/1/∞ system with inventory according to Definition 4.2.4 the stochastic queueing-inventory process Z from Definition 4.2.1 is a homogeneous Markov process . Z is ergodic if and only if $\lambda < \mu$. If Z is ergodic then it has a unique limiting and stationary distribution of product form:

$$\pi(n, k) = K^{-1} \left(\frac{\lambda}{\mu} \right)^n q_k \quad \text{with } n \in \mathbb{N}, 1 \leq k \leq M, \quad (1)$$

$$\pi(n, k) = K^{-1} \left(\frac{\lambda}{\mu} \right)^n \frac{\lambda}{\nu} \quad \text{with } n \in \mathbb{N}, k=0 \quad (2)$$

$$\text{and with normalization constant } K = \frac{\mu}{\mu - \lambda} \left(\bar{P} + \frac{\lambda}{\nu} \right). \quad (3)$$

We note further that the normalization constant K factorizes in the normalization constant for the marginal queue length and for the inventory process as :

$$K = K_X \cdot K_Y \quad \text{with} \quad K_X = \frac{\mu}{\mu - \lambda} \quad \text{and} \quad K_Y = \bar{P} + \frac{\lambda}{\nu} .$$

The theorem 4.2.4 and proof was found in [28, p58]. Note that for $\nu = \infty$ the inventory is replenished instantaneously and the inventory position 0 is left immediately. Therefore, the stationary distribution has support $\mathbb{N} \times \{1, 2, \dots, M\}$ and is given by (1) with $K = \frac{\mu}{\mu - \lambda} \bar{P}$.

The strong restriction in our present model which described above (queueing is that we regulate reordering and admission of the customers only via the inventory level. Customers are only rejected (and lost), when the physical inventory level reaches zero [3]. A more sophisticated policy would include into the decision procedure information on the actual queue length at the feasible decision instant. The gain of posing our restriction on the reorder policy is the result of Theorem 4.2.5 [28].

There is a policy specified which determines at each decision point whether a replenishment order is placed or not, and how many items are ordered. We assume that there is always at most one outstanding order.

We will first carry out the calculations for the system with arbitrary random order size out of $\{1, 2, \dots, M\}$ and reorder point 0 then show the important measures of system performance. And we compute the steady state of the system for standard simple to implement policies, and then use the equilibrium probabilities to minimize asymptotic costs or maximize the overall profit = (revenue – costs).

4.2.6 Measures of System Performance

We are interested in stationary characteristics of the queueing-inventory system. These are long-run characteristics as well. Note that stationarity is always assumed in the classical inventory theory as well. Having determined the stationary distribution, we can compute several measures of operating characteristics for the system explicitly.

- The steady state on-hand inventory distribution of $Y = (Y(t), t \geq 0)$ is:

$$P(Y = k) = \begin{cases} K_Y^{-1} \frac{\lambda}{\nu} & , \text{ for } k = 0 \\ K_Y^{-1} q_k & , \text{ for } 1 \leq k \leq M \end{cases} \quad (4)$$

Here we have denoted by Y a random variable distributed like the stationary inventory distribution.

- The marginal steady state queue length distribution of $X = (X(t), t \geq 0)$ is equal to the steady state queue length distribution in the classical $M / M / 1 / \infty$ - FCFS system with the same parameters λ and μ . Therefore the mean number of customers in system is:

$$\bar{L}_0 = \frac{\lambda}{\mu - \lambda}.$$

And so that \bar{L} is the same as in the classical $M/M/1/\infty$ -system with

parameters λ, μ . Where \bar{L} is the mean number of the waiting customers,

$$\bar{L} = \frac{\lambda^2}{\mu (\mu - \lambda)}.$$

- The stationary average on-hand inventory position is given by:

$$\bar{I} = \sum_{k=1}^M k \sum_{n=0}^{\infty} \Pi(n, k) = K_Y^{-1} \sum_{k=1}^M k q_k \quad (5)$$

- The mean number of replenishments per time unit (reorder rate) is:

$$\lambda_R = \frac{\lambda}{\bar{p} + \frac{\lambda}{\nu}}. \quad (6)$$

- The mean number of customers arriving per unit time is:

$$\lambda_A = \bar{p} \lambda_R = \frac{\bar{p} \lambda \nu}{\bar{p} \nu + \lambda}. \quad (7)$$

- The β -service level is a quantity-oriented service measure describing the proportion of demands that are met from stock without accounting for the duration of a stockout. β -service levels are widely used in practice [34].

The definition of the β -service level is standard and can be found in

[27] and [31].

$$\beta\text{-service level} = 1 - \frac{\lambda}{\bar{p} \nu + \lambda}. \quad (8)$$

- The average number of lost sales incurred per unit of time is given by:

$$\overline{LS} = \frac{\lambda^2}{\bar{p} \nu + \lambda}. \quad (9)$$

- The expected number of lost sales per cycle is given by:

$$\overline{LS}_c = \frac{\overline{LS}}{\lambda_R} = \lambda/v. \quad (10)$$

- From little's formula the customers mean sojourn time \overline{w}_0 is:

$$\overline{w}_0 = \frac{\overline{L}_0}{\lambda_A} = \frac{\overline{p} \nu + \lambda}{\overline{p} \nu (\mu - \lambda)}. \quad (11)$$

- The mean waiting time \overline{w} is:

$$\overline{w} = \frac{\overline{L}}{\lambda_A} = \frac{(\overline{p} \nu + \lambda) \lambda}{\overline{p} \nu (\mu - \lambda) \mu}. \quad (12)$$

All of above performance measures proved and discussed in detail and can be found in [28].

Note that only \overline{w}_0 and \overline{w} depend on the service rate μ . They are naturally larger than the mean sojourn time and mean waiting time of classical M/M/1- ∞ system respectively. Some performance measures are not dependent on λ and ν individually but only on their proportion λ/ν , e.g. \overline{I} , \overline{LS}_c and β . Concerning the influence of F_p we observe that several performance measures only depend on the first moment of F_p like λ_R , \overline{LS} , β and \overline{w} or are completely independent of F_p like \overline{LS}_c . \overline{I} depends on the first and second moment of F_p . Hence, for two systems which have the same parameters λ , ν and μ but different order size distributions F_p and \overline{F}_p with the same mean \overline{p} only \overline{I} will be different.

4.2.7 Cost Structure and Overall Profit

There are costs connected with operating the system originating from both, the queueing of the customers and from holding inventory at the system. We have a fixed holding cost h per item and time unit in the inventory, a fixed ordering cost k for each replenishment order, a shortage cost ℓ per unit of lost sales, a cost ν_w per customer and time unit in the waiting room, and a cost ν_1 per customer and time unit in service. Whenever a customer's service is completed, a revenue R is paid to the system.

The cost structure of the $M / M / 1$ queueing system with inventory under lost sales is [28]:

$$TC = \lambda_R \cdot k + \bar{I} \cdot h + \overline{LS} \cdot \ell + \bar{L} \cdot \nu_w + \rho \cdot \nu_1 \quad (13)$$

Where the mean costs that occur in steady state per time unit are :

- $\lambda_R \cdot k$, the fixed costs associated to replenishment orders that occur with reorder rate λ_R ,
- $\bar{I} \cdot h$, the holding costs for inventory of mean size \bar{I} ,
- $\overline{LS} \cdot \ell$, the shortage costs for the mean number of lost sales \overline{LS} ,
- $\bar{L} \cdot \nu_w$, the waiting costs for the mean number \bar{L} of waiting customers,
- $\rho \cdot \nu_1$, the costs for the mean number ρ of customers in service.

The revenue obtained by the system's service is per time unit:

- $\lambda_A \cdot R$, the amount of money obtained from the served customers per time unit, which is proportional to the throughput.

$$\text{The overall profit} = \text{TC} - (\lambda_A \cdot R) \quad (14)$$

The our goal is to obtain the optimal policy by minimize asymptotic costs or maximize the overall profit for the M / M / 1 queueing system with inventory under lost sales in our study.

4.3 Examples for the Replenishment Order Size Distribution

In this section we investigate four examples for the replenishment order size distribution. We consider the fixed order size Q , which yields an $(0, Q)$ -policy and the system with uniform, binomial and geometry distribution order sizes on $\{1, \dots, Q\}$. In all cases holds $M = Q$. The performance measures for these examples, which could be obtained from equations(4.2.6) are summarized as in:

4.3.1 Deterministic Order Size

Fixed (deterministic) order quantities are described by using one-point distributions for the order size distribution. Let us assume that the order size is fixed and equal to $Q \in N$, then for all $k \in \{1, \dots, M\}$ we have $p_k = \delta_{kQ}$. Then

$$q_k = \begin{cases} 1 & , k \in \{1, \dots, Q\} \\ 0 & , otherwise \end{cases} \quad , \quad \text{The mean order size is } \bar{p} = Q.$$

The most important performance measures are :

- The stationary average on-hand inventory position is given by:

$$\bar{I} = \frac{Q}{Q + \frac{\lambda}{\nu}} \frac{Q + 1}{2}, \quad (15)$$

- The mean number of replenishments per time unit(reorder rate) is:

$$\lambda_R = \frac{\lambda}{Q + \frac{\lambda}{\nu}}, \quad (16)$$

- The mean number of customers arriving per unit time is:

$$\lambda_A = Q \cdot \lambda_R, \quad (17)$$

- The β -service level is:

$$\beta = \frac{Q}{Q + \frac{\lambda}{\nu}}, \quad (18)$$

- The average number of lost sales incurred per unit of time is given by:

$$\overline{LS} = \frac{\lambda^2}{Q \nu + \lambda}, \quad (19)$$

- The expected number of lost sales per cycle is given by:

$$\overline{LS}_c = \frac{\overline{LS}}{\lambda_R} = \lambda/\nu, \quad (20)$$

- From little's formula the customers mean sojourn time \overline{w}_0 is:

$$\overline{w}_0 = \frac{\overline{L}_0}{\lambda_A} = \frac{Q \nu + \lambda}{Q \nu (\mu - \lambda)}, \quad (21)$$

- The mean waiting time \bar{w} is:

$$\bar{w} = \frac{\bar{L}}{\lambda_A} = \frac{(Q + \nu + \lambda) \lambda}{Q + \nu (\mu - \lambda) \mu}, \quad (22)$$

- The steady state on-hand inventory distribution is:

$$P(Y=k) = \begin{cases} \frac{\lambda}{Q + \nu + \lambda} & , \quad k = 0 \\ \frac{\nu}{Q + \nu + \lambda} & , \quad 1 \leq k \leq Q \end{cases}. \quad (23)$$

4.3.2 Uniformly Distributed Order Size

Let the size of a replenishment order be equally distributed on $\{1, \dots, Q\}$, then its uniformly distributed on $\{1, \dots, Q\}$. Hence $p_k = \begin{cases} \frac{1}{Q} & , \quad k \in \{1, \dots, Q\} \\ 0 & , \quad otherwise \end{cases}$,

The mean order size is $\bar{p} = \frac{Q + 1}{2}$,

And $q_k = \sum_{h=k}^Q p_h = \frac{Q + 1 - k}{Q}$. The most important performance

measures are:

- The stationary average on-hand inventory position is given by:

$$\bar{I} = \frac{(Q+2) (Q+1) \nu}{3(Q+1)\nu + 6\lambda}, \quad (24)$$

- The mean number of replenishments per time unit(reorder rate) is:

$$\lambda_R = \frac{2\lambda\nu}{(Q+1)\nu + 2\lambda}, \quad (25)$$

- The mean number of customers arriving per unit time is:

$$\lambda_A = \frac{Q + 1}{2} \cdot \lambda_R, \quad (26)$$

- The β -service level is:

$$\beta = \frac{(Q+1)\nu}{(Q+1)\nu + 2\lambda}, \quad (27)$$

- The average number of lost sales incurred per unit of time is given by:

$$\overline{LS} = \frac{2\lambda^2}{(Q+1)\nu + 2\lambda}, \quad (28)$$

- The expected number of lost sales per cycle is given by:

$$\overline{LS}_c = \frac{\overline{LS}}{\lambda_R} = \lambda/\nu, \quad (29)$$

- From little's formula the customers mean sojourn time \overline{w}_0 is:

$$\overline{w}_0 = \frac{\overline{L}_0}{\lambda_A} = \frac{(Q+1)\nu + 2\lambda}{\nu(Q+1)(\mu-\lambda)}, \quad (30)$$

- The mean waiting time \overline{w} is:

$$\overline{w} = \frac{\overline{L}}{\lambda_A} = \frac{((Q+1)\nu + 2\lambda)\lambda}{\nu(Q+1)(\mu-\lambda)\mu}, \quad (31)$$

- The steady state on-hand inventory distribution is:

$$P(Y=k) = \begin{cases} \frac{2\lambda}{(Q+1)\nu + 2\lambda} , & k=0 \\ \frac{2\nu[Q+1-k]}{[(Q+1)\nu + 2\lambda]Q} , & 1 \leq k \leq Q \end{cases} . \quad (32)$$

4.3.3 Binomial Distributed Order Size

Let the size of a replenishment order is $k \in \{1, \dots, M\}$, the probability of binomial distribution is $p_k = \binom{Q}{k} p^k (1-p)^{Q-k}$, where p denote to the lost sales case of shortage, $(1-p)$ denote to without shortage state. Also p must be $0 < p \leq 1$. And $q_k = \sum_{h=k}^Q p_h$. The mean order size is $\bar{p} = Qp$.

The most important performance measures are:

- The stationary average on-hand inventory position is given by:

$$\bar{I} = \frac{Q\nu(Q-1)p^2 + 2Qp\nu}{2Qp\nu + 2\lambda}, \quad (33)$$

- The mean number of replenishments per time unit(reorder rate) is:

$$\lambda_R = \frac{\lambda\nu}{Qp\nu + \lambda}, \quad (34)$$

- The mean number of customers arriving per unit time is:

$$\lambda_A = Qp \cdot \lambda_R, \quad (35)$$

- The β -service level is:

$$\beta = \frac{Qp\nu}{Qp\nu + \lambda}, \quad (36)$$

- The average number of lost sales incurred per unit of time is given by:

$$\overline{LS} = \frac{\lambda^2}{Qp\nu + \lambda}, \quad (37)$$

- The expected number of lost sales per cycle is given by:

$$\overline{LS}_c = \frac{\overline{LS}}{\lambda_R} = \lambda/\nu, \quad (38)$$

- From little's formula the customers mean sojourn time \overline{w}_0 is:

$$\overline{w}_0 = \frac{\overline{L}_0}{\lambda_A} = \frac{(Qp\nu + \lambda)}{Qp\nu(\mu - \lambda)}, \quad (39)$$

- The mean waiting time \overline{w} is:

$$\overline{w} = \frac{\overline{L}}{\lambda_A} = \frac{(Qp\nu + \lambda)\lambda}{Qp\nu(\mu - \lambda)\mu}, \quad (40)$$

- The steady state on-hand inventory distribution is:

$$P(Y=k) = \begin{cases} \frac{\lambda}{Qp\nu + \lambda}, & k=0 \\ \frac{\nu}{Qp\nu + \lambda} q_k, & 1 \leq k \leq M \end{cases}. \quad (41)$$

4.3.4 Geometry Distributed Order Size

Let the size of a replenishment order is $k \in \{1, \dots, M\}$, the probability of geometry distribution is $p_k = p (1-p)^{k-1}$, where p denoted to lost sales case of shortage, $(1-p)$ denote to without shortage state.

Also p must be $0 < p \leq 1$. And $q_k = (1-p)^{k-1} - (1-p)^Q$.

The mean order size is $\bar{p} = \frac{1 - (Q+1)(1-p)^Q + Q(1-p)^{Q+1}}{p}$. The

most important performance measures are:

Let we have :

- $a = 1 - (Q+1)(1-p)^Q + Q(1-p)^{Q+1}$,
- $b = (Q^2 + 3Q + 2)(1-p)^Q$,
- $c = (3Q^2 + 7Q + 2)(1-p)^{Q+1}$,
- $d = (3Q^2 + 5Q)(1-p)^{Q+2}$,
- $e = (Q^2 + Q)(1-p)^{Q+3}$. Then
- The stationary average on-hand inventory position is given by:

$$\bar{I} = \frac{v(2p - b + c - d + e)}{2(p)^2 (va + \lambda p)}, \quad (42)$$

- The mean number of replenishments per time unit(reorder rate) is:

$$\lambda_R = \frac{\lambda v p}{va + \lambda p}, \quad (43)$$

- The mean number of customers arriving per unit time is:

$$\lambda_A = \frac{a}{p} \cdot \lambda_R, \quad (44)$$

- The β -service level is:

$$\beta = \frac{av}{av + \lambda p}, \quad (45)$$

- The average number of lost sales incurred per unit of time is given by:

$$\overline{LS} = \frac{p\lambda^2}{av + \lambda p}, \quad (46)$$

- The expected number of lost sales per cycle is given by:

$$\overline{LS}_c = \frac{\overline{LS}}{\lambda_R} = \lambda/v, \quad (47)$$

- From little's formula the customers mean sojourn time \overline{w}_0 is:

$$\overline{w}_0 = \frac{\overline{L}_0}{\lambda_A} = \frac{(av + \lambda p)}{av(\mu - \lambda)}, \quad (48)$$

- The mean waiting time \overline{w} is:

$$\overline{w} = \frac{\overline{L}}{\lambda_A} = \frac{(av + \lambda p)\lambda}{av(\mu - \lambda)\mu}, \quad (49)$$

- The steady state on-hand inventory distribution is:

$$P(Y=k): = \begin{cases} \frac{\lambda p}{av + \lambda p}, & k = 0 \\ \frac{vp((1-p)^{k-1} - (1-p)^0)}{av + \lambda p}, & 1 \leq k \leq M \end{cases} . \quad (50)$$

4.4 Numerical Results

In this section we discuss the effect of some parameters on our above examples and its performance measures to obtain the optimal policy of M / M / 1 - ∞ queueing systems with inventory under lost sales. We will rely on the tables and graphs which doing through computerized programs.

4.4.1 Total Cost Algorithm for all Performance Measures

Given: v, λ, μ, Q , satisfying all constraints on previous definitions;

Given: p , binomial, geometry;

Calculate performance measures;

$$TC = \lambda_R \cdot k + \bar{I} \cdot h + \bar{LS} \cdot \ell + \bar{L} \cdot v_w + \rho \cdot v_1;$$

$$AP = TC - (\lambda_A \cdot R);$$

End

We can compute exact value for all performance measures by programs as on Vb.net that came with our thesis.

Note: you can find the program through attached CD in our thesis.

Figure 4.1: Interface for VB.NET program

4.4.1.1 Example:

Sales company based on the sale of one type of games for children. If the arrival rate of customers to fund officers 30 customers per hour, the service rate of 35 customers per hour, and the lead time rate 0.1 unit per day . Depending on the following data :

Order size (Q) = 500 units per day, fixed cost (K) = 50 \$ per order,

Holding cost (H) = 0.02 \$ per unit per day,

Shortest cost (S) = 2 \$ per day, waiting cost in queue (VW) = 5 \$ per day,

P (probability that lost sales case of shortage) = 0.4,

Waiting cost in system (VS) = 8 \$ per day, Revenue to the system (R) = 2000 \$ per day. Determine the optimal policy for ordering the type of games.

Based on the data of the problem and through vb.net, we can obtain the solution for our distributions that we studied as :

Deterministic

The screenshot shows a VB.NET application window titled "Order Size Distributions". The interface is divided into several sections:

- Distribution:** Radio buttons for "Deterministic" (selected), "Uniform", "Binomial", and "Geometric".
- Round to:** A dropdown menu set to "8".
- Inputs:**
 - Order Size (Q) = 500
 - Arrival Rate = 30
 - Service Rate = 35 (Unit: In/ Hour)
 - Lead Time Rate (V) = 0.24 (Unit: Hour)
 - Input P (Success) = [Empty]
- Input Cost in Dollar (\$):**
 - Fixed Cost (K) = 50
 - Holding Cost (H) = 0.02
 - Shortest Cost (S) = 2
 - Waiting Cost in Queue (Vw) = 5
 - Waiting Cost in system (Vs) = 8
 - Revenue to the System (R) = 2000
- Outputs:**
 - Traffic Intensity (Utilization Factor) P^* = 0.85714286
 - The Mean Number of Customer's in System (L) = 6
 - The Mean Order Size (M) = 500
 - Reorder Rate per Time Unit (Lr) = 0.048
 - The B - Service Level = 0.8
 - The Stationary Average on - hand Inventory Position (I) = 200.4
 - The Customer's Mean Sojourn Time (W0) = 0.25
 - The Customer's Mean Waiting Time (W) = 0.21428571
 - The Average Number of Lost Sales per Unit of Time (Ls) = 6
 - The Expexted Number of Lost Sales per Cycle (Lsc) = 125
 - The Steady State on-hand Inventory Distribution of (Y) K 0 = 0.2
- Output Cost:**
 - Total Cost = 50.97942857
 - Over All Profit = 96 (Unit: Hour)

Buttons for "Calculate" and "Exit" are located at the bottom of the window.

Figure 4.2: Deterministic in VB.NET

Uniform

Order Size Distributions

Distribution

Deterministic
 Uniform
 Binomial
 Geometric

Inputs

Order Size (Q) =
 Service Rate =
 Arrival Rate =
 Lead Time Rate (V) =

Input Cost in Dollar (\$)

Fixed Cost (K) =
 Holding Cost (H) =
 Shortest Cost (S) =
 Waiting Cost in Queue (Vw) =
 Waiting Cost in system (Vs) =
 Revenue to the System (R) =

Output Cost

Total Cost =
 Over All Profit =

Outputs

Traffic Intensity (Utilization Factor) P* =
 The Customer's Mean Sojourn Time (W0) =
 The Mean Number of Customer's in System (L) =
 The Customer's Mean Waiting Time (W) =
 The Mean Order Size (M) =
 The Average Number of Lost Sales per Unit of Time (Ls) =
 Reorder Rate per Time Unit (Lr) =
 The Expexted Number of Lost Sales per Cycle (Lsc) =
 The B - Service Level =
 The Steady State on-hand Inventory Distribution of (Y) =
 The Stationary Average on - hand Inventory Position (I) =

Binomial

Order Size Distributions

Distribution

Deterministic
 Uniform
 Binomial
 Geometric

Inputs

Order Size (Q) =
 Service Rate =
 Arrival Rate =
 Lead Time Rate (V) =

Input Cost in Dollar (\$)

Fixed Cost (K) =
 Holding Cost (H) =
 Shortest Cost (S) =
 Waiting Cost in Queue (Vw) =
 Waiting Cost in system (Vs) =
 Revenue to the System (R) =

Output Cost

Total Cost =
 Over All Profit =

Outputs

Traffic Intensity (Utilization Factor) P* =
 The Customer's Mean Sojourn Time (W0) =
 The Mean Number of Customer's in System (L) =
 The Customer's Mean Waiting Time (W) =
 The Mean Order Size (M) =
 The Average Number of Lost Sales per Unit of Time (Ls) =
 Reorder Rate per Time Unit (Lr) =
 The Expexted Number of Lost Sales per Cycle (Lsc) =
 The B - Service Level =
 The Steady State on-hand Inventory Distribution of (Y) =
 The Stationary Average on - hand Inventory Position (I) =

Figure 4.4: Binomial in VB.NET

Geometric

The screenshot shows the 'Order Size Distributions' application window. The 'Distribution' section has 'Geometric' selected. The 'Inputs' section contains the following values: Order Size (Q) = 500, Arrival Rate = 30, Service Rate = 35, Lead Time Rate (V) = 0.24, Input P (Success) = 0.4, and a unit of 'Hour'. The 'Input Cost in Dollar (\$)' section includes: Fixed Cost (K) = 50, Holding Cost (H) = 0.02, Shortest Cost (S) = 2, Waiting Cost in Queue (Vw) = 5, Waiting Cost in system (Vs) = 8, and Revenue to the System (R) = 2000. The 'Outputs' section displays various performance metrics: Traffic Intensity (Utilization Factor) P* = 0.85714286, The Mean Number of Customer's in System (L) = 6, The Mean Order Size (M) = 142.89969699, Reorder Rate per Time Unit (Lr) = 0.11198221, The B - Service Level = 0.53340746, The Stationary Average on - hand Inventory Position (I) = 0, The Customer's Mean Sojourn Time (W0) = 0.37494789, The Customer's Mean Waiting Time (W) = 0.3213839, The Average Number of Lost Sales per Unit of Time (Ls) = 13.99777619, The Expedted Number of Lost Sales per Cycle (Lsc) = 125, The Steady State on-hand Inventory Distribution of (Y) K 0 = 0.46659254. The 'Output Cost' section shows a Total Cost of 66.16609143 and an Over All Profit of 223.96441905. The unit is set to 'Hour'. 'Calculate' and 'Exit' buttons are visible at the bottom.

Figure 4.5: Geometric in VB.NET

From all above figures, we obtained the total cost and overall profit which gives the optimal policy for the company, and we conclude that the best suitable distribution is the geometric distribution order size in this problem.

4.4.2 Impact Q, P Parameters on TC of Uniform, Binomial and Geometric

We investigate six examples that are grouped into two groups by different case for p success lost sales. Each set is composed of 3 subsets by different inventory sizes $Q = 1:50; 1:100; 1:1000$. For all examples let $\mu = 1$. Then we take an examples for fixed Q with different p as $0 < p \leq 1$.

Group(i): if $0 < p \leq 0.5$,

- Example 4.1: let $Q = 1 : 50$, $\lambda = 0.2, \nu = 0.1$, show figure 4.6 and table 4.1, the result of this figure obtained by mat lab program shown in appendix A.

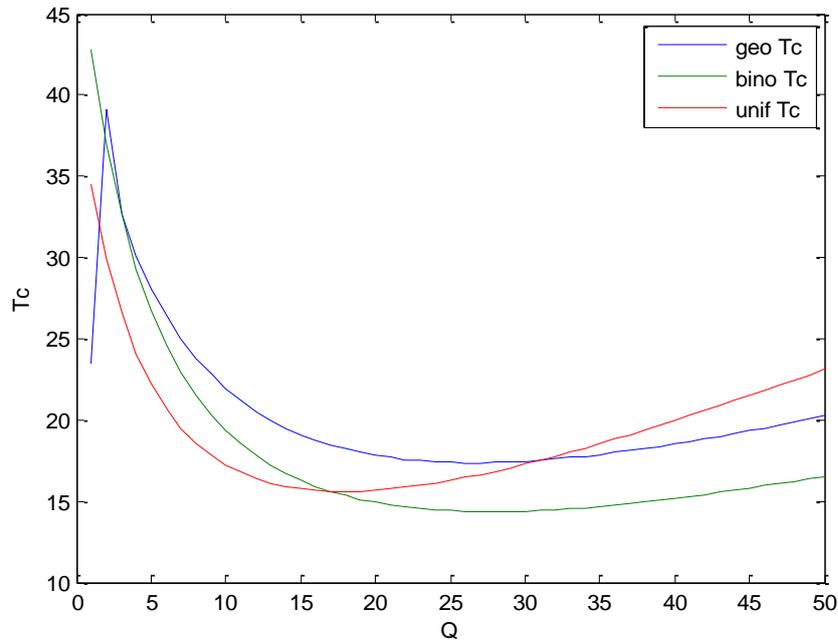


Figure 4.6: Total cost respect to the Q parameter as $Q = 1:50, 0 < p \leq 0.5$

Table 4.1: Total cost respect to the Q parameter as $Q = 1:50, 0 < p \leq 0.5$

Distribution \ Result	geometric	binomial	uniform
Total cost(Q)[min]	17.3646	14.3561	15.5908
Overall profit[max]	859.3417	892.5602	904.1639

- Example 4.2: let $Q = 1 : 100$, $\lambda = 0.2, \nu = 0.1$.

Show figure 4.7 and table 4.2

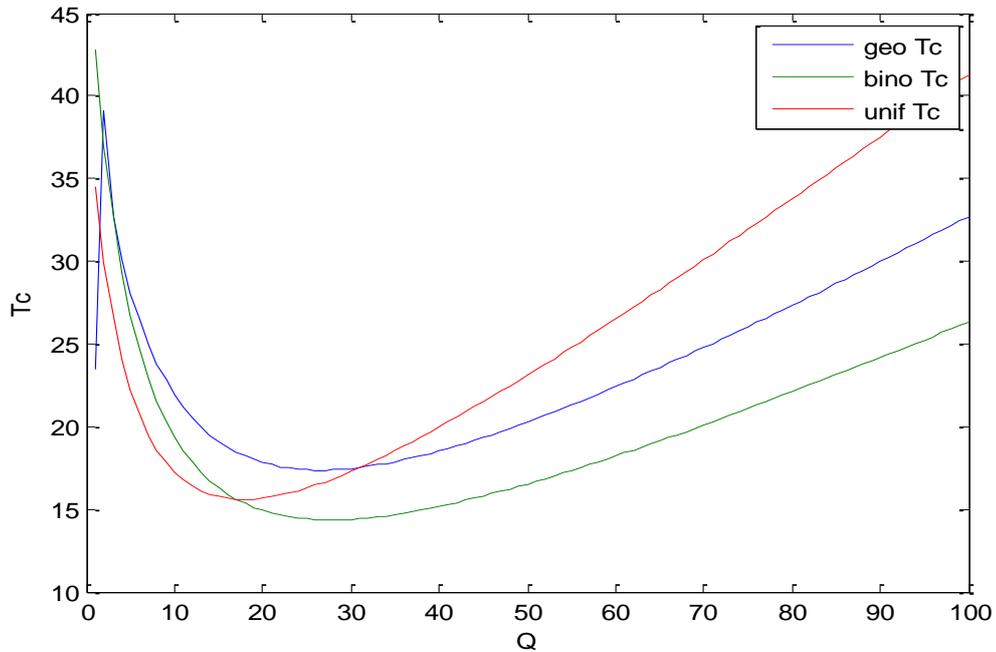


Figure 4.7: Total cost respect to the Q parameter as $Q = 1:100$, $0 < p \leq 0.5$

Table 4.2: Total cost respect to the Q parameter as $Q = 1:100$, $0 < p \leq 0.5$

Distribution \ Result	geometric	binomial	uniform
Total cost(Q)[min]	17.3646	14.3561	15.5908
Overall profit[max]	902.5113	926.0685	920.6254

- Example 4.3: let $Q = 1 : 1000$, $\lambda = 0.2$, $\nu = 0.1$.

Show figure 4.8 and table 4.3

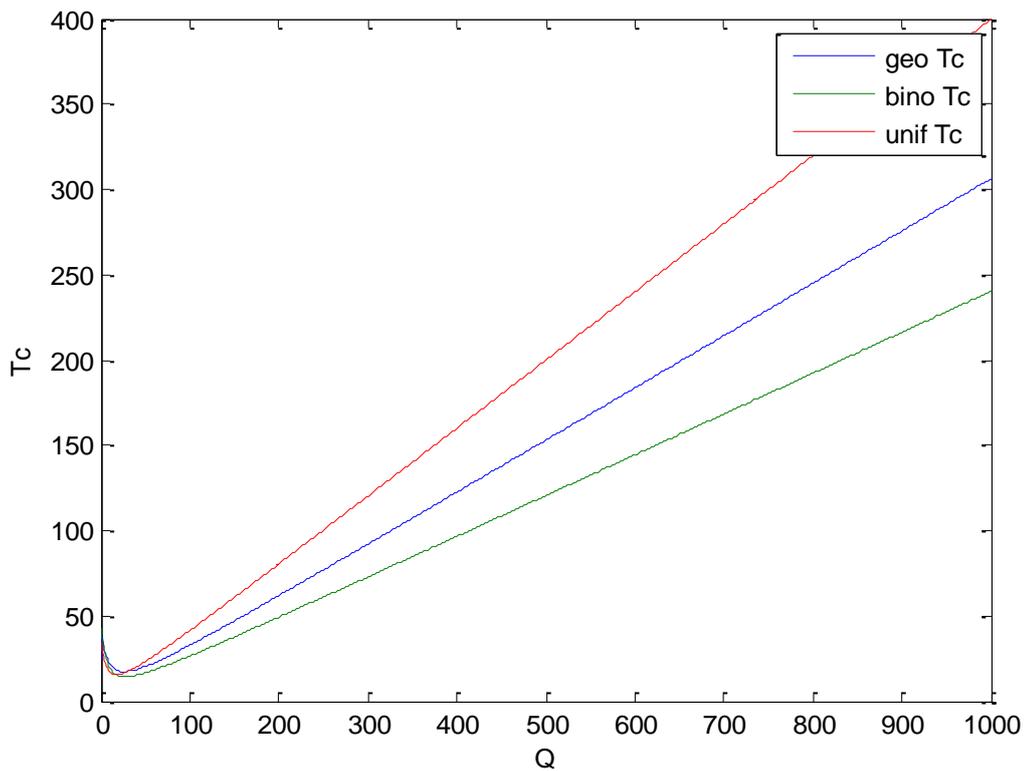


Figure 4.8: Total cost respect to the Q parameter as $Q = 1:1000$, $0 < p \leq 0.5$

Table 4.3: Total cost respect to the Q parameter as $Q = 1:1000$, $0 < p \leq 0.5$

Distribution	geometric	binomial	uniform
Result			
Total cost(Q)[min]	17.3646	14.3561	15.5908
Overall profit[max]	908.7211	930.2926	920.6254

Results group(i) : if $0 < p \leq 0.5$, the best optimal policy is to choice binomial distribution for replenishment order size Q . Because the curve of binomial TC is MIN. for all increasing in Q .

This result is stay true until we change λ, ν parameters, to confirm it show below figures (4.9, 4.10, 4.11), figure 4.9 when increasing λ , fix ν ,

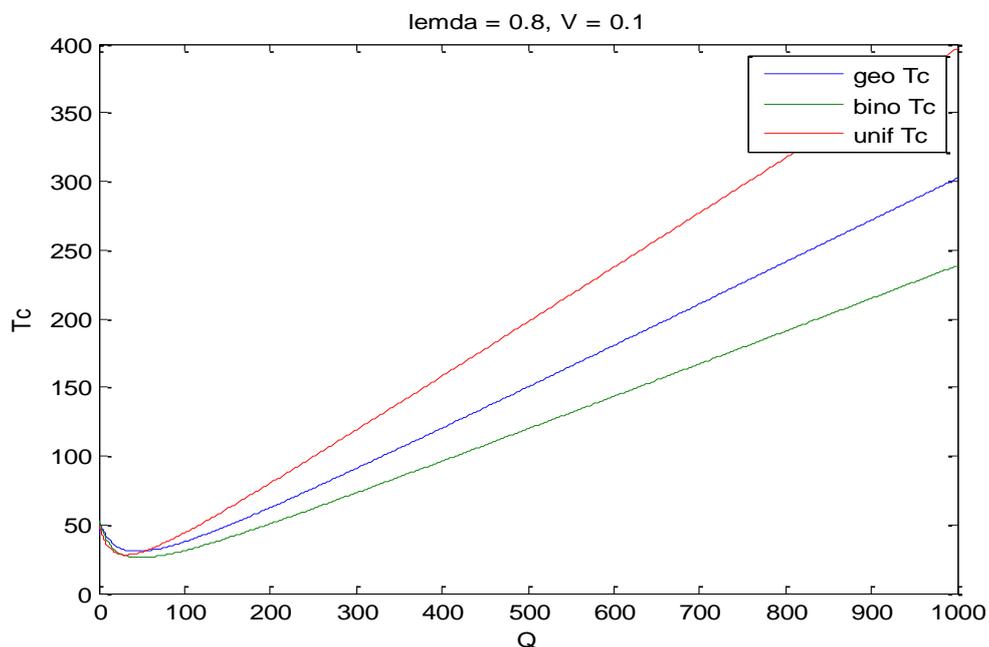


Figure 4.9: Total cost respect to the Q parameter as $Q = 1:1000$, $0 < p \leq 0.5$, inc lemnda

Figure 4.10 when increasing λ , increasing ν ,

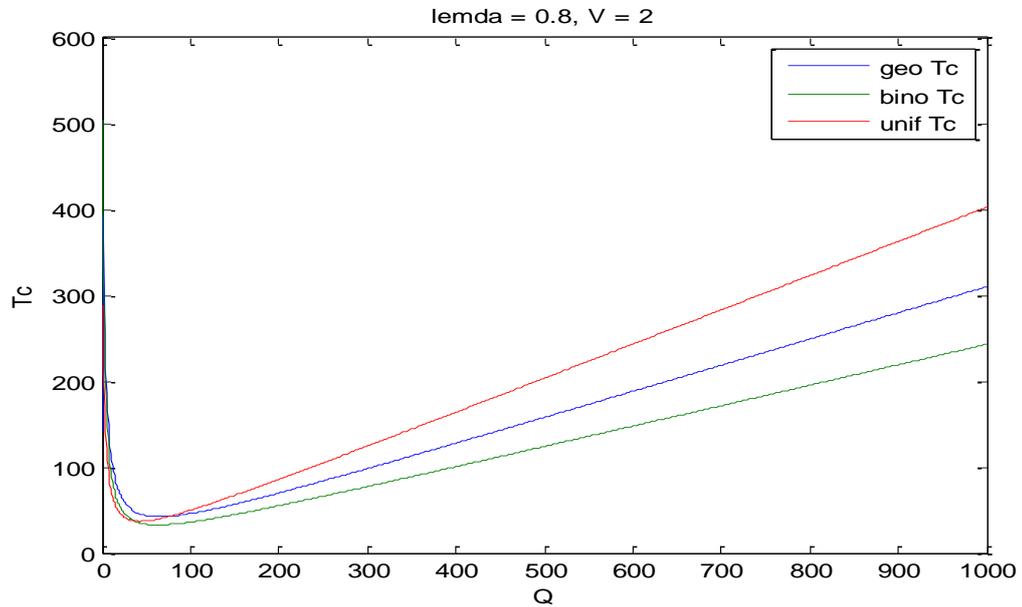


Figure 4.10: Total cost respect to the Q parameter as $Q = 1:1000$, $0 < p \leq 0.5$, inc V

Figure 4.11 when decreasing λ , fix ν ,

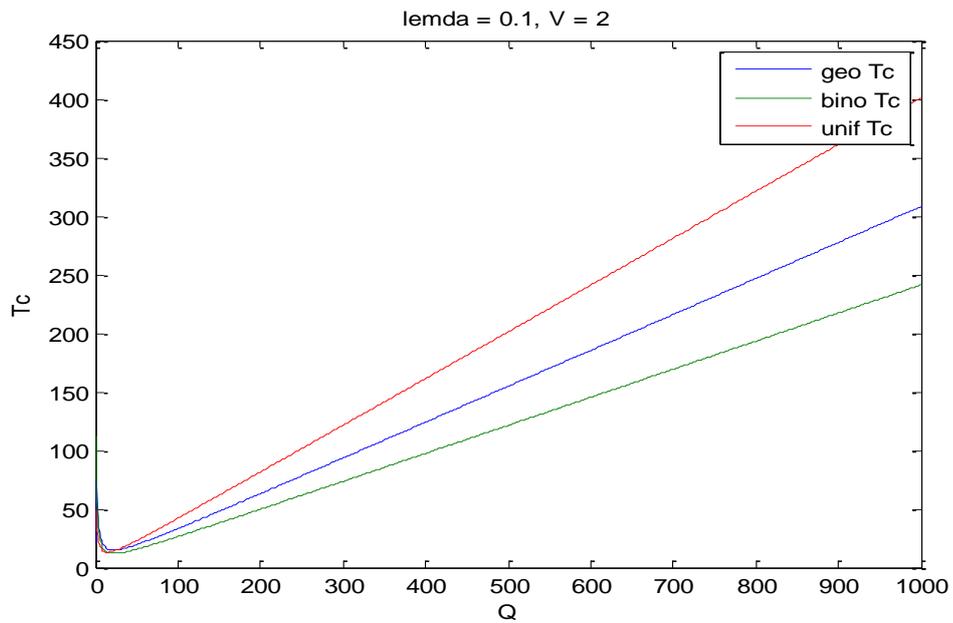


Figure 4.11: Total cost respect to the Q parameter as $Q = 1:1000$, $0 < p \leq 0.5$, dec

lemda,V

If we change the parameters λ, ν , the results of our above case remain the same and not change.

Group(ii): if $0.5 < p \leq 1$,

- Example 4.4: let $Q = 1 : 50$, $\lambda = 0.2, \nu = 0.1$, show figure 4.12 and table 4.4

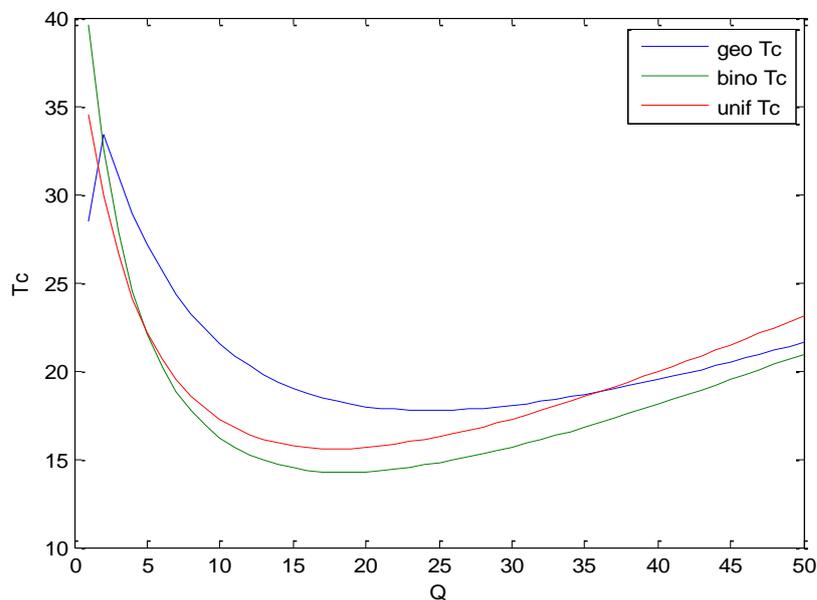


Figure 4.12: Total cost respect to the Q parameter as $Q = 1:50$, $p > 0.5$

Table 4.4: Total cost respect to the Q parameter as $Q = 1:50$, $p > 0.5$

Distribution	geometric	binomial	uniform
Result			
Total cost(Q)[min]	17.7528	14.2575	15.5908
Overall profit[max]	862.8597	916.5375	904.1639

- Example 4.5: let $Q = 1 : 100$, $\lambda = 0.2, \nu = 0.1$.

Show figure 4.13 and table 4.5

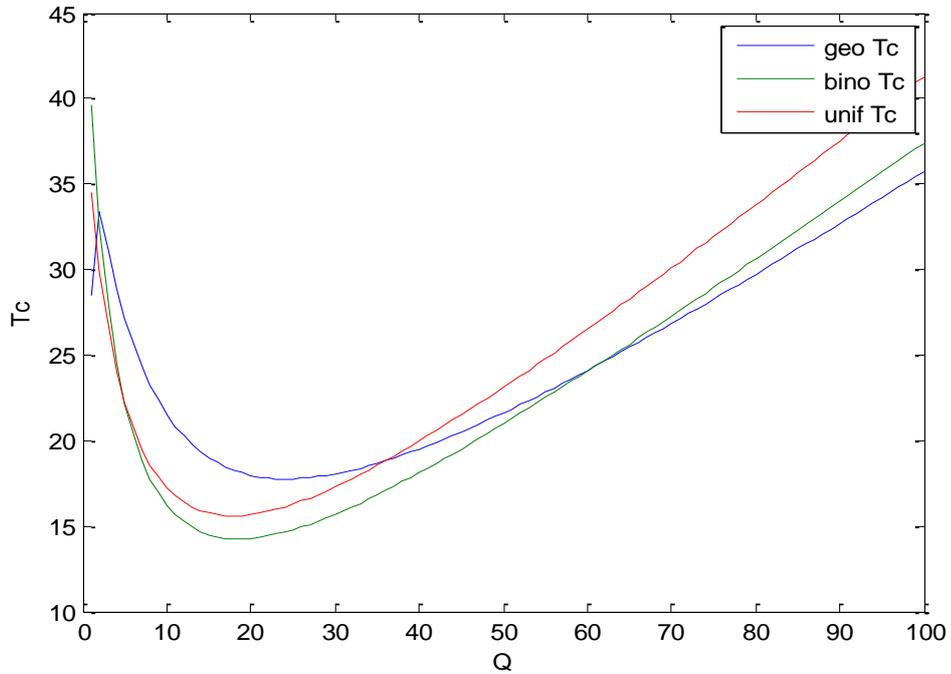


Figure 4.13: Total cost respect to the Q parameter as $Q = 1:100$, $p > 0.5$

Table 4.5: Total cost respect to the Q parameter as $Q = 1:100$, $p > 0.5$

Distribution \ Result	geometric	binomial	uniform
Total cost(Q)[min]	17.7528	14.2575	15.5908
Overall profit[max]	902.1315	930.4082	920.6254

- Example 4.6: let $Q = 1 : 1000$, $\lambda = 0.2, \nu = 0.1$.

Show figure 4.14 and table 4.6

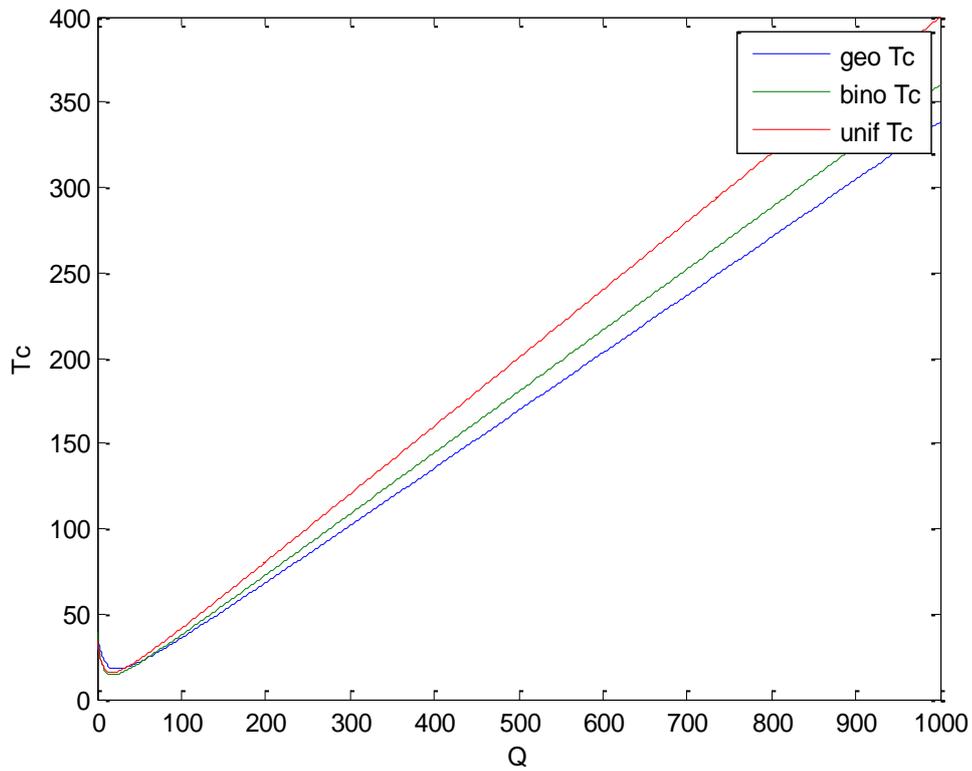


Figure 4.14: Total cost respect to the Q parameter as $Q = 1:1000$, $p > 0.5$

Table 4.6: Total cost respect to the Q parameter as $Q = 1:1000$, $p > 0.5$

Distribution	geometric	binomial	uniform
Result			
Total cost(Q)[min]	17.7528	14.2575	15.5908
Overall profit[max]	906.2471	930.4082	920.6254

Results group(ii) : for $0.5 < p \leq 1$, we have two cases:

- if Q is a small number the best optimal policy is to choice binomial distribution for small replenishment order size Q . since the curve of binomial TC is MIN. for small Q .
- But we must choice geometric distribution for large replenishment order size Q as we shown in figures 4.5, 4.6.

This result is stay true until we change λ, ν parameters. As we doing above.

If we fixed Q and change p as $0 < p \leq 1$. We have two examples.

Example 4.7: if $0 < p \leq 1$, $Q = 50$, fixed λ, ν , also $\mu = 1$, show figure 4.15, the result of this figure obtained by mat lab program shown in appendix B.

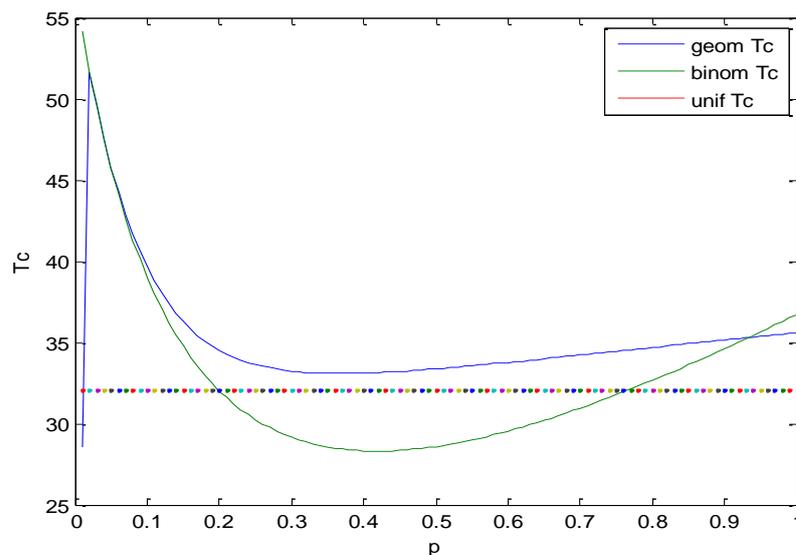


Figure 4.15: Total cost respect to the P parameter as $Q = 50$ fixed, $0 < p \leq 1$

We show that uniform TC is fixed, binomial TC is the MIN cost that is the optimal policy for small Q order sizes.

Example 4.8: if $0 < p \leq 1$, $Q = 100$, fixed λ, ν , also $\mu = 1$, show figure 4.16

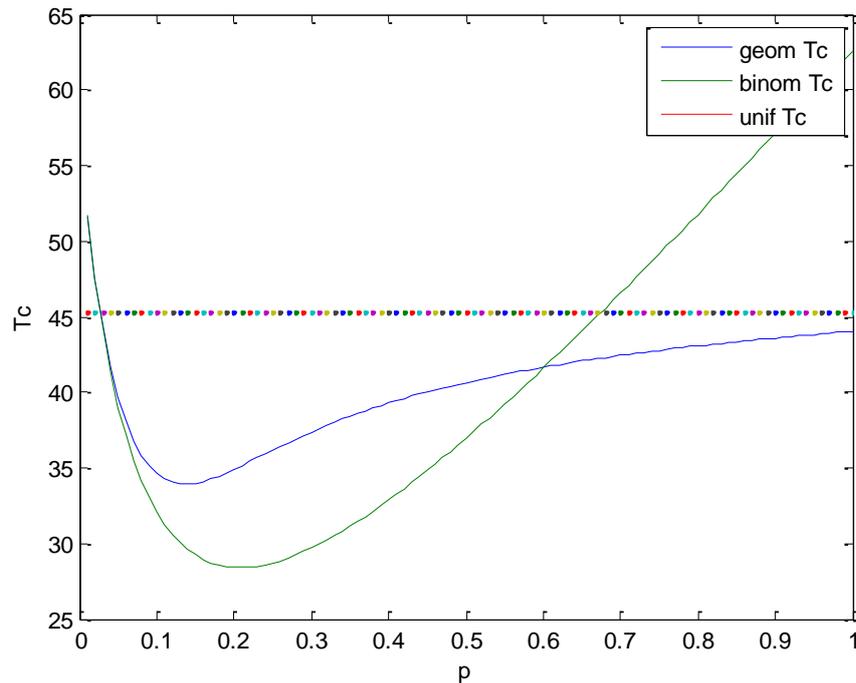


Figure 4.16: Total cost respect to the P parameter as $Q = 100$ fixed, $0 < p \leq 1$

We show that uniform TC is fixed, binomial TC is the MIN cost that is the optimal policy for small Q order sizes. Then more $Q(100:1000:\dots)$ take geometric TC is the best optimal policy.

4.4.3 Results for our Research

For all above examples we seen that the best optimal policy in decision making is to choice the binomial distribution replenishment order size generally. This gives us less cost and more profit to success our system. But if we have large Q order size we take geometric distribution.

In other words, we think that the binomial distribution replenishment order size is the best if we have models with slow items, unless we take the geometric distribution replenishment order size.

4.5 Single Server System with Inventory and Backordering

In the case of exponential inter-arrival and service times and zero lead times a reorder point $r > 0$ is suboptimal if customer demand is backordered and inventory holding costs are involved [3]. An optimal order policy for that system only places an order when the inventory level drops to zero and the number of customers in the system exceeds some threshold value [28].

For the case of backordering customers which arrive during a stock out and prescribed fixed order size with non-zero exponential lead times, the optimal policy is of threshold type such that with given inventory level the reorder decision depends on the queue length [2]. The method to prove this is stochastic dynamic optimization [36](no steady state analysis seems to be possible up to now) [28].

Chapter Five

Conclusions

In this thesis, we have studied $M / M / 1-\infty$ queueing systems with different inventory control models under lost sales. Customers are of stochastically identical behavior. To the server there is a Poisson- λ -arrival stream, $\lambda > 0$. Customers request an amount of service time which is exponentially distributed with mean 1. Service is provided with intensity $\mu > 0$. The replenishment lead time is exponentially distributed with parameter $\nu > 0$. All are independent family of random variables.

We discussed various definitions for our system found in the thesis, Then we investigate four examples for the replenishment order size distribution. We consider the fixed order size Q , which yields an $(0, Q)$ -policy and the system with uniform, binomial and geometry distribution order sizes on $\{1, \dots, Q\}$. In all cases holds $M = Q$.

We find the performance measures for these examples, and we put the cost structure, Then we do a comparison between these distributions in order to determine the best policy in the system studied. Through the study of the effect of changing parameters on the each distribution, and then selecting the lowest cost and highest profit can be obtained for the given distributions, we used computerized mathematical programs to clarification our thesis by graphs and tables. The result of our aim obtained, if we have models with slow items in order, the binomial distribution replenishment

order size is the best suitable distribution selection, unless we take the geometric distribution replenishment order size.

REFERENCES

- [1] Adan, I., Resing, J., **Queueing Theory F. 2002**, pp7-27.
- [2] Berman, O., and Kim, E., **Dynamic order replenishment policy in internet-based supply chains.**, Mathematical Models of Operations Research 53 (2001) 371–390.
- [3] Berman, O., and Kim, E., **Stochastic models for inventory management at service facilities.**, Stochastic Models 15(4) (1999) 695–718.
- [4] Berman, O., and Sapna, K.P., **Inventory management at service facilities for systems with arbitrarily distributed service times.**, Stochastic Models 16 (3,4) (2000) 343–360.
- [5] Berman, O., and Sapna, K.P., **Optimal control of service for facilities holding inventory.**, Computers and Operations Research 28 (2001) 429–441.
- [6] Berman, O., and Sapna, K.P., **Optimal service rates of a service facility with perishable inventory items.**, Naval Research Logistics, 49 (2002) 464–482.
- [7] Buchanan, D.J., Love, R.F., **A (Q,R) Inventory model with lost sales and Erlang-distributed lead times**, Naval Research Logistics Quarterly 32 (1985) 605–611.

- [8] Buzacott, J.A., Shanthikumar, J.G., **Stochastic models of manufacturing Systems.**, Prentice Hall, Englewood Clifs, 1993.
- [9] Cheng, F.M., Sethi, S.P., **Optimality of state-dependent (s,S) Policies in inventory models with markov-modulated demand and lost sales**, Production and Operations Management 8 (1999) 183– 192.
- [10] Cohen, J.W. , **The single server queue**, North-Holland, Amsterdam, 1982.
- [11] Duda, R.O., Hart, P.E., Stork, D.J., **Pattern Classification.**, (2nd ed), September 3, 1997.
- [12] <http://businessmanagementcourses.org/Lesson21QueuingTheory.pdf>.
- [13] <http://businessmanagementcourses.org/Lesson22QueuingTheory.pdf>.
- [14] http://en.wikipedia.org/wiki/Inventory_control_system.
- [15] http://en.wikipedia.org/wiki/Quality_of_service.
- [16] [http://en.wikipedia.org/wiki/Uniform_distribution_\(discrete\)](http://en.wikipedia.org/wiki/Uniform_distribution_(discrete)).
- [17] <http://fleece.ucsd.edu/~tjavidi/ECE272A/topic5.pdf>.
- [18] <http://inventorysystem.org/>.
- [19] Klein, P., **Time-homogeneous Markov chains with finite state space in discrete time.**, University of western Ontario.

- [20] Lieberman, G., Hiller, F., **Introduction to Operation Research., Fourth Edition.**
- [21] Lipsky, L., **Queueing Theory: A linear Algebraic Approach., 1992.**
- [22] Little, J.D., **A proof of the queueing formula $L = \lambda W$.**, Opns. Res., 9(1961), pp.383-387.
- [23] Mohebbi, E., Posner, M.J.M., **A continuous-review inventory system with lost sales and variable lead time.**, Naval Research Logistics, 45 (1998) 259–278.
- [24] Nick, T., **International Applied Business Research.**, Annual Conference Proceedings, Puerto Rico, March 2004.
- [25] Noble, J., **An Introduction to Markov Chains and Queueing Theory.**, 2010, p 67-79.
- [26] Ross, S.M., **Introduction to probability models, 6th ed.**, Academic Press, London, 1997.
- [27] Schneider, H., **Effect of service-levels on order-points or orderlevels in inventory models.**, International Journal of Production Research 19 (1981) 615–631.
- [28] Schwarze, M., **M/M/1 Queueing Systems With Inventory.**, 2006 springer 54-78.

- [29] Schwarze, M., Daduna, H., **Queueing systems with inventory management with random lead times and with backordering.**, 2006 springer 383-414.
- [30] Shamblin, J., Stevens, G., **Operations Research: A Fundamental Approach., 1974.**
- [31] Silver, E.A., Peterson, R., **Decision Systems for Inventory Management and Production Planning.**, John Wiley and Sons, Inc., Chichester—New York—Brisbane—Toronto—Singapore, 1985.
- [32] Stidham, S., **A last word on $L = \lambda W$.**, Opns. Res., 22 (1974), pp. 417-421.
- [33] Taha, H., **Operations Research: An Introduction., Eighth Edition 2007.**
- [34] Tempelmeier, H., **Inventory service-levels in the customer supply chain.**, Operations Research Spektrum 22 (2000) 361–380.
- [35] Tersine, R.J., **Principles of Inventory and Materials Management.** 4th Ed., PTR Prentice Hall, Englewood Cliffs, N.J., 1994.
- [36] Varaiya, P., J-P wets, **Stochastic dynamic optimization.**, Approaches and computations, 1989.

- [37] White, J.A., Schmidt, J.W., Bennett, G.K., **Analysis of Queueing Systems.**, 1975, pp 18-50.
- [38] Willing, A., **Ashort Introduction to Queueing Theory.**, J.1999, pp3-18.
- [39] Wolf, R.W., **Poisson arrivals see time averages**, Opns. Res., 30 (1982), pp. 223- 231.
- [40] Wolf, R.W. , **Stochastic modeling and the theory of queues**, Prentice-Hall, London,1989.
- [41] Zukerman, M., **Introduction to Queueing Theory and Stochastic Teletraffic Models.**, 2000-2010.
- [42] Lecture Notes , **Dr.Mohammad N. Asad** ,AL-njah university ,2007-2010 .

Appendixes

Appendix A

Total cost overall profit (different Q, change P)

Appendix B

Total cost overall profit (fixed Q, different P)

جامعة النجاح الوطنية

كلية الدراسات العليا

نظام السيطرة على المخزون باستخدام نظام الصف (M/M/1)

"حالة عدم كفاية المخزون"

إعداد

عماد رمزي محمد جمعه

إشراف

د. محمد نجيب أسعد

قدمت هذه الأطروحة استكمالاً لمتطلبات درجة الماجستير في الرياضيات المحوسبة
بكلية الدراسات العليا في جامعة النجاح الوطنية في نابلس، فلسطين.

2011

ب

نظام السيطرة على المخزون باستخدام نظام الصف (M/M/1)

"حالة عدم كفاية المخزون"

إعداد

عماد رمزي محمد جمعة

إشراف

د. محمد نجيب أسعد

الملخص

تتناول هذه الأطروحة دراسة نظام السيطرة على المخزون باستخدام نظام الصف (M/M/1) وذلك ضمن حالة فقدان المخزون وعدم كفايته، وفق الشروط العشوائية والبواسونية المعتمدة للنظام. حيث يتم تدبير كميات المواد المناسبة وفقا للمواصفات المعينة في الوقت المناسب والمكان المناسب بأقل تكلفة ممكنة، وربط ذلك بنظام صفوف الانتظار. من خلال هذا الربط ينتج نظاما متكاملًا متقدما يستخدم فيه معادلات رياضية وطرق إحصائية محوسبة وأدوات متعددة.

في هذه الأطروحة ناقشنا تعريفات مختلفة لنظام السيطرة على المخزون المذكور، ثم درسنا بعض الأمثلة على توزيعات تجديد حجم الطلب، ثم قمنا بحساب مقاييس الأداء لكل توزيع منفصل تم التحدث عنه، ووضعنا اقتران التكلفة، لتحقيق أقل تكلفة.

أجرينا في الأطروحة - من خلال برامج محوسبة على الماتلاب وبرنامج VB.net - مقارنة بين تلك التوزيعات لمعرفة التوزيع المناسب الذي يعطي أفضل سياسة مثلى للنظام المذكور والذي يحقق الهدف (أقل تكلفة وأعلى ربح).

من أهم نتائج هذه الدراسة :

1 إيجاد مقاييس الأداء لنظام السيطرة على المخزون باستخدام نظام الصف (M/M/1) في

حالة عدم كفاية المخزون. وتوسيع ذلك لتوزيعات منفصلة مرتبطة بالنظام المذكور.

2 إيجاد اقتران التكلفة المرتبط بالنظام المذكور، واقتران الربح.

ت

3 -التوصل إلى أن التوزيع الحدي هو الأفضل لتحقيق الهدف والسياسة المثلى في حالة " أن حركة الكميات والطلب بطيئة"، بينما ان كانت غير ذلك فيكون التوزيع الهندسي هو الأفضل بين التوزيعات المختارة في دراستنا.