THE UNIVERSITY OF CALGARY

MODELING AND PERFORMANCE ANALYSIS OF QUEUEING SYSTEMS

by

Yun Wang

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE
DEGREE OF MASTER OF SCIENCE

DEPT. OF ELECTRICAL & COMPUTER ENGINEERING

CALGARY, ALBERTA
NOVEMBER, 1992

National Library of Canada

Acquisitions and Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale du Canada

Direction des acquisitions et des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

ISBN 0-315-83277-0

Canada

THE UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled, "MODELING AND PERFORMANCE ANALYSIS OF QUEUEING SYSTEMS", submitted by YUN WANG in partial fulfillment of the requirements for the degree of Master of Science.

Wah Chun Chan

Dr. W. C. Chan, Supervisor and Chairman
Dept. of Electrical and Computer Engineering

B Sesay

Dr. A. Sesay
Dept. of Electrical and Computer Engineering

David Irvine-Halliday

Dr. D. Irvine-Halliday
Dept. of Electrical and Computer Engineering

Dr. X. Mao
Dept. of Mechanical Engineering

Date: 1993-02-23

ABSTRACT

Studies in the modeling of input traffic and the performance analysis of queueing systems with bursty input based on the principle of maximum entropy and of queueing theory are presented. The method of entropy maximization is applied to study both the single server and multiserver queueing systems. Then, two types of bursty input traffic are investigated. For a bulk data input, two equivalent arrival processes are obtained. For a doubly stochastic Poisson input, an approximation by a two-state Markov modulated Poisson process and the associated interarrival time distribution are determined. Finally the performance analysis of queueing systems with these two bursty inputs is investigated. Results for the mean delay, the mean queue length, the waiting time distribution and the state probability distribution are derived. Comparisons of theoretical results with simulation results show good accuracy of the modeling of the input traffic and the approaches employed in the performance analysis of queueing systems.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my supervisor Dr. W-C. Chan for his invaluable guidance and suggestions as well as continual support and encouragement throughout the course of this work, without which the work presented here could never have been completed.

My appreciation also goes to the staff of the Department of Electrical & Computer Engineering for their help during the whole period of my study and research in the Department. The financial assistance received from the Department is gratefully acknowledged.

Lastly, I would like to thank the many graduate students in the Department for their friendly assistance and thoughtful discussions with me during this research.

To My Parents.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

G/G/1	single server queue with general interarrival time distribution and general service time distribution
M/G/1	single server queue with exponential interarrival time distribution and general service time distribution
G/M/1	single server queue with general interarrival time distribution and exponential service time distribution
M/M/1	single server queue with exponential interarrival time distribution and exponential service time distribution
M/D/1	single server queue with exponential interarrival time distribution and constant service time
IDI	index of dispersion for intervals
IDC	index of dispersion for counts
G^X	general batch input
M^X	batch-Poisson input
DSPP	doubly stochastic Poisson process
MMPP	Markov-modulated Poisson process
$M^X/G/1$	single server queue with batch-Poisson input and general service time distribution
$M^X/M/1$	single server queue with batch-Poisson input and exponential service

	time distribution
$M^X/D/1$	single server queue with batch-Poisson input and constant service time
$G^X/M/1$	single server queue with general batch input and exponential service time distribution
MMPP/M/1	single server queue with Markov-modulated Poisson input and exponential service time distribution
N	number of customer in system
L	message length
T	packet interarrival time
S	service time
D	packet delay in a queueing system
W	packet waiting time in a queueing system
s	number of servers
λ_c	mean message arrival rate
t_c	mean message arrival time
μ	mean service rate
τ	mean service time
ρ	traffic intensity
a	offered load
$E(X)$	mean of random variable X
Var(X)	variance of X
c_X	coefficient of variation of X

CHAPTER 1

INTRODUCTION

1.1 STATEMENT OF THE PROBLEM

With the rapid advances of telecommunication networks and the increasing demands for communication services, modeling and performance analysis of the telecommunication networks have become more and more important problems in the related areas.

The primary function of telecommunication networks is to provide a communication path between user devices connected to the networks. Contention for resources in a telecommunication network can be modeled as a network of queues, each consisting of service stations with random input traffic.

Performance analysis of communication networks is concerned with the nature and characteristics of traffic flow in the networks. Important quantities of analysis are the number of messages or packets at each service station, the queue length, the message or packet delay, the throughput and other parameters of interest. These quantities form a basis for assessing the functional effectiveness of the network. Thus, to carry out quantitative performance analysis, mathematical models that interrelate the important parameters of traffic flow must be employed. The mathematical framework of queueing theory provides one important type of technique that is frequently used for this purpose.

Queueing systems deal with processes in which customers arrive or are generated, wait their turn for service, are serviced, and then depart. Many of the access protocols for networks involve such a sequence, where messages correspond to the customers in the processes. Thus, approximate queueing models can be used to study the communication networks and develop quantitative measures of performance.

A communication network may be regarded as a collection of interconnected nodes and links with different kinds of facilities that provide communications. Using the queueing models, each network node may be represented by a single queue.

A queueing system is completely characterized by three essential features: the input process, the queue discipline, and the service mechanism. For the queue discipline, the most natural queue discipline is that the customers form a queue and wait for service according to the order of arrival. This is called the first-come-first-served or first-in-first-out (FIFO) queue discipline. The service mechanism is concerned with the distribution function of the length of service times. In a telecommunication network, it deals with the time to process a message over a channel or through a device and is determined by the length of the input message. The traffic in a network is typically nonuniform or stochastic in nature. At any point in the network, the arrival times of the basic unit (character, packet, message) are random variables. So the nature of the input traffic to any nodes in a network is a major factor in determining the performance of the network.

Many models of queueing systems assume that arrivals occur according to a Poisson process. Intuitively, the Poisson process may be characterized by the properties that events occur one at a time and do not depend on the past history of events. The models with Poisson input are often mathematically tractable. However, experimental results and extensive studies [1]-[31] show that the wide variety of traffic supported by the modern telecommunication networks have different traffic characteristics. Some traffic, such as data, is highly bursty, while some traffic, such as voice and video, is continuous and correlated. These measured results indicate that the conventional Poisson assumption is inaccurate or inadequate for modeling the real network input traffic. Hence, there is a need for more accurate input traffic modeling for the performance analysis of such networks.

For the queueing systems with non-Poisson input traffic, even the simplest models involving bursty and correlated traffic tend to be difficult to solve analytically. As a consequence, many attempts to resolve this issue have recently been made.

A queueing system can be described by the state of the system. The state of the queueing system is characterized by a unique probability distribution called stationary probability distribution under statistical equilibrium. From the state probability distribution of the queueing system various performance measures of interest can be obtained.

In queueing theory, the common way to obtain the stationary state probability distribution of a queueing system is to solve the differential-difference equations which describe the dynamic system state behavior by employing the properties of statistical equilibrium of Markov processes. However, except in a few simple cases, such as the queueing models represented by the birth and death process, the explicit results are difficult to represent analytically. A lot of efforts have been made in proposing bounds and approximations for the more complex cases [32]-[66]. Many of these approximations are based on the partial knowledge of the first two moments of the distributions. However, even in the presence of empirical data, the characterization of these distributions involves a degree of arbitrariness which may cause a significant variation in the performance metrics [66],[38]. To overcome this shortage, a method using entropy maximization has been studied and employed.

1.2 REVIEWS OF PREVIOUS RESEARCH

Many studies have been published regarding the investigation of traffic process in different networks and the methods of characterizing and representing them approximately.

Results in [1]-[3] show that the data traffic in a packet network is bursty. A common model for data traffic is the batch process. The investigation of batch-arrival queueing models are presented in [4]-[11]. There are other kinds of data traffic processes which are discussed in [12], [13]. An arrival process of packets from a voice source is fairly complex due to the strong correlation among arrivals. In [14]-[16], the correlated generation of voice packets within a call is modeled by an Interrupted Poisson Process. Another common approach for modeling aggregated arrivals from N voice sources is to use a two-state Markov modulated Poisson process [21], [22]. Performance analyses of packet voice communication systems are given in [14]-[22]. In [23] and [24], two input traffic models for video sources are proposed. One is the continuous-state Autoregressive process and the other is the discrete-state, continuous time Markov process. The performance analyses of packet video communications are discussed in [23] and [26]. With the development of the integrated services digital networks, the integration of voice, data, video and other traffic into a network has received considerable attention. Studies in this issue are presented in [27]-[31].

1.3 THESIS OUTLINE

In chapter 2, the concepts of maximum entropy and minimum cross entropy are introduced and the principles of maximum entropy and minimum entropy are applied to study the single server queueing system, the Erlang loss system and the Erlang delay system. State proability distributions of these queueing systems are derived. The second moment of the state is calculated for some special systems.

Chapter 3 is devoted to the characterization and modeling of input traffic. Two kinds of input traffic are examined. One is the batch data process which represents the packet arrival process in computer communication networks, and the other is the Markov modulated Poisson process which represents a doubly stochastic Poisson process for packet arrival process in a packet switching system. The packet interarrival time distributions are derived and the statistical properties of the traffic models, such as the burstiness and correlation are discused. Finally, numerical results are presented.

In Chapter 4 performances of queues with batch arrival process or Markov modulated Poisson arrival process are studied based on the principle of maximum entropy and on the G/M/1 model. Numerical results are also provided.

Chapter 5 draws the main conclusions of the work and recommends some topics for further research.

CHAPTER 2

MAXIMUM ENTROPY ANALYSIS OF QUEUEING SYSTEMS

2.1 INTRODUCTION

Entropy maximization and cross-entropy minimization are general approaches to inferring a probability distribution from constraints which incompletely or partially characterize that distribution. The principle of maximum entropy has been shown[39] to be a uniquely correct, self-consistent method of inference for estimating probability distributions given information in the form of mean value.

Entropy maximization was first proposed as a general inference procedure by Jaynes[40] although it has historical roots in physics [41]. It has been applied in a remarkable variety of fields[42]-[49], including statistical mechanics and thermodynamics, reliability estimation, traffic networks, queueing theory and computer system modeling, system simulation, system modularity, spectral analysis and general probabilistic problem solving.

Utilization of the principle of maximum entropy in systems modeling has been made by various authors[49]-[60]. The analyses of queueing problem by entropy maximization are twofold. First, we shall show that many well-known formulae of queueing theory can be derived by means of entropy maximization. As such we shall show that the maximum entropy formalism can provide a framework for the analysis of queueing systems. Second, we shall use the resulting estimates of the distributions in system modeling and performance analysis.

In this chapter we shall introduce the principles of maximum entropy and minimum cross-entropy and apply these principles to the analyses of single server queues and multiserver queues.

2.2 THE PRINCIPLES OF MAXIMUM ENTROPY AND MINIMUM CROSS ENTROPY

Consider a system that has a set X of possible states $\{x_0, x_1, ...\}$ which may be finite or countably infinite and $x_n, n = 0, 1, ...$ may be specified arbitrarily. The probability that the system is in state x_n is denoted by $p(x_n)$. Suppose all that is known about these state probabilities are (m+1) constraints of the form

$$\sum_{x_n \in X} p(x_n) = 1 \tag{2.1}$$

$$\sum_{x_n \in X} f_k(x_n) p(x_n) = F_k \quad, k = 1, 2, ..., m \tag{2.2}$$

where $\{F_k\}$ are the prescribed mean values defined on the set of functions $\{f_k(x)\}$. The system entropy function is defined as

$$H(p) = - \sum_{x_n \in X} p(x_n) \ln p(x_n) \tag{2.3}$$

The principle of maximum entropy states that of all the distributions satisfying the constraints given by (2.1) and (2.2), the minimally prejudiced distribution which should be chosen is the one that maximizes the entropy function (2.3).

The principle of minimum cross-entropy is a kind of generalization that applies in cases when there is prior knowledge about the system states in addition to the constraints. This principle states that, of all the distributions that satisfy the constraints, one should choose the one that minimizes the cross-entropy

$$H(p,q) = -\sum_{x_n \in X} p(x_n)\ln\{p(x_n)/q(x_n)\} \tag{2.4}$$

where $q(x_n)$ is an estimate factor of $p(x_n)$, called estimates of the state probability distribution. Maximization of entropy (2.3) is a special case of minimization of cross-entropy (2.4) when $q(x_n)$ is uniform for $x_n \in X$ [39].

Minimization of (2.4) subject to constraints (2.1) and (2.2) can be carried out using the method of Lagrange multipliers. We define the Lagrangian

$$Lg = H(p,q) - \beta_0(\sum_{x_n \in X} p(x_n) - 1) - \sum_{k=1}^{m} \beta_k(\sum_{x_n \in X} f_k(x_n)p(x_n) - F_k) \tag{2.5}$$

where β_k, k=0,1, ...,m are the Lagrange multipliers associated with the constraints.

Then the necessary conditions for a stationary point of Lg are

$$\frac{\partial Lg}{\partial p(x_n)} = 0 \tag{2.6}$$

and

$$\frac{\partial Lg}{\partial \beta_k} = 0 \tag{2.7}$$

Performing the differentiations in (2.6) and (2.7), we obtain

$$1 + \ln(p(x_n)/q(x_n)) + \beta_0 + \sum_{k=1}^{m} \beta_k f_k(x_n) = 0 \tag{2.8}$$

$$\sum_{x_n \in X} p(x_n) = 1 \tag{2.9}$$

and

$$\sum_{x_n \in X} f_k(x_n)p(x_n) = F_k \quad , k = 1, 2, ..., m \tag{2.10}$$

Solving (2.8) for $p(x_n)$ yields

$$p(x_n) = \frac{1}{Z_p} q(x_n) exp\{-\sum_{k=1}^{m} \beta_k f_k(x_n)\} \tag{2.11}$$

where

$$\begin{aligned} Z_p &= exp\{1+\beta_0\} \\ &= \sum_{x_n \in X} q(x_n) exp\{-\sum_{k=1}^{m} \beta_k f_k(x_k)\} \end{aligned} \tag{2.12}$$

Substituting (2.11) into (2.2), we get

$$\sum_{x_n \in X} q(x_n) exp\{-\sum_{k=1}^{m} \beta_k f_k(x_n)\} = F_k \quad , k = 1, 2, ..., m \tag{2.13}$$

From (2.12) and (2.13) we can determine Z_p and hence the Lagrange multipliers β_k. Then $p(x_n)$ are given by (2.11).

2.3 SINGLE SERVER QUEUES

The G/G/1 queue represents an infinite capacity queueing system with general independent input, general service time distribution and a single server. The M/G/1 queue represents a queueing system with Poisson arrivals and a general service time distribution. The G/M/1 queue is the dual of the M/G/1 queue and has a general arrival pattern and a single exponential server. These models are of great value in the performance analysis of complex queueing systems, such as computer and flexible manufacturing systems modelled as general queueing networks.

Analysis of a single server queue based on the principle of maximum entropy has been carried out by several authors[52]-[56]. Particularly, D.D. Kouvatsos has obtained many theoretical results for single server queueing systems. In this part, first we shall present the results for the G/G/1 queue based on entropy maximization obtained by Kouvatsos, then we shall apply those results to the M/G/1 queue and the G/M/1 queue, and compare them with the exact results obtained from queueing theory.

2.3.1 The G/G/1 Queue

Consider a stable first-come first-served (FCFS) G/G/1 queue. Suppose that the queue is in steady-state and the state of the queueing system is defined by the number of customers N (being served and waiting in the system). A system is said to be in state n if $N = x_n = n$. Let p_n be the equilibrium state probability that there are n customers in the system, i.e. $p_n = p\{N = n\}$, and λ be the mean arrival rate, in customers/second, μ the mean service rate, in customers/second, c_s^2 the squared service time coefficient of variation.

For the G/G/1 queue the constraints are:

(a) Normalization

$$\sum_{i=0}^{\infty} p_n = 1 \tag{2.14}$$

(b)Utilization

$$p_0 = 1 - \rho \tag{2.15}$$

where

$$\rho = \frac{\lambda}{\mu} \tag{2.16}$$

(c) Mean

$$\sum_{i=0}^{\infty} np_n = E(N) \tag{2.17}$$

By using the solution method described in section 2.2, we have

$$p_n = \frac{1}{Z_p} q_n x^{h(n)} y^n \quad , n = 0, 1, \ldots \tag{2.18}$$

where

$$h(n) = \begin{cases} 1 & , n = 0 \\ 0 & , n = 1, 2, \ldots \end{cases} \tag{2.19}$$

and

$$x = e^{-\beta_1} \tag{2.20}$$

$$y = e^{-\beta_2} \tag{2.21}$$

β_1 and β_2 are the Lagrange multipliers.

Since we have no prior information about the states of the system we assume uniform prior estimates, $q_n = 1$ for all n and write (2.18) as

$$p_n = \frac{1}{Z_p} x^{h(n)} y^n \tag{2.22}$$

Substituting this p_n into (2.14)-(2.17), we obtain Z_p, x and y as

$$Z_p = \frac{E(N) - \rho}{\rho^2} \tag{2.23}$$

$$x = \frac{(1-\rho)(E(N) - \rho)}{\rho^2} \tag{2.24}$$

and

$$y = \frac{E(N) - \rho}{E(N)} \tag{2.25}$$

Then p_n is given by [56]

$$p_n = \begin{cases} 1 - \rho & , n = 0 \\ \frac{\rho^2}{E(N)} \left(\frac{E(N) - \rho}{E(N)} \right)^{n-1} & , n \geq 1 \end{cases} \tag{2.26}$$

This result is the state probability distribution of a single server queue with known first moment of the system state.

2.3.2 The M/G/1 Queue

For the M/G/1 queue, we use the Pollaczek-Khinchin formula[61] for the mean number of customer E(N) in the system

$$E(N) = \rho + \rho^2 \frac{1 + c_s^2}{2(1-\rho)} \tag{2.27}$$

Substituting E(N) into (2.26) we have

$$p_n = \begin{cases} 1-\rho & ,n=0 \\ 2(1-\rho)\rho^n \frac{(1+c_s^2)^{n-1}}{(2-\rho-\rho c_s^2)^2} & ,n \geq 1 \end{cases} \tag{2.28}$$

Expression (2.28) provides an approximation of p_n for an M/G/1 system with known average arrival rate and the first two moments of the service time. From (2.28) we can calculate the variance of the random variable N

$$Var(N) = \frac{\rho^2(2-\rho+\rho c_s^2)^2}{4(1-\rho^2)} + \frac{2-\rho^2+\rho^2 c_s^2}{2(1-\rho)} \tag{2.29}$$

As an example, consider the M/M/1 queue, where the service time distribution is exponential and $c_s^2 = 1$, from (2.28) and (2.29) we have

$$p_n = (1-\rho)\rho^2 \tag{2.30}$$

and

$$Var(N) = \frac{\rho}{(1-\rho)^2} \tag{2.31}$$

which yields the exact classical result for the M/M/1 queue [61].

As another example, consider the M/D/1 queue, where the service time is constant and $c_s^2 = 0$, from (2.28) and (2.29) we have

$$p_n = \begin{cases} 1-\rho & ,n=0 \\ 2(1-\rho)\left(\frac{\rho}{2-\rho}\right)^n & ,n \geq 1 \end{cases} \tag{2.32}$$

and

$$Var(N) = \frac{\rho^2(2-\rho)}{4(1-\rho^2)} + \frac{2-\rho^2}{2(1-\rho)} \tag{2.33}$$

The classic result for Var(N) of M/D/1 queue is [62]

$$Var(N) = \frac{1}{1-\rho^2}\left(\rho - \frac{3\rho^2}{2} + \frac{5\rho^3}{6} - \frac{\rho^4}{12}\right) \tag{2.34}$$

From (2.33) and (2.34) we note that there is a difference between the maximum entropy solution and the classic solution.

2.3.3 The G/M/1 Queue

For the G/M/1 queue, the mean waiting time is equal to

$$W = \frac{\sigma}{\mu(1-\sigma)} \tag{2.35}$$

where σ is the root of the equation

$$\sigma = A^*(\mu - \mu\sigma) \tag{2.36}$$

where $A^*(\cdot)$ is the Laplace-Stieltjes transform of the interarrival time distribution function A(t). By means of Little's formula[63], the average number of customers in the system is given by

$$E(N) = \lambda(W + \frac{1}{\mu}) = \frac{\rho}{1-\sigma} \tag{2.37}$$

Substituting E(N) into (2.26), we get

$$p_n = \begin{cases} 1-\rho & , n = 0 \\ \rho(1-\sigma)\sigma^{n-1} & , n \geq 0 \end{cases} \tag{2.38}$$

which is the exact classic result [61].

2.4 MULTISERVER QUEUES

We shall use the maximum entropy method to derive the state probability distribution for the Erlang loss system and the Erlang delay system.

2.4.1 The Erlang Loss System

Suppose that in an Erlang loss system there are s servers with service rate μ and customers arrive according to a Poisson process with rate λ. If an arriving customer finds all servers busy, then the customer will be rejected.

Let $\{p_n\}$, n=0,1,...,s be the steady-state probability distribution of having n customers in the system at any moment. Assume that some information about the state probabilities is known and expressed in the following constraints:

(a) The normalization condition

$$\sum_{n=0}^{s} p_n = 1 \tag{2.39}$$

(b) The mean number of customers in the system

$$\sum_{n=0}^{s} np_n = E(N) \tag{2.40}$$

For the Erlang loss system, the condition on conservation of traffic holds

$$\sum_{n=0}^{s} np_n = a(1 - p_s) \tag{2.41}$$

or

$$\sum_{n=0}^{s} n\mu p_n = \lambda(1 - p_s) \tag{2.42}$$

where $a = \lambda/\mu$.

In order to determine the probability distribution $\{p_n\}$ by the method of entropy maximization, we formulate the problem as follows:

$$\text{Minimize} \quad H(p,q) = -\textstyle\sum_{k=0}^{s} p_n \ln\{p_n/q_n\} \tag{2.43}$$

subject to the constraints (2.39) and (2.40).

The optimization problem can be solved by the method of undetermined Lagrange's multipliers leading to the solution

$$p_n = \frac{1}{Z_p} q_n x^n \quad , n = 0, 1, ..., s \tag{2.44}$$

Thus we have

$$p_0 = \frac{q_0}{Z_p} \tag{2.45}$$

From (2.39) we have

$$Z_p = \sum_{n=0}^{s} q_n x^n \tag{2.46}$$

and (2.40)

$$\begin{aligned} E(N) &= \frac{\sum_{n=0}^{s} n q_n x^n}{\sum_{n=0}^{s} q_n x^n} && (2.47) \\ &= \frac{\sum_{n=0}^{s-1} (n+1) q_{n+1} x^{n+1}}{\sum_{n=0}^{s} q_n x^n} && (2.48) \end{aligned}$$

Using (2.41), we get

$$\begin{aligned} E(N) &= a(1-p_s) && (2.49) \\ &= a \sum_{n=0}^{s-1} p_n && (2.50) \\ &= \frac{a \sum_{n=0}^{s-1} q_n x^n}{\sum_{n=0}^{s} q_n x^n} && (2.51) \end{aligned}$$

Comparing (2.48) and (2.51), we find q_n and q_{n+1} have the following relation

$$q_{n+1} = \frac{a}{n+1}\frac{1}{x}q_n \quad ,n = 0, 1, ..., s-1 \tag{2.52}$$

or

$$q_n = \frac{a^n}{n!}\frac{1}{x^n}q_0 \quad ,n = 0, 1, , ..., s \tag{2.53}$$

Substituting (2.53) into (2.44) we have

$$p_n = \frac{a^n}{n!}p_0 \quad ,n = 0, 1, ..., s \tag{2.54}$$

and

$$p_0 = \frac{1}{\sum_{n=0}^{s} \frac{a^n}{n!}} \tag{2.55}$$

Expressions (2.54) and (2.55) are the exact solution known as the Erlang distribution [61].

2.4.2 The Erlang Delay System

In an Erlang delay system, the number of states of the system is infinite. If an arriving customer finds all the servers busy, the customer will wait in the queue until service is available.

For the Erlang delay system, we assume the following constraints:

(a) The normalization condition

$$\sum_{n=0}^{\infty} p_n = 1 \tag{2.56}$$

(b) The mean number of customers in the system

$$\sum_{n=0}^{\infty} np_n = E(N) \tag{2.57}$$

Moreover, suppose p_n, n=0,1,...,s-1 are given. For the Erlang delay system

$$p_n = p_0 \frac{a^n}{n!} \quad , n = 0, 1, ..., s-1 \tag{2.58}$$

and

$$p_0 = \sum_{k=0}^{s-1} \frac{a^k}{k!} + \frac{a^s}{s!} \frac{1}{1 - a/s} \tag{2.59}$$

where $a = \lambda/\mu$.

The maximum entropy solution for p_n under the constraints (2.56)-(2.58) is

$$p_n = \begin{cases} \frac{1}{Z_p} q_n e^{-\alpha_n} x^n & , 0 \leq n \leq s-1 \\ \frac{1}{Z_p} q_n x^n . & , n \geq s \end{cases} \tag{2.60}$$

where

$$\frac{1}{Z_p} = e^{-1-\beta_1} \tag{2.61}$$

$$x = e^{-\beta_2} \tag{2.62}$$

and β_1, β_2 and $\alpha_j, j = 0, 1, ..., s-1$ are the Lagrange multipliers associated with the constraints (2.56)-(2.58), respectively.

From Takahashi [64], for a given interarrival time distribution $F_A(\cdot)$ and service time distribution $F_S(\cdot)$ with rational Laplace-Stieltjes transforms $F_A^*(s)$ and $F_S^*(s)$, respectively,

$$\frac{p_{n+1}}{p_n} = y < 1 \quad \text{if n is sufficiently large} \tag{2.63}$$

where

$$y = F_A^*(sk) \tag{2.64}$$

and k is the unique positive root satisfying the characteristic equation

$$F_A^*(sk) F_S^*(-k) = 1 \tag{2.65}$$

For $p_n, n \geq s$, of (2.60)

$$y = \frac{q_{n+1}}{q_n} x \tag{2.66}$$

From (2.64) we know that y is a constant which is independent of n. This means that $\{q_n\}$ in (2.60) should be either constant or geometric because in both cases q_{n+1}/q_n is independent of n. For both cases (2.60) can be written as

$$p_n = \begin{cases} \frac{1}{Z_p} e^{-\alpha_n} x^n & ,0 \leq n \leq s-1 \\ \frac{1}{Z_p} x^n & ,n \geq s \end{cases} \tag{2.67}$$

Substituting p_n of (2.54) into (2.56)-(2.58), we obtain [10]

$$x = \frac{E(N) - s(1-p_0) + \sum_{i=1}^{s-1}(s-1)p_i}{E(N) - (s-1)(1-p_0) + \sum_{i=1}^{s-1}(s-i-1)p_i} \tag{2.68}$$

$$Z_p = \frac{(1 - \sum_{i=0}^{s-1} p_i)^2 \Psi}{[E(N) - s(1-p_0) + \sum_{i=1}^{s-1}(s-i)p_i]^s} \tag{2.69}$$

where

$$\Psi = [E(N) - (s-1)(1-p_0) + \sum_{i=1}^{s-1}(s-i-1)p_i]^{s-2} \tag{2.70}$$

When the input process is Poisson with rate λ and exponential service time distribution with service rate μ, we have

$$F_A^*(s) = \frac{\lambda}{\lambda + s} \tag{2.71}$$

and

$$F_S^*(s) = \frac{\mu}{\mu + s} \tag{2.72}$$

where s is the Laplace-Stieltjes transform variable. Substituting (2.71) and (2.72) into (2.65) to solve for k and substituting this k into (2.64) to solve for y, we find

$$y = \frac{a}{s} \tag{2.73}$$

where s is the number of servers.

Then substituting this y and $\{p_n\}, 0 \leq n \leq s-1$, of (2.58) into (2.71) and (2.72) and using (2.67) we obtain the state probabilities [10]

$$p_n = \begin{cases} p_0 \frac{a^n}{n!} & 0 \leq n \leq s-1 \\ p_0 \frac{s^s}{s!} \left(\frac{a}{s}\right)^n & n \geq s \end{cases} \tag{2.74}$$

Note that this result is known as the state probability distribution of the Erlang delay system.

2.5 SUMMARY

In this chapter, we have applied the principle of maximum entropy to analyze queueing systems and determined the equilibrium state probability distribution for several queueing systems. We have presented the maximum entropy solution for the state probability distribution of the G/G/1 queue obtained by Kouvatsos, and used the method of minimum cross-entropy with the estimate factor of the distribution involved to derive for the first time the state probability distributions for the Erlang loss system and the Erlang delay system respectively.

From these results, it can be seen that the solution method presented is a general method for determining the state distributions when only partial prior information in the form of mean value about the system state is available. The maximum entropy solution is the approximation of the queueing system performance analysis. The accuracy of the approximation depends on the prior information provided. Generally the approximations are the least-biased choices for the given information.

CHAPTER 3

CHARACTERIZATION AND MODELING OF INPUT TRAFFIC

3.1 INTRODUCTION

A Poisson process is a good approximation for the input process of customers to a queueing system if customers arrive one at a time and if the arrival of one customer does not affect the probability of future arrivals. These conditions are frequently met, for example, by the arrival process of telephone calls to a central office, since there is a large number of potential callers each of whom calls infrequently. It is important to note that queueing models which assume Poisson arrivals can often be solved analytically. However, in some situations a Poisson process may not be sufficient as a good approximation for input processes.

Due to the development of data networks and the Integrated Services Digital Networks (ISDN), various communication services are available, such as data, voice and video, etc., each having different traffic characteristics. For example, data traffic input process has quite irregular or bursty statistics and may not be adequately modeled by a Poisson process, while traffic like voice and image is lengthy and steady and exhibits high correlation, and the aggregate packet arrival process resulting from the superposition of the streams from many voice sources is quite complicated, possessing a certain burstiness that leads to surprisingly large packet delays in the multiplexer under heavy loads. In order to evaluate the performance of such networks, it is imperative to appropriately model and characterize the input traffic and to establish the relations of the input source parameters with network parameters.

In this chapter, we shall study two kinds of input traffic and develop mathematical models for their representations. The first kind of traffic is expressed by a batch process. We shall apply the principle of maximum entropy to establish an interarrival time distribution function. The second kind of traffic concerns the bursty arrival of packets to a node in a packet-switching network. We shall represent the traffic by a Markov- Modulated Poisson process.

3.2 INDEXES OF DISPERSION

Since the first analysis of data traffic in computer communication networks in the mid- and late 1970's, which showed that packet arrival processes are highly variable, researchers have frequently described data traffic in computer communication networks as "bursty". Yet a precise definition of burstiness is not available in the literature. Most researchers seem to invoke the term bursty when confronted with processes having nonexponential interarrival time distributions. The vagueness surrounding the concept of burstiness stems both from its use to denote different types of variability in many disparate situations and from the difficulty of characterizing in meaningful ways the capricious nature of packet arrivals.

In [65] R. Gusella introduced an approach to characterize the variability of measured packet arrival processes with indexes of dispersion. Indexes of dispersion have long been known in the statistics community as a powerful tool in the analysis of the second-order properties of point processes. R. Gusella demonstrated that indexes of dispersion are valuable and valid tools for characterizing the variability of packet arrival processes.

3.2.1 The Index of Dispersion for Intervals (IDI)

Let $\{X_k,\ k \geq 1\}$ represent the sequence of interarrival times of an arrival process. We assume that $\{X_k,\ k \geq 1\}$ is stationary, by which we mean that the joint distribution of $(X_{i+1},\ X_{i+2},\ldots,X_{i+k})$ is independent of i for all k. Let $S_k = X_{i+1} + X_{i+2} + \cdots + X_{i+k}$ denote the sum of k consecutive interarrival times. The index of dispersion for intervals is defined by [65]

$$\begin{aligned} c_k^2 &= \frac{kVar(S_k)}{E^2(S_k)} = \frac{Var(S_k)}{kE^2(X)} \\ &= \frac{kVar(X) + 2\sum_{j=1}^{k-1}(k-j)cov(X_i, X_{i+1})}{kE^2(X)} \\ &= c_1^2 + (k-1)\rho_k \end{aligned} \tag{3.1}$$

where Var(X) and E(X) are the common variance and common mean of the X_k respectively, c_1^2 is the squared coefficient of variation of a single interarrival time, $\rho_k = cov(X_i, X_{i+k})/Var(X)$ is the autocorrelation coefficient.

For k=1, $c_k^2 = c_1^2$. For $k \geq 1$, c_k^2 is k times the squared coefficient of variation of S_k. For a Poisson process, $c_k^2 = c_1^2 = 1$. For a renewal process, $\rho_k = 0$, so $c_k^2 = c_1^2$. If the process is bursty, c_1^2 is usually larger than 1. For a nonrenewal process, $\rho_k \geq 0$, so $c_k^2 \geq c_1^2$, c_k^2 reveals the relationship between the variability and the correlation among successive interarrival times in the aggregate packet arrival process. It measures the cumulative covariance (normalized by the square of the mean) among k consecutive interarrival times. The notion of cumulative covariance seems to be very important for the multiplexer application, because the exceptionally large packet delays under heavy loads are due not only to high values of c_1^2 but also to the cumulative effect of many small individual covariances. Looking for fluctuations in the IDI sequence c_k^2,

$k \geq 1$ is a good way to test deviation from the renewal property. In [66], [67] and [68], the sequence c_k^2 is used as the basis for calculating the variability parameter to approximately characterize the arrival process.

3.2.2 The Index of Dispersion for Counts (IDC)

Let N(t) denote the counting process associated with an arrival process. Then N(t) is equal to the number of arrivals in an interval of length t. The index of dispersion for counts is defined as

$$I_t = \frac{Var(N(t))}{E(N(t))}) \tag{3.2}$$

For a Poisson process, $I_t = 1$. In general, I_t will not be constant for renewal processes in which counts in disjoint intervals are correlated. I_t is an alternate way to evaluate the variability of point processes from the perspective of packet arrivals. It can be proved that the limits of the IDI and IDC are equal, i.e. [65]

$$\lim_{k \to \infty} c_k^2 = \lim_{t \to \infty} I_t \tag{3.3}$$

3.3 BATCH PROCESSES

In computer communication networks, packet switching techniques are widely used, where messages are divided into smaller pieces called packets, each of which has a maximum length. Since the message length is a random quantity, a message consists of a random number of packets. In this case, we shall say the arrival process of packets forms a batch arrival process with random batch size. Batch process is one of the models that is often used to represent bursty data traffic.

3.3.1 General Batch Process

Consider a general batch process satisfying the following conditions:

1. Message arrivals follow a stationary and orderly input process with mean arrival rate λ_c or mean interarrival time $t_c = 1/\lambda_c$.

2. Each message consists of a random number L of packets with probability $p_i = P\{L = i\}$, i=1,2,..., and mean E(L).

3. Let T be a mixed random variable denoting the packet interarrival time with probability density function $f_T(t)$. The range of T is from 0^- to $+\infty$.

4. Let T_1 be a random variable denoting the interarrival time of messages.

5. Define t_p as the length of time between the arrival instant of the first bit of a given packet and the arrival instant of the first bit of the previous packet of the same message.

6. Let T_2 be a random variable denoting the total time duration of all the packets in a message.

In terms of t_p, there are two general cases:

(a) t_p is constant.

(b) t_p is variable.

Note that t_p is resulted from the processing time of a packet in a packet switching office or a node of a computer network. When the packet lengths in bits are the same, t_p is constant. When the packet lengths in bits are not the same, t_p is variable. For most practical cases t_p is constant, so we shall consider case (a) only in the following sections. It is interesting to note that the limiting case where $t_p = 0$, can be used to represent the traffic with batch arrivals.

3.3.2 Equivalent Packet Arrival Process I

We shall derive the interarrival time distribution function for a batch arrival process with constant t_p by the principle of maximum entropy.

According to the conditions of 1–6 in section 3.3.1, we have the following relations:

$$E(T_2) = t_p E(L) \tag{3.4}$$

$$E(T_1) = \frac{1}{\lambda_c} \tag{3.5}$$

and

$$E(T) = \frac{1}{\lambda_c E(L)} \tag{3.6}$$

Note that

$$P\{T < t_p\} = 0 \tag{3.7}$$

and

$$P\{T = t_p\} = 1 - \frac{\alpha}{E(L)} \tag{3.8}$$

where

$$\alpha = P\{T_1 > T_2\} \tag{3.9}$$

Since $E(L) > 1$ and $P\{T_1 > T_2\} \leq 1$, then $\alpha < E(L)$.

The probability density function $f_T(t)$ can be written as

$$f_T(t) = (1 - \frac{\alpha}{E(L)})\delta(t - t_p) + f_c(t - t_p)U(t - t_p) \tag{3.10}$$

where $f_c(t)$ denotes the continuous part of $f_T(t)$, $\delta(t)$ is the Dirac delta function, and $U(t)$ is the unit step function.

In order to find $f_c(t)$, we note that $f_T(t)$ is subject to the following constraints:

(a) Normalization condition

$$\int_0^\infty f_T(t)dt = 1 - \frac{\alpha}{E(L)} + \int_{t_p}^\infty f_c(t-t_p)dt = 1 \tag{3.11}$$

or

$$\int_0^\infty f_c(u)du = \frac{\alpha}{E(L)} \tag{3.12}$$

(b) Mean interarrival time condition

$$\int_0^\infty tf_T(t)dt = t_p(1 - \frac{\alpha}{E(L)}) + \int_{t_p}^\infty tf_c(t-t_p)dt = \frac{1}{\lambda_c E(L)} \tag{3.13}$$

or

$$\int_0^\infty (u+t_p)f_c(u)du = \frac{1-(E(L)-\alpha)t_p\lambda_c}{\lambda_c E(L)} \tag{3.14}$$

Define the entropy function

$$H = -\int_0^\infty f_c(u)\ln f_c(u)du \tag{3.15}$$

We determine $f_c(u)$ by maximizing the entropy function H subject to the constraints in (3.12) and (3.14). This is an isoperimetric problem of the calculus of variations. We can solve the optimization problem by introducing the Lagrange multipliers γ_k, k=0,1 and forming the Lagrangian

$$Lg = -f_c(u)\ln f_c(u) - \gamma_0 f_c(u) - \gamma_1(u+t_p)f_c(u) \tag{3.16}$$

Setting

$$\frac{\partial Lg}{\partial f_c(u)} = -\ln f_c(u) - 1 - \gamma_0 - \gamma_1(u+t_p) = 0 \tag{3.17}$$

leads to

$$f_c(u) = G_p e^{-\gamma_1 u} \tag{3.18}$$

where

$$G_p = e^{-1-\gamma_0-\gamma_1 t_p} \tag{3.19}$$

Using (3.12) and (3.14), we obtain G_p and γ_1 as

$$G_p = \frac{\alpha^2 \lambda_c}{E(L)(1 - E(L)t_p\lambda_c)} \tag{3.20}$$

and

$$\gamma_1 = \frac{\alpha\lambda_c}{1 - E(L)t_p\lambda_c} \tag{3.21}$$

Substituting G_p and γ_1 into (3.18), we obtain

$$f_T(t) = (1 - \frac{\alpha}{E(L)})\delta(t - t_p) + \frac{\alpha}{E(L)}\gamma_1 e^{-\gamma_1(t-t_p)}U(t - t_p) \tag{3.22}$$

Then the packet interarrival time distribution function $F_T(t)$ is given by

$$F_T(t) = (1 - \frac{\alpha}{E(L)}e^{-\gamma_1(t-t_p)})U(t - t_p) \tag{3.23}$$

It follows that the squared coefficient of variation of the packet interarrival time is

$$c_T^2 = \frac{1}{\alpha}(1 - \lambda_c t_p E(L))^2(2E(L) - \alpha) \tag{3.24}$$

It remains to determine α. There are two ways to find α. One way is to find α by measurement. Another way is by (3.9).

When we use the measurement method, we first find the probability $P\{T = t_p\}$ by measurement, then we calculate α by (3.8)

$$\alpha = E(L)(1 - P\{T = t_p\}) \tag{3.25}$$

If we use the second method, we have to assume that the probability density function of message interarrival time T_1 is given by $f_1(t)$, or the corresponding distribution function is $F_1(t)$, and the probability density function of total packet duration T_2 is $f_2(t)$. Note that T_1 and T_2 are independent random variables. Then α can be calculated as follows:

$$\begin{aligned} \alpha &= 1 - P\{T_1 \le T_2\} \\ &= 1 - \int_0^\infty \int_0^u f_1(v) f_2(u) dv du \\ &= 1 - \int_0^\infty F_1(u) f_2(u) du \end{aligned} \tag{3.26}$$

Since

$$f_2(t) = \sum_{n=0}^{\infty} P\{L = n\} \delta(t - nt_p) \tag{3.27}$$

Substituting (3.27) into (3.26), we obtain

$$\alpha = 1 - \sum_{n=0}^{\infty} P\{L = n\} F_1(nt_p) \tag{3.28}$$

For example, if message arrivals follow a Poisson process with arrival rate λ_c, and the message size has the geometric distribution

$$p\{L = n\} = p^{n-1}(1 - p) \tag{3.29}$$

By means of (3.28) and (3.29) and the relation of $p = 1/E(L)$, we have

$$\alpha = \frac{(E(L) - 1)e^{-\lambda_c t_p}}{E(L) - e^{-\lambda_c t_p}} \tag{3.30}$$

Thus we have established an equivalent packet interarrival time distribution function (3.23) or density function (3.22) for a batch arrival process. We see that $f_T(t)$ in (3.22) is a generalized exponential density function and is expressed in terms E(L), t_c and t_p.

Now we consider the limiting case where $t_p \to 0$. When $t_p \to 0$, we have from (3.28), (3.20) (3.21) and (3.24) respectively,

$$\alpha = 1 \tag{3.31}$$

$$G_p = \frac{\lambda_c}{E(L)} \tag{3.32}$$

$$\gamma_1 = \lambda_c \tag{3.33}$$

and

$$c_T^2 = 2E(L) - 1 \tag{3.34}$$

It follows that

$$f_T(t) = (1 - \frac{1}{E(L)})\delta(t) + \frac{\lambda_c}{E(L)} e^{-\lambda_c t} U(t) \tag{3.35}$$

and

$$F_T(t) = (1 - \frac{1}{E(L)} e^{-\lambda_c t}) U(t) \tag{3.36}$$

The results given in (3.31) to (3.36) are identical with those obtained by Wu [10].

3.3.3 Equivalent Packet Arrival Process II

In this section we shall establish another equivalent packet interarrival time distribution function using the same method with a different constraint, the variance of the interarrival time, Var(T). We consider the limiting case where $t_p = 0$.

Since $t_p = 0$, we have

$$P\{T = 0\} = 1 - \frac{1}{E(L)} \tag{3.37}$$

and

$$f_T(t) = (1 - \frac{1}{E(L)})\delta(t) + f_c(t)U(t) \tag{3.38}$$

For t_p=0, we have from (3.30), $\alpha = 1$. From the normalization condition (3.12) and the mean condition (3.14) we have

$$\int_0^\infty f_c(t)dt = \frac{1}{E(L)} \tag{3.39}$$

and

$$\int_0^\infty t f_c(t)dt = \frac{1}{\lambda_c E(L)} \tag{3.40}$$

Now we introduce the second moment constraint as

$$\int_0^\infty (t - \frac{1}{\lambda_c E(L)})^2 f_T(t)dt = Var(T) \tag{3.41}$$

or

$$\int_0^\infty t^2 f_c(t)dt = E(T^2) \tag{3.42}$$

where $E(T^2)$ is given by

$$E(T^2) = \frac{1}{\lambda_c^2 E(L)}(c_T^2 + 1) \tag{3.43}$$

Note that c_T^2 can be determined by measurement.

Furthermore, as $t_p = 0$, for a batch process with an arbitrary message interarrival time probability density function $f_1(t)$ and mean message length E(L), the density function $f_T(t)$ of the batch process can be written as

$$f_T(t) = (1 - \frac{1}{E(L)})\delta(t) + \frac{1}{E(L)} f_1(t)U(t) \tag{3.44}$$

From (3.44) we have

$$E(T) = \frac{E(T_1)}{E(L)} \tag{3.45}$$

$$E(T^2) = \frac{E(T_1^2)}{E(L)} \tag{3.46}$$

and

$$c_T^2 = E(L)(c_{T_1}^2 + 1) - 1 \tag{3.47}$$

Expression (3.47) shows that for $E(L) > 1$ the squared coefficient of variance of the interarrival time of a batch arrival process expressed by c_T^2 is greater than E(L) times that of the message arrival process expressed by $c_{T_1}^2$. The larger the E(L), the greater the c_T^2 of the batch arrival process.

In addition, we can obtain c_T^2 by (3.47) if we know $c_{T_1}^2$, the squared coefficient of variation of the message interarrival time. So (3.47) provides another way to determine c_T^2.

Suppose that the message interarrival time is exponential with rate λ_c. Then the density function $f_1(t)$ becomes

$$f_1(t) = \lambda_c e^{-\lambda_c t} \tag{3.48}$$

and

$$c_{T_1}^2 = 1 \tag{3.49}$$

Substituting $f_1(t)$ and $c_{T_1}^2$ into (3.44) and (3.47) respectively, we have

$$f_T(t) = (1 - \frac{1}{E(L)})\delta(t) + \frac{\lambda_c}{E(L)} e^{-\lambda_c t} U(t) \tag{3.50}$$

and

$$c_T^2 = 2E(L) - 1 \tag{3.51}$$

which are exactly the results given in (3.35) and (3.34).

Essentially the problem of finding an approximate density function $f_T(t)$ reduces to the problem of finding an approximate density function $f_1(t)$. When the message arrival process is Poisson, the maximum entropy solution for the interarrival time distribution under the mean arrival time constraint is exact. But for a non-Poisson process, the solution is only an approximation. If we want to get a more accurate solution for $f_T(t)$ or $f_1(t)$, we have to introduce more constraints as we do in this section.

In order to determine the density function $f_c(t)$ by the maximum entropy method subject to the conditions (3.39), (3.40) and (3.42), we define the entropy function

$$H = -\int_0^\infty f_c(t) \ln f_c(t) dt \tag{3.52}$$

and the Lagrangian

$$Lg = -f_c(t)\ln f_c(t) + \beta_0 f_c(t) + \beta_1 t f_c(t) + \beta_2 t^2 f_c(t) \tag{3.53}$$

By maximizing (3.53) we get

$$\begin{aligned} f_c(t) &= Z_p e^{\frac{\beta_1^2}{4\beta_2}} e^{-\beta_2(t+\frac{\beta_1}{2\beta_2})^2} \\ &= Z_p e^{\frac{\gamma_1^2}{2\gamma_2}} e^{-\frac{(t+\gamma_1)^2}{2\gamma_2}} \end{aligned} \tag{3.54}$$

where

$$\gamma_1 = \frac{\beta_1}{2\beta_2} \tag{3.55}$$

and

$$\gamma_2 = \frac{1}{2\beta_2} \tag{3.56}$$

Using the constraints(3.39), (3.40) and (3.42), we obtain the following set of equations

$$(\gamma_1 + t_c) e^{\frac{\gamma_1^2}{2\gamma_2}} \int_{\gamma_1}^{\infty} e^{-\frac{t^2}{2\gamma_2}} dt - (t_c + E(L)E(T^2)) = 0 \tag{3.57}$$

$$\gamma_2 = \gamma_1 t_c + E(L)E(T^2) \tag{3.58}$$

and

$$Z_p = \frac{\gamma_1 + t_c}{E(L)(\gamma_1 t_c + E(L)E(T^2))} \tag{3.59}$$

Resorting to numerical methods, we can solve these equations, and then find Z_p, γ_1 and γ_2. Substituting them into (3.54) we can determine the second equivalent interarrival time probability density function. The density function in (3.54) is a normal-like distribution with mean and variance determined by the mean message length E(L), the mean interarrival time $t_c/E(L)$ and the second moment of interarrival time $E(T^2)$.

3.3.4 Numerical Results And Comparision

In this section we will present some numerical results from simulations and results from theoretical analyses in section 3.3.2 and 3.3.3.

First we discuss the relation of α with E(L) and t_p and the relation of c_T with E(L) and t_p, which are presented in (3.30) and (3.24) respectively.

Fig. 3.1 shows α as a function of E(L) for given t_c and $E(T_2)$, i.e. $E(L)t_p$. The results show that for given t_c and $E(T_2)$, α almost keeps the same value when E(L) or t_p varies, and α is less for greater $E(T_2)$. These agree with the definition of α given in (3.9).

Fig. 3.2 shows the curve of the coefficient of variation of interarrival time, c_T, versus E(L) under given t_c and $E(T_2)$. Fig. 3.3 shows c_T as a function of E(L) for given t_c and t_p. Fig. 3.4 shows the curve of c_T versus t_p under given t_c and E(L), and From these figures we can see the effect of E(L) on c_T and the effect of nonzero t_p on c_T.

When t_p is equal to zero or $E(T_2) = t_pE(L) = 0$, c_T increases as E(L) increases, see the solid curves in Fig. 3.2 and Fig. 3.3. When t_p is not equal to zero and $E(T_2)$ is kept unchanged, c_T increases while E(L) increasing, see the dotted curve in Fig. 3.2. If $E(T_2)$ increases as E(L) increases for given t_p, c_T increases with E(L) increasing at first but decreases after E(L) approaches a certain value, see Fig. 3.3. From (3.24) and Fig. 3.3 we can see that c_T decreases to zero when

$$E(L) = \frac{t_c}{t_p} \tag{3.60}$$

i.e.

$$E(T) = \frac{1}{\lambda_c E(L)} = t_p \tag{3.61}$$

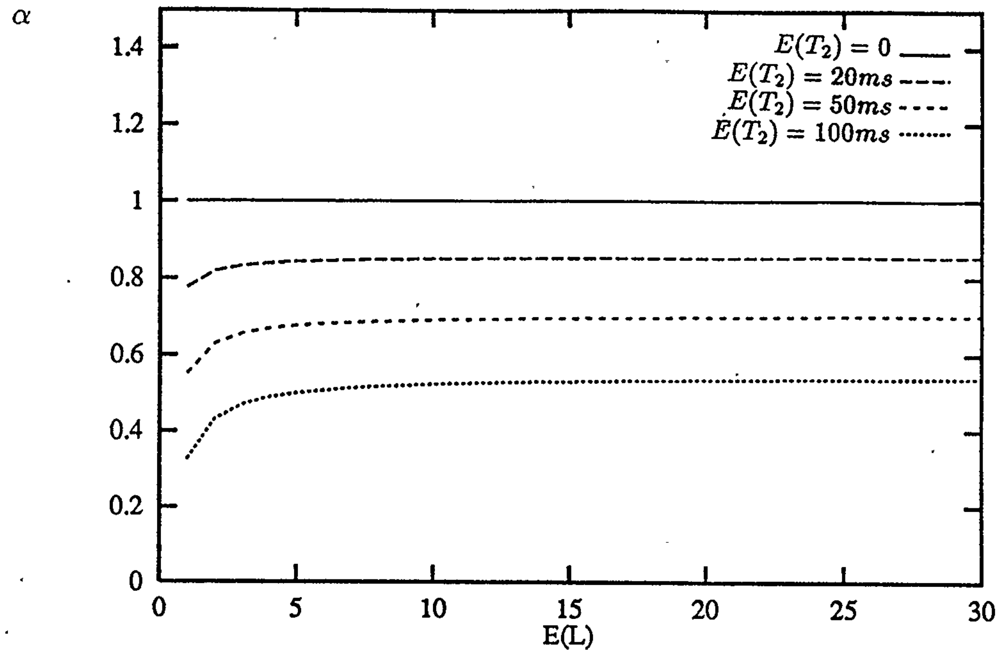

Figure 3.1. Parameter α, $t_c = 120$ms

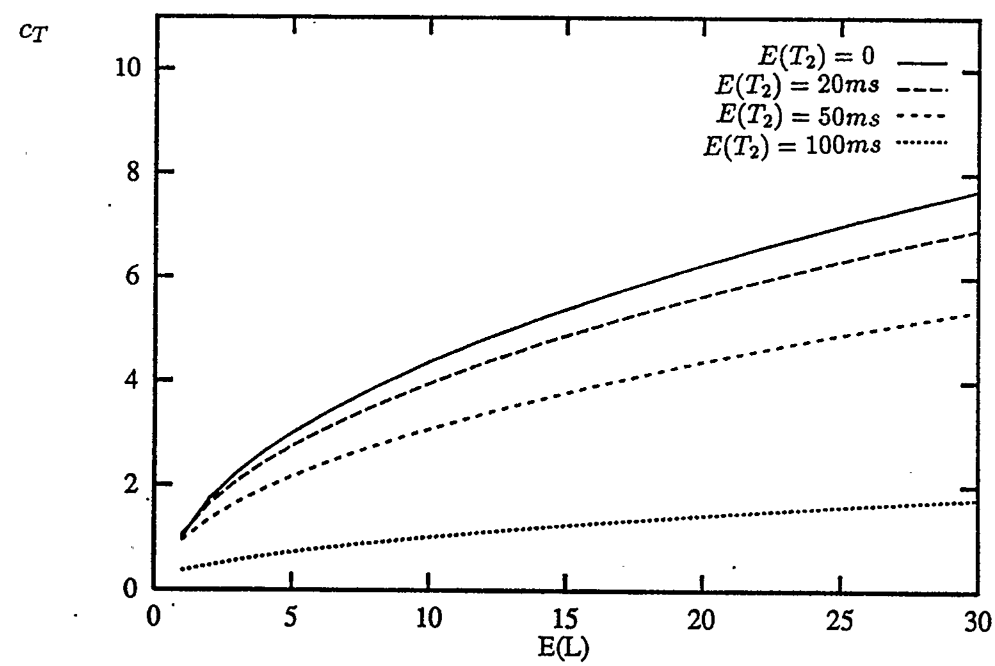

Figure 3.2. Coefficient of Variation of Packet Interarrival Time, $t_c = 120$ms

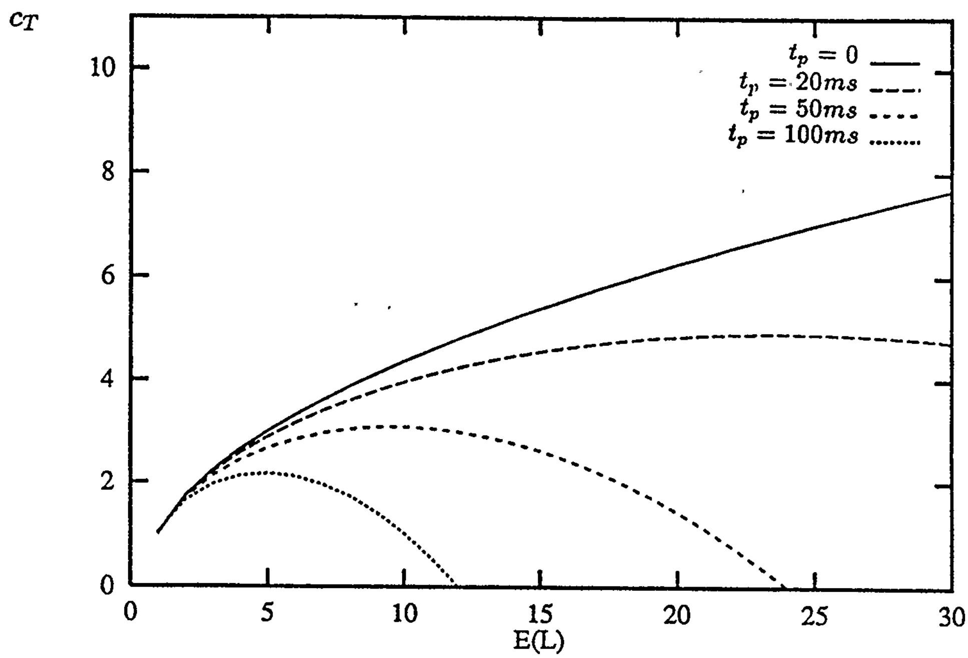

Figure 3.3. Coefficient of Variation of Packet Interarrival Time, $t_c = 120$ms

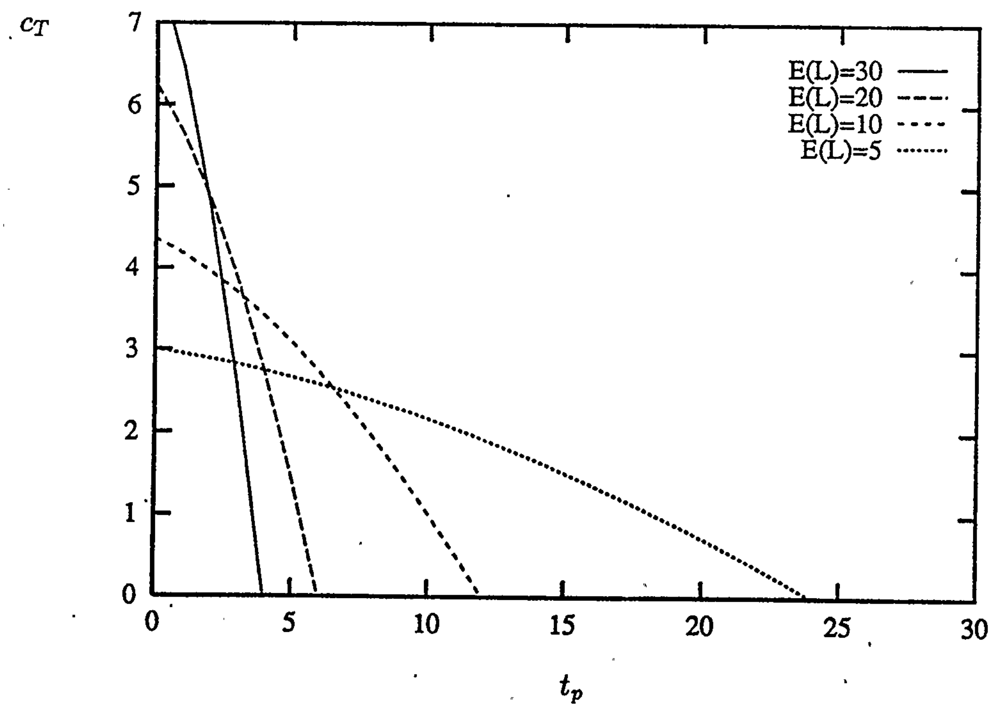

Figure 3.4. Coefficient of Variation of Packet Interarrival Time, $t_c = 120$ms

or

$$E(T_2) = t_c \tag{3.62}$$

The reason for c_T decreasing with E(L) increasing is that, after E(L) increases to a certain value, the packet interarrival time T decreases to and tends to t_p, and the number of packet interarrival times which are equal to t_p is much greater than the number of packet interarrival times which are not equal to t_p. Under this condition, the variance of T decreases, so does the coefficient of variation. At the point $E(L) = t_c/t_p$, as (3.61) establishes, the variance of T decreases to zero, so the coefficient of variation decreases to zero. The increase of t_p also decreases the variance of packet interarrival time for given E(L), so c_T decreases when t_p increases for given E(L), see Fig. 3.4.

Now we consider the interarrival time distributions for the batch arrival processes with Poisson or non-Poisson message arrivals.

Fig. 3.5 and Fig. 3.6 show the continuous part of the interarrival time density function $f_c(t)$ for the batch processes with Poisson message arrivals and with t_p equal to 5ms and 10ms, respectively. In both figures two theoretical curves obtained from (3.22) are shown. For the curve with α_c we determine the parameter α in (3.22) by (3.30). For the curve with α_m we determine α by obtaining $P\{T = t_p\}$ from simulation first and then calculating (3.25).

In table 3.1 and table 3.2 the simulation results and computation results of $P\{T = t_p\}$, α and c_T as functions of E(L) are provided. In both tables, $P\{T = t_p\}_s$ and c_{Ts} are obtained from simulation results, $P\{T = t_p\}_c$ and c_{Tc} are calculated by (3.8) and (3.24) respectively with α_c, and c_{Tm} is obtained from (3.24) with α_m. Comparing the results provided in Fig. 3.5, Fig. 3.6, table. 3.1 and table. 3.2, we see that

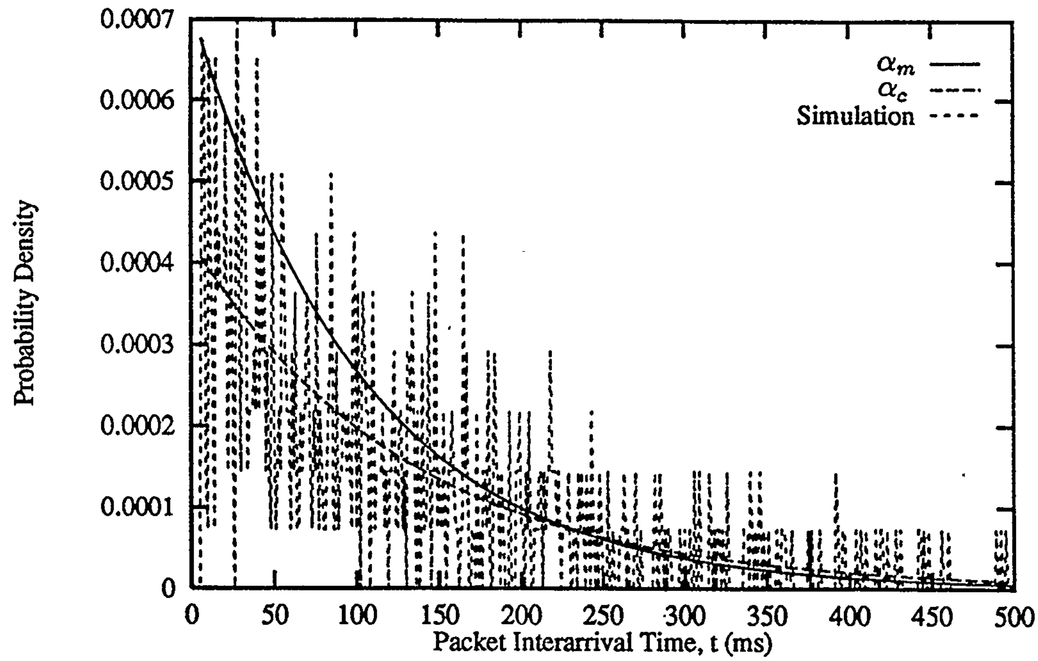

Figure 3.5. Packet Interarrival Time Probability Density of M^X Input with $t_p = 5$ms $t_c = 120ms, E(L) = 10$

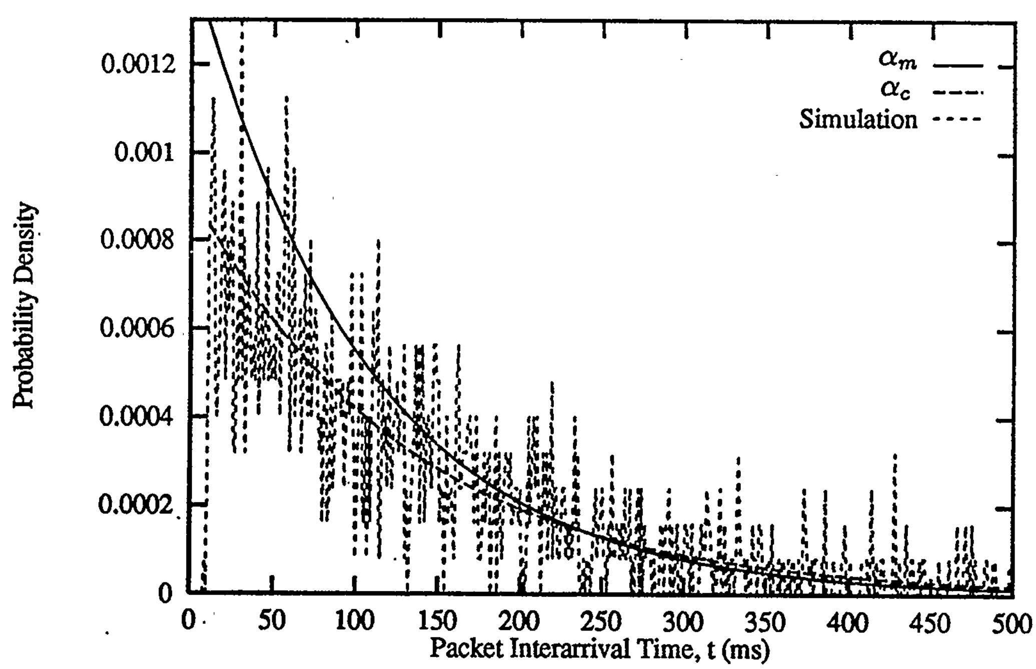

Figure 3.6. Packet Interarrival Time Probability Density of M^X Input with $t_p = 10$ms $t_c = 120ms, E(L) = 5$

Table 3.1. Results of $P\{T = t_p\}$, α and c_T with M^X Input, $t_c = 120$ms, $t_p = 2$ms

E(L)	$P\{T = t_p\}_s$	$P\{T = t_p\}_c$	α_m	α_c	c_{Ts}	c_{Tm}	c_{Tc}
5	0.8288	0.8170	0.8559	0.9151	3.040	2.996	2.888
8	0.9011	0.8906	0.7909	0.8749	3.829	3.800	3.604
10	0.9180	0.9150	0.8204	0.8499	4.134	4.029	3.956
15	0.9521	0.9471	0.7932	0.7339	4.782	4.786	4.551
16	0.9557	0.9511	0.7088	0.7828	4.941	4.872	4.630

the interarrival time density function given in (3.22) matches favorably with the simulation results. We can say that (3.22) or (3.23) is a good approximation for the interarrival time of the batch input process with Poisson message arrival and nonzero t_p. The results with α_m are more closer to the simulation results than those with α_c. It indicates that the more exact the α is, the more closer the result we could obtain from (3.22) comparing to the simulation result.

In the case where t_p is equal to zero, when the message arrival process is Poisson, the interarrival time density function given by (3.22) can exactly represent the batch arrival process. This can be seen from Fig. 3.7, Fig. 3.8 and table 3.3. In table 3.3, $P\{T = t_p\}_s$ and c_{Ts} are obtained from simulation results. $P\{T = t_p\}_c$ and c_{TI} are calculated by (3.8) and (3.24) respectively.

If $t_p = 0$ and the message arrival process is not Poisson, the representation of the interarrival time probability density of the batch process by (3.22) is not so good. Therefore (3.54) should be used.

In Fig. 3.9 and Fig. 3.10 the message arrival processes are assumed to be uniformly distributed over the interval of 0 to $2t_c$. The interarrival time density given by (3.22) and (3.54) are plotted in both figures. In table 3.4 we provide a comparison of c_T and $P\{T = 0\}$. $P\{T = 0\}$ obtained from both formaulae are the same with value

Table 3.2. Results of $P\{T = t_p\}$, α and c_T with M^X Input, $t_c = 120$ms, $t_p = 5$ms

E(L)	$P\{T = t_p\}_s$	$P\{T = t_p\}_c$	α_m	α_c	c_{Ts}	c_{Tm}	c_{Tc}
5	0.8607	0.8380	0.6963	0.8099	2.854	2.894	2.667
8	0.9198	0.9082	0.7342	0.6413	3.636	3.262	3.040
10	0.9475	0.9309	0.6910	0.5252	3.545	3.551	3.084
15	0.9699	0.9598	0.6025	0.4515	2.989	2.697	2.619
16	0.9767	0.9632	0.5874	0.3728	2.225	2.437	2.303

Table 3.3. Results of $P\{T = t_p\}$ and c_T with M^X Input, $t_c = 120$ms, $t_p = 0$

E(L)	$P\{T = t_p\}_s$	$P\{T = t_p\}_c$	c_{Ts}	c_{TI}
5	0.8247	0.80	3.250	3.0
8	0.8884	0.8750	4.045	3.872
10	0.9065	0.90	4.281	4.359
15	0.9370	0.9333	5.307	5.385
16	0.9388	0.9375	5.583	5.568

Table 3.4. Results of $P\{T = t_p\}$ and c_T with G^X Input, $t_c = 120$ms, $t_p = 0$

E(L)	$P\{T = t_p\}_s$	$P\{T = t_p\}_c$	c_{Ts}	c_{TI}	c_{TII}
5	0.8071	0.80	2.401	3.0	2.380
8	0.8926	0.8750	3.137	3.873	3.109
10	0.9201	0.90	3.602	4.359	3.511
15	0.9342	0.9333	4.419	5.385	4.358
16	0.9397	0.9375	4.535	5.568	4.509

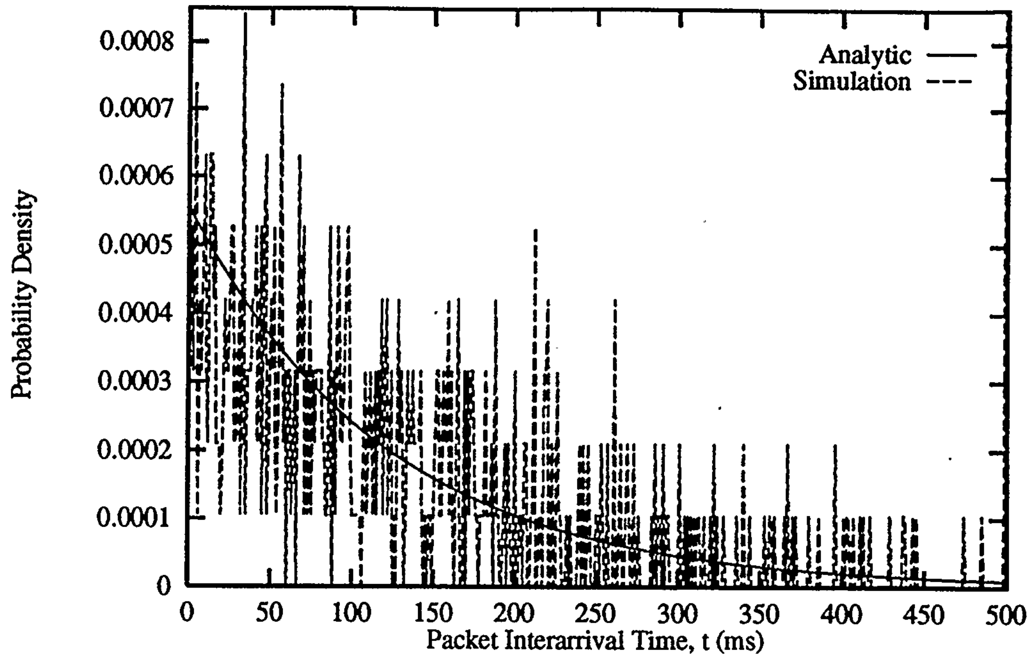

Figure 3.7. Packet Interarrival Time Probability Density of M^X Input with $t_p = 0$ $t_c = 120ms, E(L) = 10$

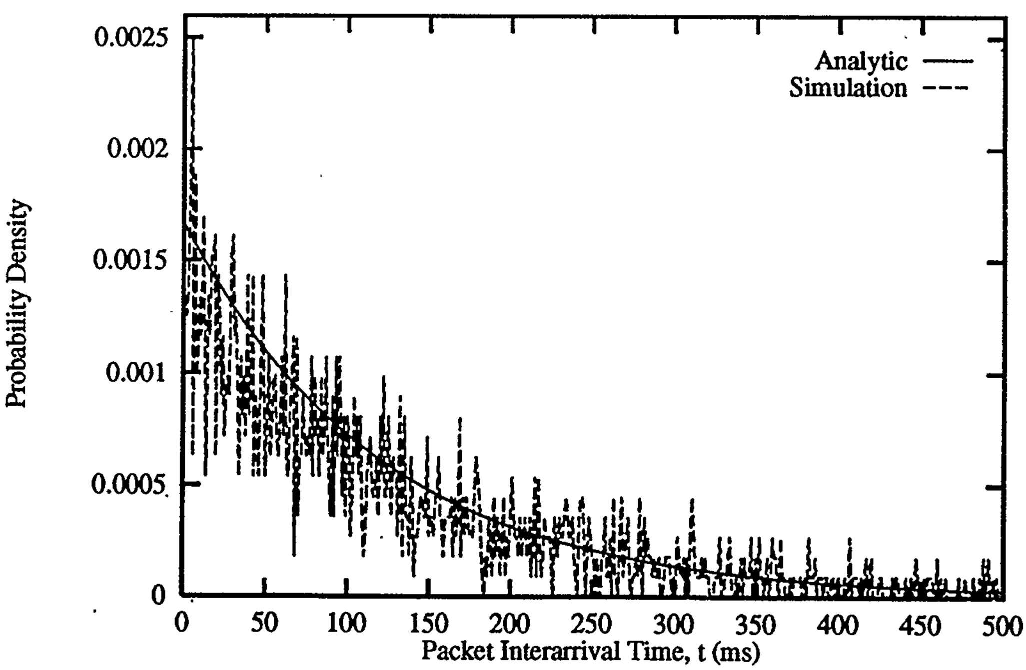

Figure 3.8. Packet Interarrival Time Probability Density of M^X Input with $t_p = 0$ $t_c = 120ms, E(L) = 5$

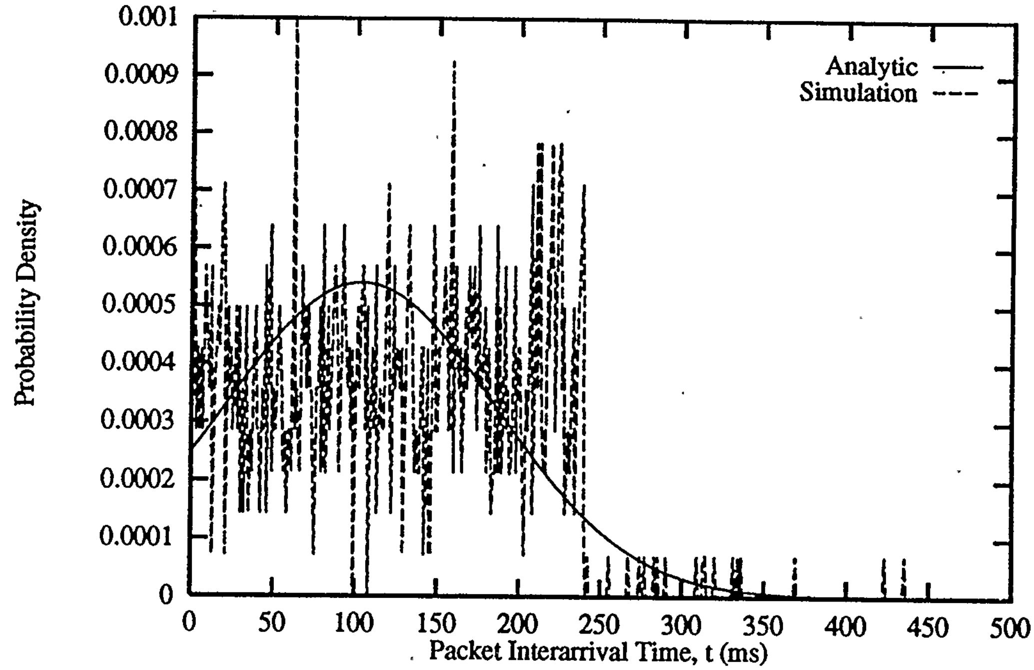

Figure 3.9. Packet Interarrival Time Probability Density of G^X Input with $t_p = 0$
$t_c = 120ms, E(L) = 10$

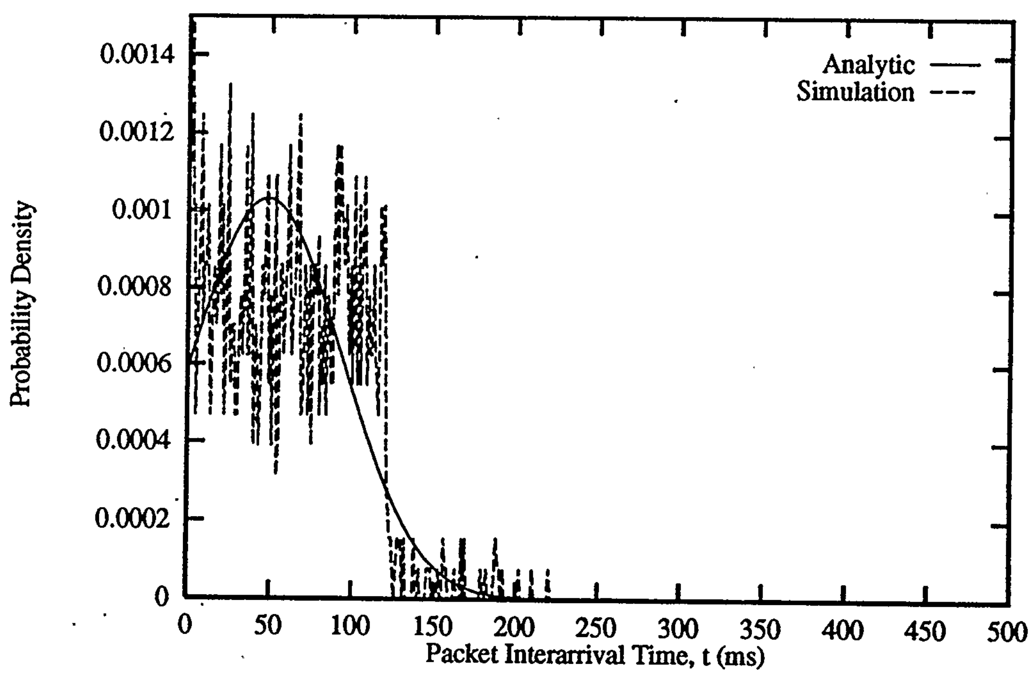

Figure 3.10. Packet Interarrival Time Probability Density of G^X Input with $t_p = 0$
$t_c = 60ms, E(L) = 10$

of $1 - 1/L$ and accurate, but c_T are quite different. In the table c_{TI} is calculated by (3.24) and c_{TII} is obtained from (3.47). Fig. 3.9 , Fig. 3.10 and table 3.4 show that the results obtained from (3.54) matches the simulation results much better than those obtained from (3.22).

From the above results we can see that when the message arrival process is Poisson with random message length, (3.22) yields very accurate results. If the message process is not Poisson, (3.54) is better than (3.22) in the sense that the variance of the interarrival time obtained from (3.54) is more close to that of the simulated batch arrival processes than (3.22).

3.4 DOUBLY STOCHASTIC POISSON PROCESS AND MARKOV-MODULATED POISSON PROCESS

In this section we shall study the arrival process of packets to a node in a packet -switching network. When calls are connected by the switching circuits, the route through which packets travel for a particular call is fixed for the duration of the call. We shall show that if an individual call in its holding time generates packets according to a Poisson process, then the instantaneous packet arrival rate at any node is equal to the sum of the rates for the calls routed through the node, which varies randomly with time. In this case, the arrivals are correlated and the traffic is said to be bursty.

3.4.1 Traffic Model

The arrival process of packets can be generated in the following way [12]. Consider an Erlang delay system with s servers and service rate μ_1, see Fig. 3.11, where call requests arrive at the system following a Poisson process with rate λ_c. When a call setup is initiated, packets are generated according to a Poisson process with a rate

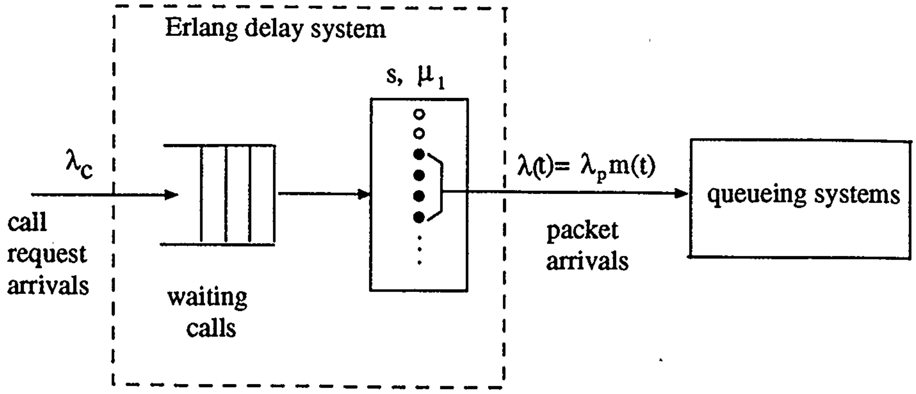

Figure 3.11. Packet Arrival Process

λ_p in the duration of the call holding time. Let m(t) denote the number of calls in progress at time t. Then the packet generating rate is

$$\lambda(t) = \lambda_p m(t) \tag{3.63}$$

where m(t) is the number of busy servers at time t. Thus m(t) is a random function fluctuating in unit steps between 0 and s. The packet stream generated then is offered to a node of the switching network.

The traffic considered here is thus a nonhomogeneous Poisson process with rate $\lambda_p m(t)$. In general the packet stream to each node is different with different traffic parameters λ_c, s, μ_1 and λ_p.

3.4.2 Doubly Stochastic Poisson Process

The packet stream described in section 3.4.1 is in fact a doubly stochastic Poisson process with rate $\lambda(t)$ which itself is a realization of a stationary, continuous-time stochastic process. Since $\lambda(t) = \lambda_p m(t)$, the statistical properties of $\lambda(t)$ are determined by m(t).

As is well known, for the Erlang delay system, m(t) has stationary probabilities

$\{p_n$, n=0,1,...,s$\}$ which is given by

$$p_n = \begin{cases} \frac{a_1^n}{n!}p_0 & ,\text{n=0,1,...,s-1} \\ \frac{a_1^s}{s!}\frac{p_0}{1-\frac{a_1}{s}} & ,\text{n=s} \end{cases} \tag{3.64}$$

where

$$p_0 = \sum_{k=0}^{s-1} \frac{a_1^k}{k!} + \frac{a_1^s}{s!}\frac{1}{1-\frac{a_1}{s}} \tag{3.65}$$

and

$$a_1 = \frac{\lambda_c}{\mu_1} \tag{3.66}$$

is the offered load of the Erlang delay system. Thus p_n is the equilibrum probability that n servers are busy. With this choice of absolute probabilities, m(t) is a strictly stationary process, whose mean, variance and third moment are, respectively,

$$m_1 = a_1 \tag{3.67}$$

$$\sigma^2 = a_1(1 - p_s) \tag{3.68}$$

and

$$m_3 = \sigma^2 + a_1(3a_1 - 2sp_s + a_1^2 - a_1 p_s) \tag{3.69}$$

The covariance function of m(t), R(t), can be expressed as [69]

$$\begin{aligned} R(t) &= E\{[m(u+t) - m_1][m(u) - m_1]\} \\ &\simeq \sigma^2 e^{-\frac{t}{\tau_c}} \end{aligned} \tag{3.70}$$

where τ_c is the time constant defined by

$$\tau_c = \frac{1}{\sigma^2}\int_0^\infty R(t)dt \tag{3.71}$$

we have [12]

$$\tau_c = \frac{1}{\mu_1(1-p_s)} \tag{3.72}$$

From (3.63) the rate process $\lambda(t)$ is also a stationary counting process with random value $\lambda_j = j\lambda_p$, j=0,1,...,s. Denote the mean, variance, the third moment and the covariance function of $\lambda(t)$ by λ_{m1},λ_{m2},λ_{m3} and r(t) respectively. They are simply related to the corresponding moments and the covariance function of m(t). In terms of the moments of m(t) given by (3.67)–(3.70) we have

$$\lambda_{m1} = E(\lambda(t)) = a_1\lambda_p \tag{3.73}$$

$$\lambda_{m2} = E[\lambda(t) - \lambda_{m1}]^2 = a_1(1-p_s)\lambda_p \tag{3.74}$$

$$\lambda_{m3} = E(\lambda^3(t)) = (\sigma^2 + 3a_1^2 - 2Sp_s + a_1^3 - a_1p_s) \tag{3.75}$$

and

$$r(t) = \lambda_p^2\sigma^2 e^{-\frac{t}{\tau_c}} \tag{3.76}$$

Now we examine the packet arrival process N(t), the number of arrival packets at time t. By the definition of the doubly stochastic Poisson process, the probability of exactly k packets arriving in t is given by [70]

$$\begin{aligned} P_k(t) &= P\{N(t) = k\} \\ &= E_{\lambda(t)}\left[\sum_{k=0}^{\infty} \frac{(\int_0^t \lambda(u)du)^k}{k!} exp\{-\int_0^t \lambda(u)du\}\right] \end{aligned} \tag{3.77}$$

Then the mean number of arrival packets over the interval (0, t) is

$$E(N(t)) = E(E(N(t) \mid \lambda(t))) = E(\lambda(t)t) = \lambda_{m1}t \tag{3.78}$$

The IDC of the packet arrival process is given by [70]

$$I_t = 1 + \frac{2}{\lambda_{m1} t} \int_0^t (t-u) r(u) du \tag{3.79}$$

Substituting r(t) in (3.76) into (3.79), we get

$$I_t = 1 + \frac{2\sigma^2 \tau_c}{a_1 \lambda_p} + \frac{2\sigma^2 \tau_c^2}{a_1 \lambda_p t}(e^{-\frac{t}{\tau_c}} - 1) \tag{3.80}$$

As $t \to \infty$, we obtain

$$I_\infty = 1 + \frac{2\sigma^2}{a_1 \lambda_p} \tau_c \tag{3.81}$$

This result shows that the index of dispersion for counts I_∞ is related to both the variance σ^2 and the time constant τ_c of m(t).

From (3.2) and (3.80) we get the second moment of the number of arrival packets over the interval (0, t)

$$E(N^2(t)) = a_1^2 \lambda_p^2 t^2 + (a_1 \lambda_p + 2\sigma^2 \tau_c) t - 2\sigma^2 \tau_c^2 (1 - e^{-\frac{t}{\tau_c}}) \tag{3.82}$$

3.4.3 Markov-Modulated Poisson Process

The packet arrival process considered in section 3.4.1 is a correlated doubly stochastic process. Since the rate process $\lambda(t)$ is a fairly complicated process which is difficult for analysis, we shall present an approximating packet arrival process which is also a doubly stochastic but analyzable when offered to a queueing system and which matches the important statistical properties of the rate process $\lambda(t)$.

A Markov-modulated Poisson process (MMPP) is a doubly stochastic Poisson process where the rate process $\lambda(t)$ is determined by the state of a continuous-time Markov chain. We define the state of the Markov chain or the state of the MMPP as follows. When the rate process $\lambda(t)$ is equal to λ_j at time t, j=0,1,...,m (m is an integer), the Markov chain or the MMPP is said to be in state j. We also call this rate process a phase process. When the rate is λ_j, j=0,1,...,m, the process is said to be in phase j.

Since the packet arrival process disscused in section 3.4.1 is a doubly stochastic Poisson process with a rate process being a phase process, the arrival rate $\lambda_j = j\lambda_p$ at time t, j=0,1,...,s, we can choose an MMPP as an approximating process. We use a two-state MMPP for which simple analytic or algorithmic queueing results are available [12]. The Markov chain is in state j (j=1,2) if the arrival proces is Poisson with rate λ_j. The transition rate of state 1 and 2 are r_1 and r_2, respectively.

Denote the equilibrium probability vector of the two-state MMPP by

$$\mathbf{P} = [p_1, p_2] \tag{3.83}$$

It is known that

$$\mathbf{P} = [\frac{r_2}{r_1 + r_2}, \frac{r_1}{r_1 + r_2}] \tag{3.84}$$

The four parameters λ_1, λ_2, r_1 and r_2 completely determine the two-state MMPP.

Now we shall derive the interarrival time distribution for the two-state MMPP in terms of λ_1, λ_2, r_1 and r_2.

Let T denote the interarrival time of packets and G(t) its complementary distribution function

$$G(t) = P\{T > t\} \tag{3.85}$$

Since

$$G(t) = P\{T > t\} = P\{N_t = 0\} \tag{3.86}$$

It follows from (3.77)

$$G(t) = E_{\lambda(t)}(exp\{-\int_0^t \lambda(u)du\}) \tag{3.87}$$

where

$$\lambda(t) = \begin{cases} \lambda_1 & \text{if the Markov chain is in state 1 at t} \\ \lambda_2 & \text{if the Markov chain is in state 2 at t} \end{cases} \tag{3.88}$$

Define

$$G_1(t) = P\{T > t \mid \lambda(0) = \lambda_1\} \tag{3.89}$$

and

$$G_2(t) = P\{T > t \mid \lambda(0) = \lambda_2\} \tag{3.90}$$

Then

$$G(t) = p_1 G_1(t) + p_2 G_2(t) \tag{3.91}$$

From the stationary property of $\lambda(t)$, it follows that the quantity

$$E(exp\{-\int_h^{t+h} \lambda(u)du \mid \lambda(h)\})$$

is independent of h. By considering the possible changes of $\lambda(t)$ during a small time interval (0,h), we can write

$$G_1(t+h) = (1 - r_1 h)e^{-\lambda_1 h} G_1(t) + r_1 h e^{-\frac{\lambda_1+\lambda_2}{2} h} G_2(t) + o(h) \tag{3.92}$$

As $h \to 0$, we have the difference-differential equation

$$G_1'(t) = -(r_1 + \lambda_1)G_1(t) + r_1 G_2(t) \tag{3.93}$$

Similarly, we have

$$G_2'(t) = r_2 G_1(t) - (r_2 + \lambda_2)G_2(t) \tag{3.94}$$

The solutions to these difference-differential equations are

$$G_1(t) = \alpha_1 e^{-\beta_1 t} + \eta_1 e^{-\beta_2 t} \tag{3.95}$$

and

$$G_2(t) = \alpha_2 e^{-\beta_1 t} + \eta_2 e^{-\beta_2 t} \tag{3.96}$$

where

$$\beta_1 = \frac{1}{2}(r_1 + r_2 + \lambda_1 + \lambda_2) - \sqrt{\frac{1}{4}(r_1 + r_2 + \lambda_1 + \lambda_2)^2 - \lambda_1\lambda_2 - \lambda_1 r_2 - \lambda_2 r_1} \tag{3.97}$$

and

$$\beta_2 = \frac{1}{2}(r_1 + r_2 + \lambda_1 + \lambda_2) + \sqrt{\frac{1}{4}(r_1 + r_2 + \lambda_1 + \lambda_2)^2 - \lambda_1\lambda_2 - \lambda_1 r_2 - \lambda_2 r_1} \tag{3.98}$$

Thus

$$G(t) = \alpha e^{-\beta_1 t} + \eta e^{-\beta_2 t} \tag{3.99}$$

The interarrival time distribution function of the two-state MMPP is given by

$$\begin{aligned} F(t) &= 1 - G(t) \\ &= 1 - \alpha e^{-\beta_1 t} - \eta e^{-\beta_2 t} \end{aligned} \tag{3.100}$$

Since G(0)=1, we have

$$\alpha + \eta = 1 \tag{3.101}$$

then

$$F(t) = 1 - \alpha e^{-\beta_1 t} - (1-\alpha) e^{-\beta_2 t} \tag{3.102}$$

and the density function is

$$f(t) = \alpha\beta_1 e^{-\beta_1 t} + \beta_2(1-\alpha) e^{-\beta_2 t} \tag{3.103}$$

By using the mean interarrival time condition, we find

$$\alpha = \frac{1}{\beta_2 - \beta_1}(\beta_1\beta_2\frac{r_1 + r_2}{\lambda_1 r_2 + \lambda_2 r_1} - \beta_1) \tag{3.104}$$

From (3.103) we see that the interarrival time distribution of a two-state MMPP is a hyperexponential distribution. The squared coefficient of variation of the interarrival time is given by

$$c_T^2 = \frac{2(\frac{\alpha}{\beta_1^2} + \frac{1-\alpha}{\beta_2^2})}{(\frac{\alpha}{\beta_1} + \frac{1-\alpha}{\beta_2})^2} - 1 \tag{3.105}$$

As we mentioned before, the two-state MMPP and its interarrival time distribution are completely determined by parameters λ_1, λ_2, r_1 and r_2. However, there may be more than one way to choose them. Here we use the method by [12] to choose these four parameters such that the statistical characteristics of the rate process, λ_{m1}, λ_{m2}, λ_{m3} and τ_c are matched with those of the two-state MMPP.

We set

$$\lambda_{m1} = \frac{\lambda_1 r_2 + \lambda_2 r_1}{r_1 + r_2} \tag{3.106}$$

$$\lambda_{m2} = \frac{r_1 r_2(\lambda_1 - \lambda_2)^2}{(r_1 + r_2)^2} \tag{3.107}$$

$$\lambda_{m3} = \frac{\lambda_1^3 r_2 + \lambda_2^3 r_1}{(r_1 + r_2)^2} \tag{3.108}$$

and

$$\lambda_{m2}\tau_c = \int_0^\infty \frac{r_1 r_2(\lambda_1 - \lambda_2)^2}{(r_1 + r_2)^2} e^{-(r_1+r_2)t} dt = \frac{r_1 r_2(\lambda_1 - \lambda_2)^2}{(r_1 + r_2)^3} \tag{3.109}$$

Solving equations (3.106)–(3.109), yields the following relations [12]:

$$r_1 = \frac{1}{\tau_c(1 + \eta)} \tag{3.110}$$

$$r_2 = \frac{\eta}{\tau_c(1+\eta)} \tag{3.111}$$

$$\lambda_1 = \lambda_{m1} + \sqrt{\frac{\lambda_{m2}}{\eta}} \tag{3.112}$$

and

$$\lambda_2 = \lambda_{m1} - \sqrt{\frac{\lambda_{m2}}{\eta}} \tag{3.113}$$

where

$$\eta = 1 + \frac{\delta}{2}(\delta - \sqrt{4+\delta^2}) \tag{3.114}$$

and

$$\delta = \frac{\lambda_{m3}}{\lambda_{m2}\sqrt{\lambda_{m2}}} \tag{3.115}$$

Now we denote the number of arrivals of the two-state MMPP over the interval (0,t) by N_t. Given the four parameters of the MMPP using the same procedure as in deriving (3.78) to (3.82), we have

$$E(N_t) = \frac{\lambda_1 r_2 + \lambda_2 r_1}{r_1 + r_2} t \tag{3.116}$$

$$\begin{aligned} E(N_t^2) &= \frac{(\lambda_1 r_2 + \lambda_2 r_1)^2}{(r_1+r_2)^2} t^2 + \left[\frac{\lambda_1 r_2 + \lambda_2 r_1}{r_1+r_2} + \frac{2(\lambda_1-\lambda_2)^2 r_1 r_2}{(r_1+r_2)^3}\right] t \\ &\quad - \frac{2(\lambda_1-\lambda_2)^2 r_1 r_2}{(r_1+r_2)^4}(1 - exp\{-(r_1+r_2)t\}) \end{aligned} \tag{3.117}$$

$$\begin{aligned} I_t &= 1 + \frac{2(\lambda_1-\lambda_2)^2 r_1 r_2}{(r_1+r_2)^2(\lambda_1 r_2 + \lambda_2 r_1)} \\ &\quad - \frac{2(\lambda_1-\lambda_2)^2 r_1 r_2}{(r_1+r_2)^3(\lambda_1 r_2 + \lambda_2 r_1)t}(1 - exp\{-(r_1+r_2)t\}) \end{aligned} \tag{3.118}$$

and

$$I_\infty = 1 + \frac{2(\lambda_1-\lambda_2)^2 r_1 r_2}{(r_1+r_2)^2(\lambda_1 r_2 + \lambda_2 r_1)} \tag{3.119}$$

We find that when we use equations (3.110)–(3.113) to choose λ_1, λ_2, r_1 and r_2, the statistics in (3.116), (3.117) and (3.118) are equal to those in (3.78), (3.82) and (3.80) respectively.

3.4.4 Simulation Results

In this section we present the simulation results and compare them with analytical results obtained in section 3.4.3.

First we set up an Erlang delay queueing system with s servers as shown in Fig. 3.11. This system generates the packets process. The input of the calls to the system is Poisson with rate λ_c. The service time distribution is exponential with mean value μ_1. During each service period, packets are generated by a Poisson process with rate λ_p.

It is interesting to note that the probability density function of packet interarrival time obtained by simulation is exponential-like. See Fig. 3.12 to Fig. 3.15 as examples. Here we use the coefficient of variation of packet interarrival time to characterize the bursty nature of the packet process. The simulation results show that the coefficient of variation of packet interarrival time is greater than 1, which indicates that the packet process is indeed a bursty one. Since the coefficient of variation of an exponential distribution is equal to 1, the interarrival time distribution of the packet processes we are investigating is not an exponential distribution.

In section 3.4.3 we used a two-state MMPP to represent the packet process and showed that the interarrival time distribution of a two-state MMPP is hyperexponential. Fig. 3.12 and Fig. 3.13 show the density functions of the packet interarrival time obtained by simulation and the numerical results obtained by the MMPP model

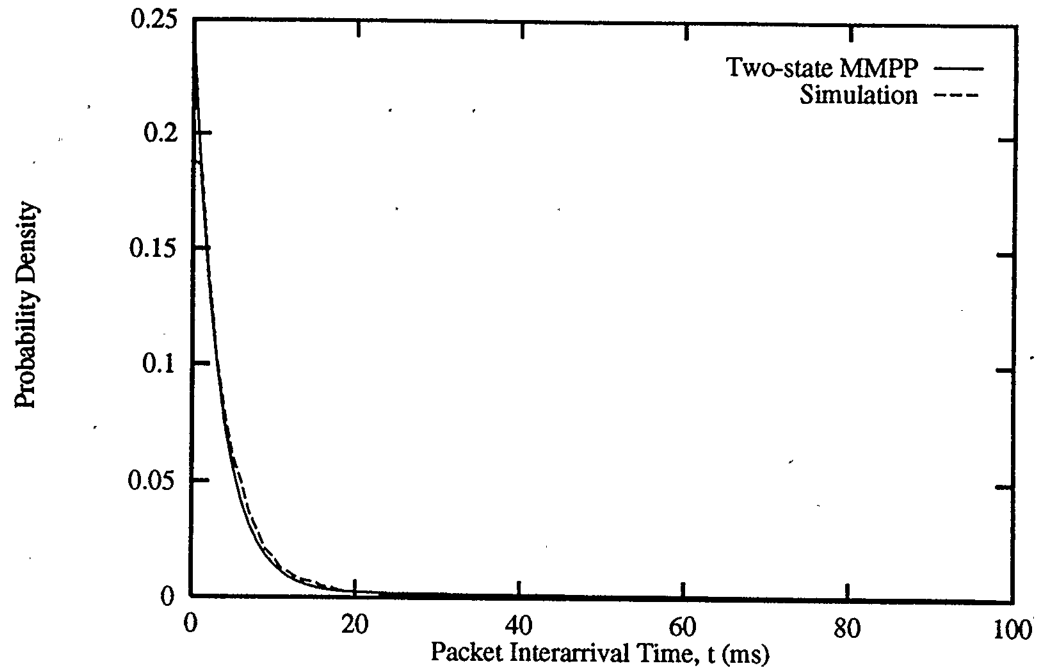

Figure 3.12. Packet Interarrival Time Probability Density, $s = 2, a_1 = 0.25, t_p = 5ms$

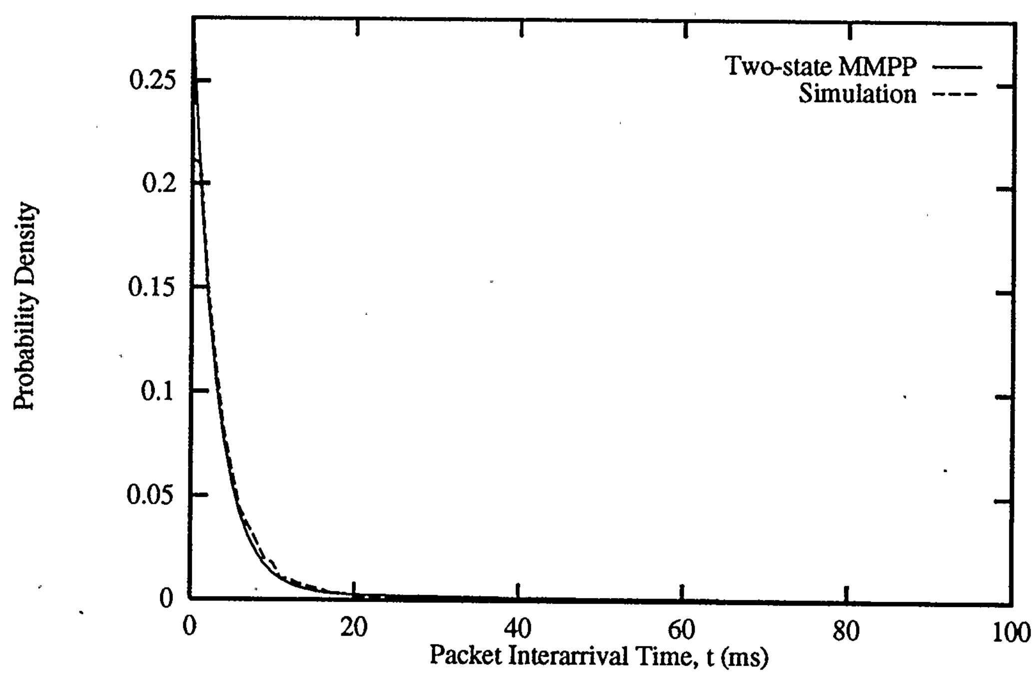

Figure 3.13. Packet Interarrival Time Probability Density, $s = 2, a_1 = 0.5, t_p = 5ms$

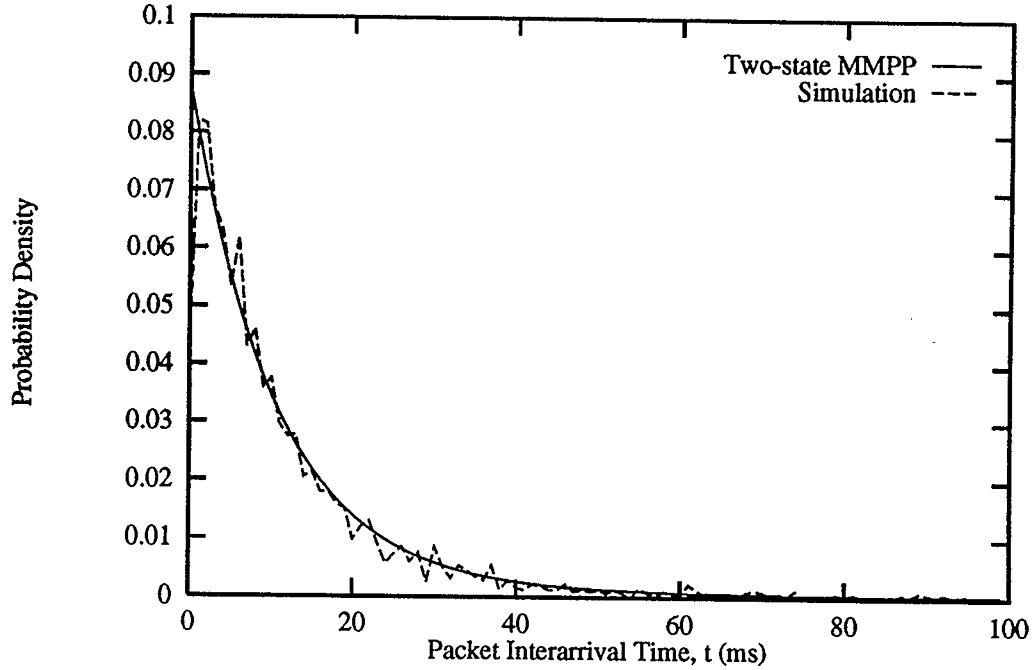

Figure 3.14. Packet Interarrival Time Probability Density, $s = 4, a_1 = 0.25, t_p = 5ms$

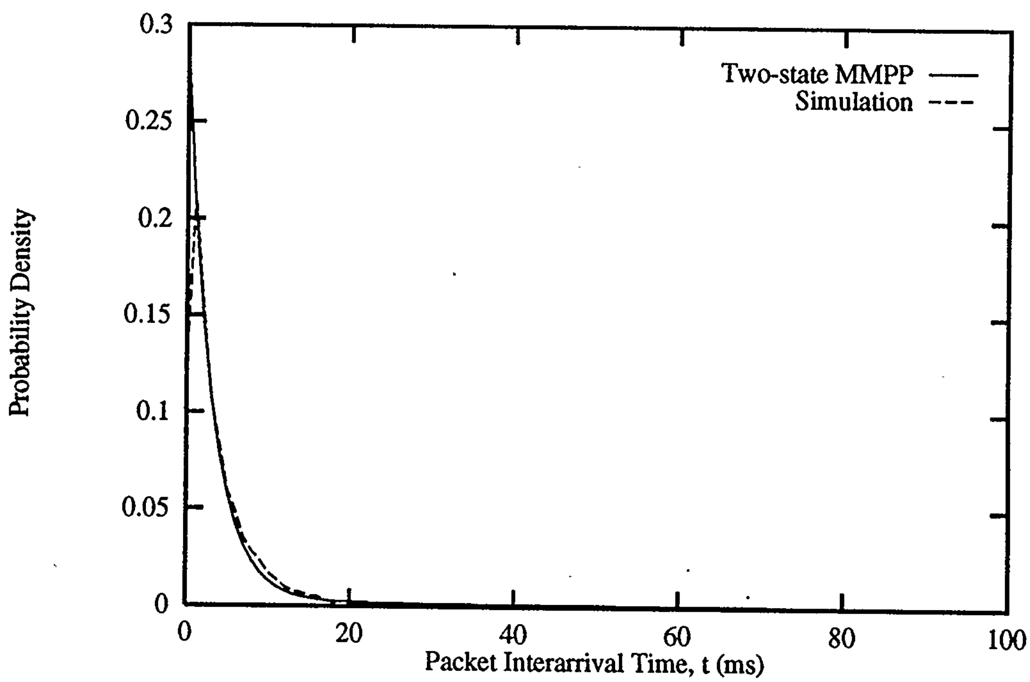

Figure 3.15. Packet Interarrival Time Probability Density, $s = 4, a_1 = 0.5, t_p = 5ms$

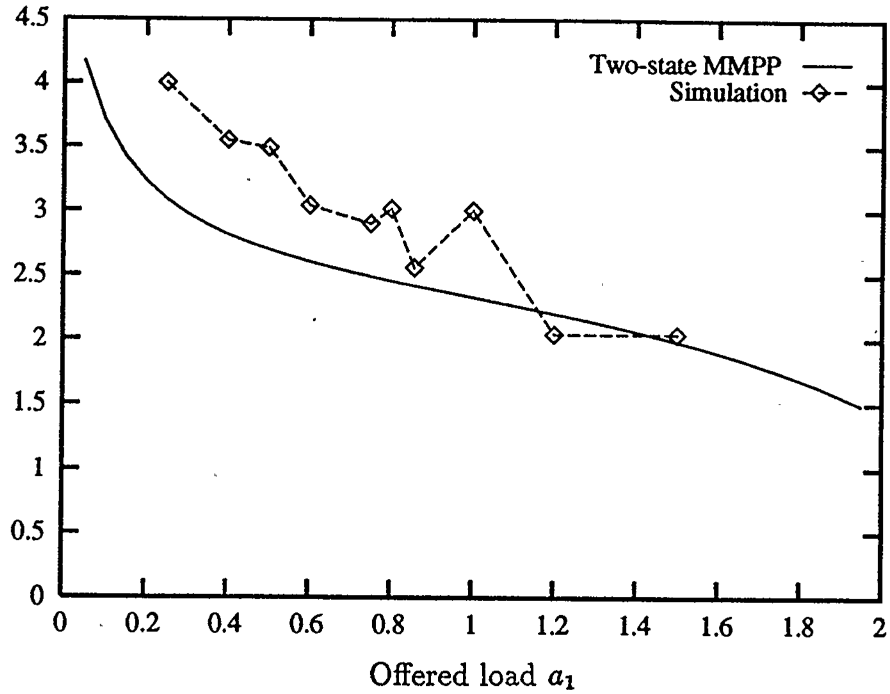

Figure 3.16. Coefficient of Variation of Packet Interarrival Time, $s = 2, \lambda_p/\mu_1 = 12$

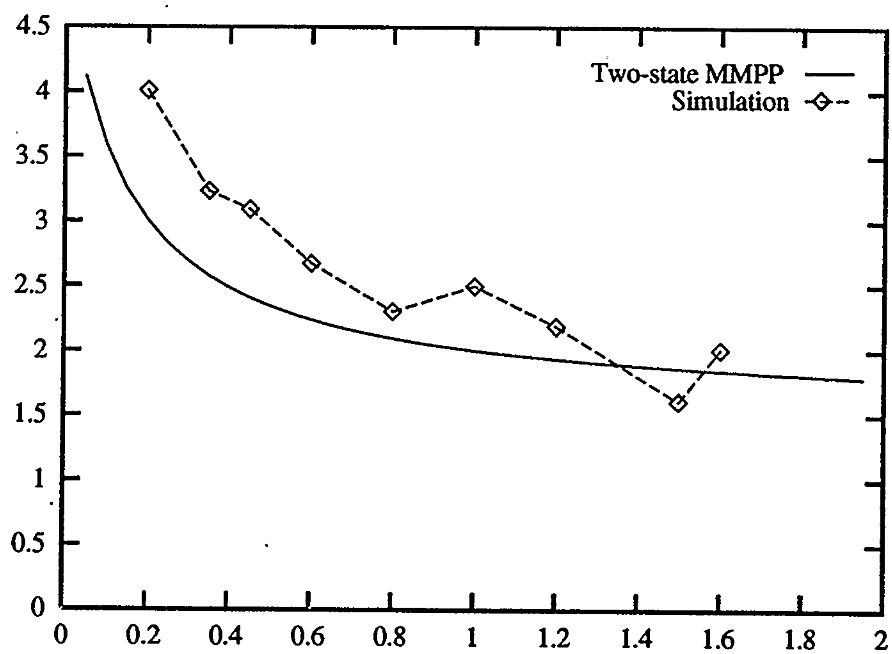

Figure 3.17. Coefficient of Variation of Packet Interarrival Time, $s = 4, \lambda_p/\mu_1 = 12$

Offered load a_1

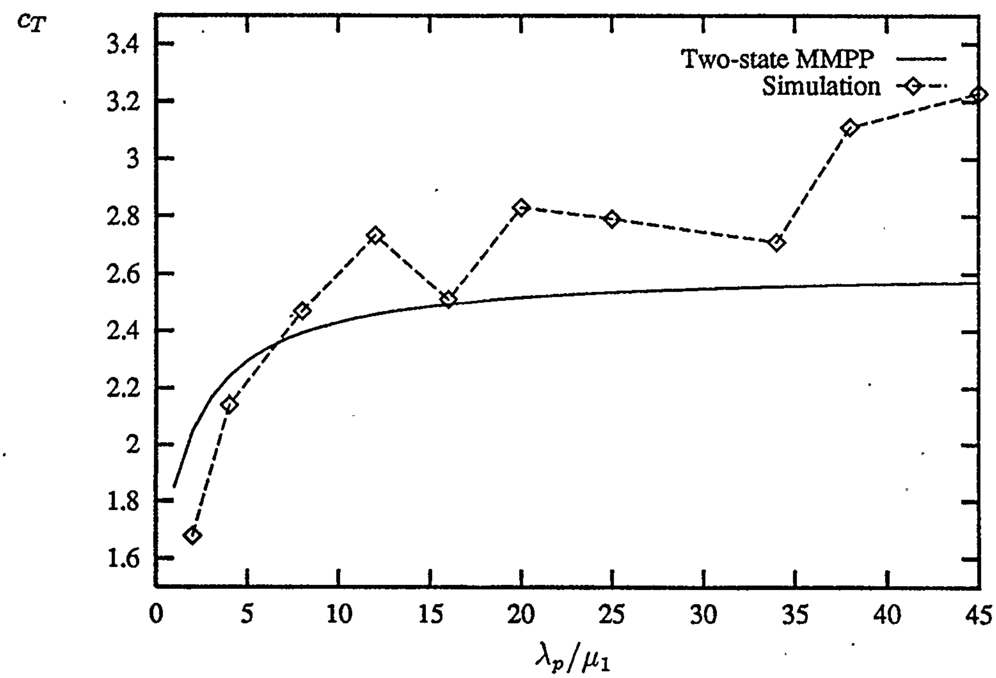

Figure 3.18. Coefficient of Variation of Packet Interarrival Time, $s = 2, a_1 = 0.8$

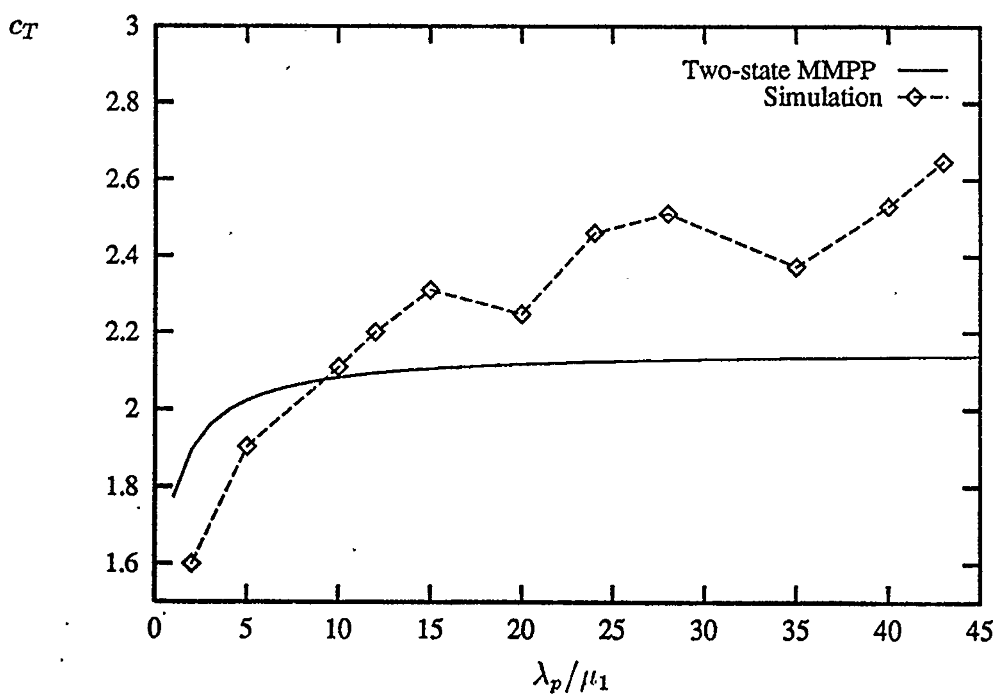

Figure 3.19. Coefficient of Variation of Packet Interarrival Time, $s = 4, a_1 = 0.8$

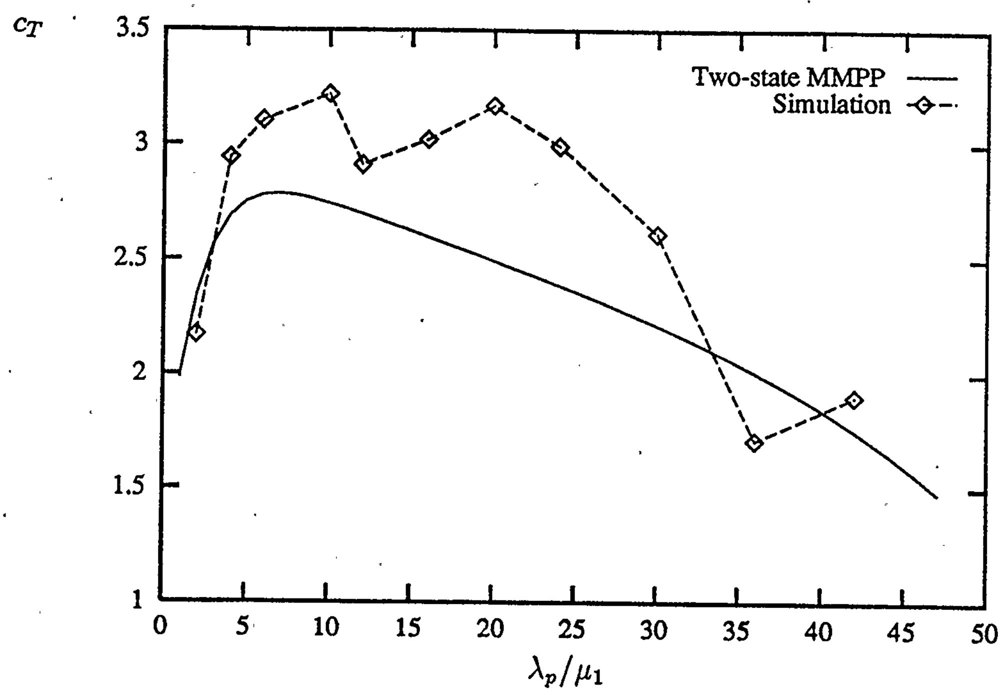

Figure 3.20. Coefficient of Variation of Packet Interarrival Time, $s = 2, t_c = 120ms$

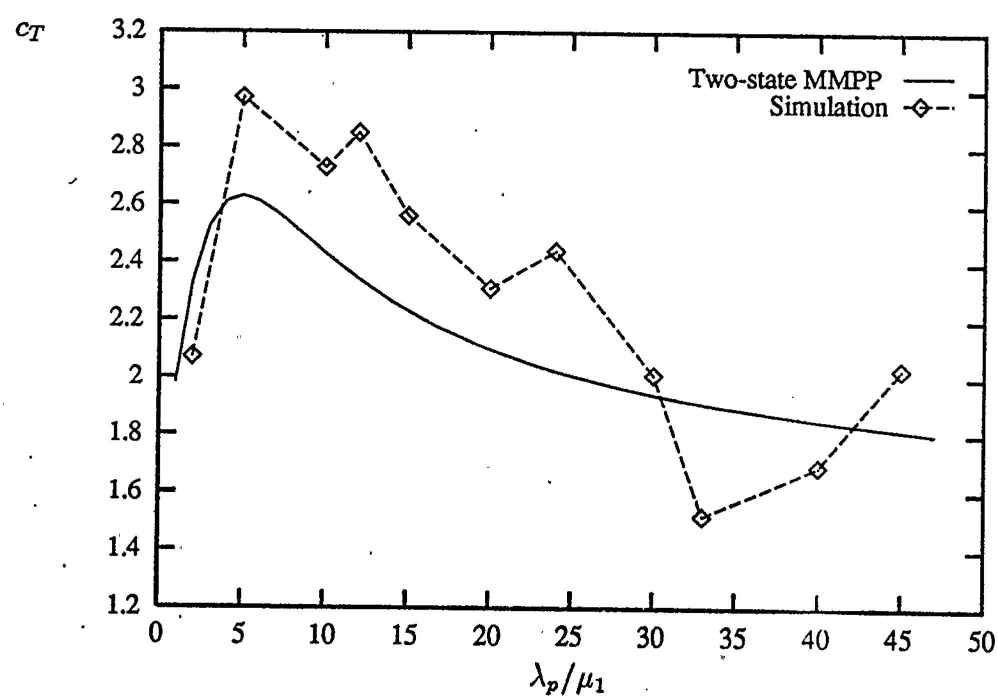

Figure 3.21. Coefficient of Variation of Packet Interarrival Time, $s = 4, t_c = 120ms$

respectively. In both figures s is equal to 2 but the mean packet rates or the offered load a_1 are different. Fig. 3.14 and Fig. 3.15 give the results for s=4.

Fig. 3.16 and Fig. 3.17 show the coefficient of variation of the packet interarrival time c_T as a function of the offered load a_1 for s=2 and s=4 respectively. In these figures, we keep the mean number of packets per call, i.e. λ_p/μ_1, unchanged. We see that c_T decreases with increasing a_1.

Fig. 3.18 and Fig. 3.19 show c_T as a function of the mean number of packets per call for s=2 and s=4 respectively. In both figures, a_1 is kept constant. The results show that, for given a_1, c_T increases while λ_p/μ_1 grows. It indicates that λ_p/μ_1 is one of the important elements which influence the burstiness of the packet process.

In Fig. 3.20 and Fig. 3.21 with s=2 and s=4 respectively, we set $t_c = 120ms$ and $t_p = 5ms$ and change μ_1. Both λ_p/μ_1 and $a_1 = \lambda_c/\mu_1$ change as μ_1 varies. The results reflect the compound effects of λ_p/μ_1 and a_1 on c_T, which are shown in Fig. 3.16 to Fig. 3.19.

Comparing the approximation with simulation results, we see that the hyperexponential distribution is a good approximation for the packet process in the sense that the density function is closer to the simulation result and the distribution can yield good approximations for the first and second moments of the packet interarrival time.

It should be pointed out that the parameters, α, β_1 and β_2 of the hyperexponential distribution (3.102) are determined by the parameters λ_1, λ_2, r_1 and r_2, so the accuracy of the approximation of the interarrival time distribution depends on the choice of these parameters.

3.5 SUMMARY

In this chapter, first, we use the maximum entropy method to develop two interarrival time distribution formulas for a batch arrival process. In formula (3.22) we use the first moment of interarrival time as a constraint. In formula (3.54) we use both the first and second moments of interarrival time as constraints. Through the analyses and comparisons we have shown that when the message arrival process is Poisson with random message length, formula (3.22) yields very accurate results for the interarrival time distribution of the batch arrival process. If the message process is not Poisson, formula (3.54) yields a better approximation.

Secondly, we use a two-state Markov-modulated Poisson process to approximate a doubly stochastic Poisson process. We also analyz the statistical properties of the traffic with the two models and obtaine an interarrival time distribution for the two-state MMPP. In the analysis it is shown that the Markov-modulated Poisson process can be used to represent the characteristics for both the burstiness and correlation of the traffic.

CHAPTER 4

PERFORMANCE ANALYSIS OF QUEUEING SYSTEMS

In this chapter we shall discuss the performances of queues with batch process or doubly stochastic Poisson process as an input. In section 4.1, queueing models with batch arrival process are studied. In section 4.2, queueing models with doubly stochastic Poisson process input are investigated. In both sections we shall utilize the analytical results of those input processes obtained in chapter 3 to derive the mean delay, the mean queue length, the waiting time distribution and the state probability distribution for the considered queues.

4.1 PERFORMANCE ANALYSIS OF QUEUES WITH BATCH ARRIVAL PROCESSES

Batch-arrival queueing models can be used in many practical situations, such as the analysis of message packetization in data communication systems. In general it is difficult to find tractable expressions for the probability distributions, such as the waiting time distribution and the state probability distribution. It is, therefore, useful to have easily computable approximations for these probabilities. In this section, we shall give approximations for the $M^X/G/1$ model and the $G^X/M/1$ model by using the principle of maximum entropy. Also we shall discuss the methods to calculate the mean delay in these queueing systems.

4.1.1 Delay in the $M^X/G/1$ Queue

The batch processes discussed in this section satisfy the conditions given in section 3.3.1. If the message arrivals follow a Poisson process with rate λ_c, the packet process is a Poisson batch arrival M^X input process.

In a $M^X/G/1$ queue, the service times of packets, S, are independent identically distributed random variables with distribution function $F_S(t)$ and the Laplace-Stieltjes transform $F_S^*(s)$. We assume $F_s(0) = 0$ and the mean service rate be μ.

For convenience we introduce the following notations. Let

. D be the delay of a packet in the queueing system.

. N_q be the queue length.

. W be the waiting time of an arbitrary test packet in a batch, W_1 the waiting time of the first packet in the batch, and W_2 the waiting time caused by packets which are in the same batch and are served before the test packet.

. $F_X(t)$ be the distribution function of X and $F_X^*(s)$ be the corresponding Laplace-Stieltjes transform.

. $B^*(s)$ denote the Laplace-Stieltjes transform of the total amount of service time required by all packets belonging to one batch.

For a $M^X/G/1$ queue with t_p equal to zero, because W is the sum of W_1 and W_2 and W_1 and W_2 are independent random variables, it has been shown that [4]

$$F_W^*(s) = F_{W_1}^*(s)F_{W_2}^*(s) \tag{4.1}$$

and

$$E(W) = E(W_1) + E(W_2) \tag{4.2}$$

where

$$F^*_{W_1}(s) = \frac{(1-\rho)s}{s - \lambda_c(1 - B^*(s))} \tag{4.3}$$

$$F^*_{W_2}(s) = \frac{1 - B^*(s)}{E(L)(1 - F^*_S(s))} \tag{4.4}$$

and

$$\rho = \frac{E(L)\lambda_c}{\mu} \tag{4.5}$$

From (4.3) and (4.4), we get, respectively

$$E(W_1) = \frac{\rho[E(L^2)/E(L) + c_S^2]}{2\mu(1-\rho)} \tag{4.6}$$

and

$$E(W_2) = \frac{E(L^2)/E(L) - 1}{2\mu} \tag{4.7}$$

where c_S^2 is the coefficient of variation of the service time.

So we have the mean waiting time of the packet

$$E(W) = \frac{\rho(1 + c_S^2) + E(L^2)/E(L) - 1}{2\mu(1-\rho)} \tag{4.8}$$

The mean delay is given by

$$E(D) = \frac{\rho(c_S^2 - 1) + E(L^2)/E(L) + 1}{2\mu(1-\rho)} \tag{4.9}$$

By means of Little's formula, we find the mean queue length

$$\begin{aligned} E(N_q) &= \lambda_c E(L) E(W) \\ &= \frac{\rho^2(1 + c_S^2) + \rho E(L^2)/E(L) - \rho}{2(1-\rho)} \end{aligned} \tag{4.10}$$

For a $M^X/G/1$ queue with nonzero t_p, after taking account of the effect of t_p we get the approximate mean waiting time of the packet

$$E(W) = \frac{\rho(1+c_S^2) + E(L^2)/E(L) - 1}{2\mu(1-\rho)} - \frac{t_p E(L)}{2}[1 - F_S(t_p)] \tag{4.11}$$

The mean delay is

$$E(D) = \frac{\rho(1+c_S^2) + E(L^2)/E(L) - 1}{2\mu(1-\rho)} - \frac{t_p E(L)}{2}[1 - F_S(t_p)] + 1/\mu \tag{4.12}$$

and the mean queue length is

$$E(N_q) = \frac{\rho^2(1+c_S^2) + \rho E(L^2)/E(L) - \rho}{2(1-\rho)} - \frac{t_p E(L)^2 \lambda_c}{2}[1 - F_S(t_p)] \tag{4.13}$$

Now we assume that the message length L has a geometric distribution

$$P(L = i) = p(1-p)^{i-1}, \quad i = 1, 2, ... \tag{4.14}$$

with mean $E(L) = 1/p$ and $E(L^2) = (2-p)/p^2$, then for the $M^X/G/1$ queue with zero t_p, the mean waiting time reduces to

$$E(W) = \frac{\rho(1+c_S^2) + (2/p - 2)}{2\mu(1-\rho)} \tag{4.15}$$

or

$$E(W) = \frac{\rho(1+c_S^2) + (2E(L) - 2)}{2\mu(1-\rho)} \tag{4.16}$$

The mean delay becomes

$$E(D) = \frac{\rho(c_S^2 - 1) + 2E(L)}{2\mu(1-\rho)} \tag{4.17}$$

and the mean queue length is

$$E(N_q) = \frac{\rho^2(1+c_S^2) + 2\rho E(L) - \rho}{2(1-\rho)} \tag{4.18}$$

For the $M^X/G/1$ queue with nonzero t_p, the mean waiting time becomes

$$E(W) = \frac{\rho(1+c_S^2)+E(L)}{2\mu(1-\rho)} - \frac{t_pE(L)}{2}[1-F_S(t_p)] \tag{4.19}$$

The mean delay is

$$E(D) = \frac{\rho(1+c_S^2)+2E(L)-2}{2\mu(1-\rho)} - \frac{t_pE(L)}{2}[1-F_S(t_p)] + 1/\mu \tag{4.20}$$

and the mean queue length is

$$E(N_q) = \frac{\rho^2(1+c_S^2)+\rho E(L)}{2(1-\rho)} - \frac{t_pE(L)^2\lambda_c}{2}[1-F_S(t_p)] \tag{4.21}$$

4.1.2 The Waiting Time Distribution for the $M^X/G/1$ Queue

We shall derive the waiting time distribution for the $M^X/G/1$ queue by means of the entropy maximization method.

First we should determine if the distribution function $F_W(t)$ has a jump at t=0. Since [7]

$$\begin{aligned} F_W(0) &= F_W^*(\infty) \\ &= F_{W_1}^*(\infty)F_{W_2}^*(\infty) \end{aligned} \tag{4.22}$$

from (4.3) and (4.4), we have

$$F_W(0) = \frac{1-\rho}{E(L)} \tag{4.23}$$

This result indicates that $F_W(t)$ has a jump of $F_W(0)$ at t=0.

Let $f_W(t)$ be the density function of packet waiting time W. Since the distribution function $F_W(t)$ has a jump of $F_W(0)$ at t=0, thus $f_W(t)$ can be written as

$$f_W(t) = F_W(0)\delta(t) + f_{W_c}(t) \tag{4.24}$$

where $f_{W_c}(t)$ denotes the continuous part of the density function.

Let the entropy function of $f_{W_c}(t)$ be defined as

$$H = -\int_0^\infty f_{W_c}(t) \ln f_{W_c}(t) dt \tag{4.25}$$

The normalization condition is given by

$$F_W(0) + \int_0^\infty f_{W_c}(t) dt = 1 \tag{4.26}$$

and the mean of W is given by

$$\int_0^\infty t f_{W_c}(t) dt = E(W) \tag{4.27}$$

By maximizing the entropy function H in (4.25) subject to the constraints (4.26) and (4.27), we get the maximum entropy solution for $f_{W_c}(t)$ as

$$f_{W_c}(t) = exp\{-1 - \gamma_0 - \gamma_1 t\} \tag{4.28}$$

where γ_0 and γ_1 are the Lagrange multipliers.

Then substituting $f_{W_c}(t)$ into (4.26) and (4.27), we get

$$exp\{-1 - \gamma_0\} = \frac{[1 - F_W(0)]^2}{E(W)} \tag{4.29}$$

and

$$\gamma_1 = \frac{1 - F_W(0)}{E(W)} \tag{4.30}$$

Inserting (4.29) and (4.30) into (4.28), we obtain the density function

$$f_{W_c}(t) = \frac{[1 - F_W(0)]^2}{E(W)} exp\{-\frac{1 - F_W(0)}{E(W)} t\} \tag{4.31}$$

and the distribution function

$$F_W(t) = 1 - [1 - F_W(0)] exp\{-\frac{1 - F_W(0)}{E(W)} t\} \tag{4.32}$$

where E(W) is given by (4.16).

Substituting (4.23) and (4.16) into (4.32), we obtain the waiting time distribution as

$$F_W(t) = 1 - \frac{E(L) - 1 + \rho}{E(L)} exp\{-\frac{[E(L) - 1 + \rho][2\mu(1 - \rho)]}{E(L)[\rho(1 + c_S^2) + 2E(L) - 2]} t\} \tag{4.33}$$

4.1.3 State Probability Distribution for the $M^X/G/1$ Queue

By means of Little's formula, we calculate the average number of packets in the system as

$$E(N) = \lambda_c E(L) E(D) \tag{4.34}$$

$$= [\rho^2(c_S^2 - 1) + 2\rho E(L)]/[2(1 - \rho)] \tag{4.35}$$

Define the number of packets in the system as the state of the $M^X/G/1$ queue. We obtain the maximum entropy (ME) state probability distribution using (2.26)

$$p_n = \begin{cases} 1 - \rho & , n = 0 \\ \frac{2\rho(1-\rho)}{\rho(c_S^2-1)+2E(L)}[\frac{\rho(c_S^2+1)+2E(L)-2}{\rho(c_S^2-1)+2E(L)}]^{n-1} & , n \geq 1 \end{cases} \tag{4.36}$$

4.1.4 Mean Delay in the $G^X/M/1$ Queue

For the $G^X/M/1$ queue, the service time distribution is exponential. We can use the G/M/1 model to analyze the $G^X/M/1$ queue. For the G/M/1 queue, we have the mean waiting time, mean delay and mean queue length as [61]

$$E(W) = \frac{\sigma}{\mu(1 - \sigma)} \tag{4.37}$$

$$E(D) = \frac{1}{\mu(1 - \sigma)} \tag{4.38}$$

and

$$E(N_q) = \frac{\rho\sigma}{(1-\sigma)} \tag{4.39}$$

where σ is the root of the functional equation

$$\sigma = F_A^*(\mu - \mu\sigma) \quad , \quad 0 < \sigma < 1 \tag{4.40}$$

and $F_A^*(s)$ is the Laplace-Stieltjes transform of the packet interarrival time distribution $F_A(t)$.

In section 3.3, we have developed two equivalent packet interarrival time density functions (3.22) and (3.54) by means of the maximum entropy principle. Now we shall use them to calculate the mean waiting time and the mean delay for the $G^X/M/1$ queue.

If we use the first moment approximation of the interarrival time for the G^X input process in formula (3.22), we have

$$F_A^*(s) = [1 - \frac{\alpha}{E(L)} + \frac{\alpha\gamma_1}{E(L)(s+\gamma_1)}]e^{-st_p} \tag{4.41}$$

where α and γ_1 are given by (3.21) and (3.30), respectively.

Then σ satisfies (4.40), or

$$\sigma = [1 - \frac{\alpha}{E(L)} + \frac{\alpha\gamma_1}{E(L)(\mu - \mu\sigma + \gamma_1)}]e^{-(\mu-\mu\sigma)t_p} \tag{4.42}$$

After solving (4.42) for σ, we can calculate the mean values given in (4.37)-(4.39).

Now we consider the limiting case $t_p \to 0$. When t_p is equal to zero, we have $\alpha = 1$ and $\gamma_1 = \lambda_c$, see (3.31) and (3.33). Then (4.42) becomes

$$\sigma = 1 - \frac{1}{E(L)} + \frac{\lambda_c}{E(L)(\mu - \mu\sigma + \lambda_c)}) \tag{4.43}$$

The solution of (4.43) is given by

$$\sigma = 1 - \frac{1-\rho}{E(L)} \tag{4.44}$$

Substituting this σ into (4.37)-(4.39), we obtain the mean values for the $G^X/M/1$ queue

$$E(W) = \frac{E(L) - 1 + \rho}{\mu(1-\rho)} \tag{4.45}$$

$$E(D) = \frac{E(L)}{\mu(1-\rho)} \tag{4.46}$$

and

$$E(N_q) = \frac{\rho E(L) - \rho + \rho^2}{(1-\rho)} \tag{4.47}$$

If we use the second moment approximation of the interarrival time for the G^X input process in formula (3.54), we have

$$F_A^*(s) = 1 - \frac{1}{E(L)} + Z_p exp\{\frac{\gamma_1^2}{2\gamma_2}\} \int_0^\infty exp\{-\frac{(t+\gamma_1)^2}{2\gamma_2} - st\}dt \tag{4.48}$$

where Z_p, γ_1 and γ_2 are given by (3.55)-(3.57), respectively. And σ satisfies (4.40), or

$$\sigma = 1 - \frac{1}{E(L)} + Z_p exp\{\frac{(\gamma_1 + \mu\gamma_2 - \mu\sigma\gamma_2)^2}{2\gamma_2}\} \int_{\gamma_1+\mu\gamma_2-\mu\sigma\gamma_2}^\infty exp\{-\frac{t^2}{2\gamma_2}\}dt \tag{4.49}$$

After solving (4.49) for σ, we can calculate the mean waiting time, mean delay and mean queue length by (4.37)-(4.39).

4.1.5 The Waiting Time Distribution and The State Probability Distribution for The $G^X/M/1$ Queue

Using the G/M/1 model, we have the waiting time distribution as [61]

$$F_W(t) = 1 - \sigma e^{-\mu(1-\sigma)t} \tag{4.50}$$

and the state probability distribution as [61]

$$p_k = \begin{cases} 1-\rho & ,k=0 \\ \rho(1-\sigma)\sigma^{k-1} & ,k \geq 1 \end{cases} \tag{4.51}$$

If we consider the case of zero t_p and use the first moment approximation of the interarrival time for the G^X input process, σ is given by (4.44). Then (4.50) and (4.51) become, respectively

$$F_W(t) = 1 - [\frac{E(L) - 1 + \rho}{E(L)}]exp\{-\frac{\mu(1-\rho)}{E(L)}t\} \tag{4.52}$$

and

$$p_k = \begin{cases} 1-\rho & ,k=0 \\ \frac{\rho(1-\rho)}{E(L)}[1 - \frac{1-\rho}{E(L)}]^{k-1} & ,k \geq 1 \end{cases} \tag{4.53}$$

4.1.6 Numerical Results

When we calculate the numerical results in this section, we will consider three special cases. The first is the $M^X/M/1$ queue where the input is a bulk Poisson process and the service time distribution is exponential. The second is the $M^X/D/1$ queue where the input is the bulk Poisson process either but the service time is constant. The third is the $G^X/M/1$ queue where the input is assumed to be uniformly distributed message arrivals with random packet length, and the service time is exponentially distributed.

Fig. 4.1 and Fig. 4.2 show the mean delay of packet in a $M^X/M/1$ queue as a function of ρ or E(L) for given t_c, t_p and $\tau = 1/\mu$ in the queue. By comparing these two figures, we find that an increase in t_p will reduce the mean delay of the packet. In Fig. 4.3 and Fig. 4.4, the mean delay is shown as a function of E(L) for given t_p, ρ and τ. We see that for the same value of ρ, an increase in E(L) will increase the mean delay of the packet in a $M^X/M/1$ queue.

In Fig. 4.5 and Fig. 4.6, the service time is constant. The effect of E(L) on the mean delay and the effect of t_p on the mean delay are the same as those in the $M^X/M/1$ queue as shown in Fig. 4.1 and Fig. 4.2.

In Fig. 4.7 and Fig. 4.8, the message arrivals of the batch input processes to the single server queues are assumed to be uniformly distributed over the interval of 0 to $2t_c$. The service time distribution is exponential. In these two figures, the mean delays calculated based on the two different interarrival time density function, formula (3.22) and formula (3.54) respectively, are presented and compared with the simulation results. We see that the results based on formula (3.54) are more accurate than those based on formula (3.22).

Fig. 4.9 and Fig. 4.10 show the waiting time probability density of the $M^X/M/1$ queue for zero t_p and 5ms, respectively. We see that the simulated results are closely matched with the theoretical results. The waiting time density function of the $M^X/D/1$ queue are shown in Fig. 4.11 and Fig. 4.12. Since we only use the first moment of the waiting time to derive the waiting time distribution for the $M^X/G/1$ queue in section 4.1.2, there is a difference between the theoretical results and the simultion results. But function $F_W(t)$ in (4.33) can be used to approximate the waiting time distribution for the $M^X/G/1$ queue.

In Fig. 4.13 and Fig. 4.14 the waiting time density function of the same queue as in Fig. 4.7 and Fig. 4.8 are presented. As expected, the results based on formula (3.54) are closer to the simulation results than those based on formula (3.22).

From the above discussions and the results we may conclude that our results obtained from the performance analysis are fairly accurate.

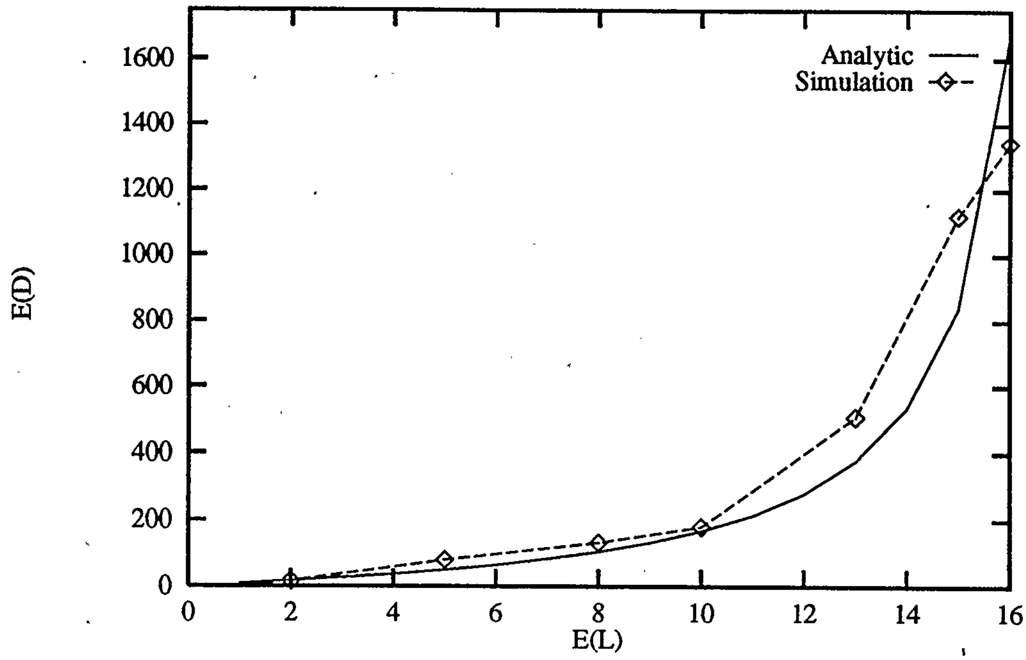

Figure 4.1. Mean Delay in the $M^X/M/1$ Queue, $t_p = 0, t_c = 120ms, \tau = 7ms$

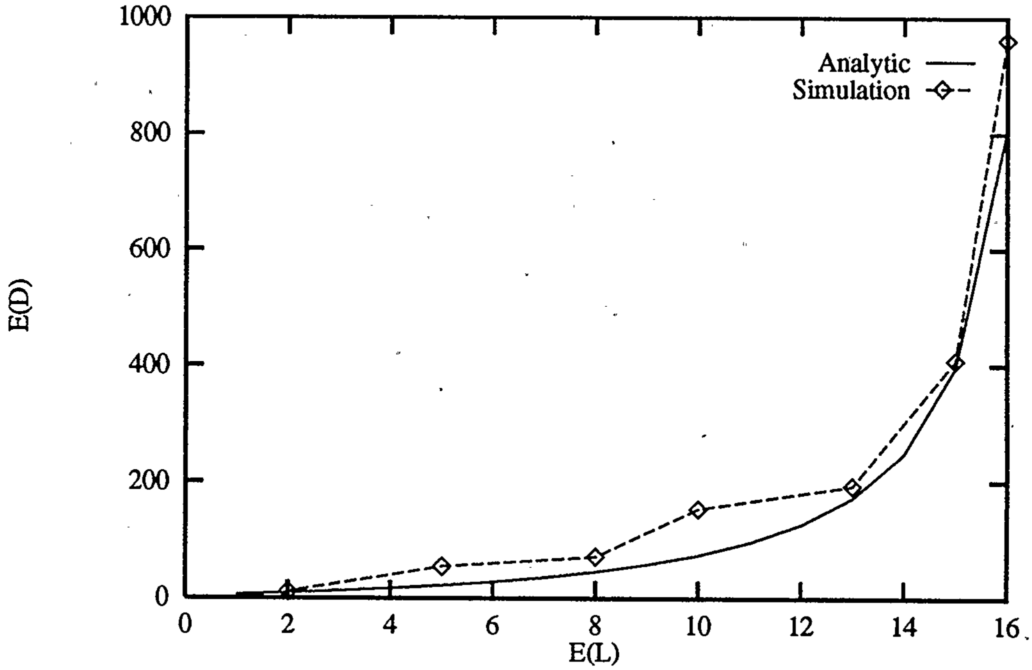

Figure 4.2. Mean Delay in the $M^X/M/1$ Queue, $t_p = 5ms, t_c = 120ms, \tau = 7ms$

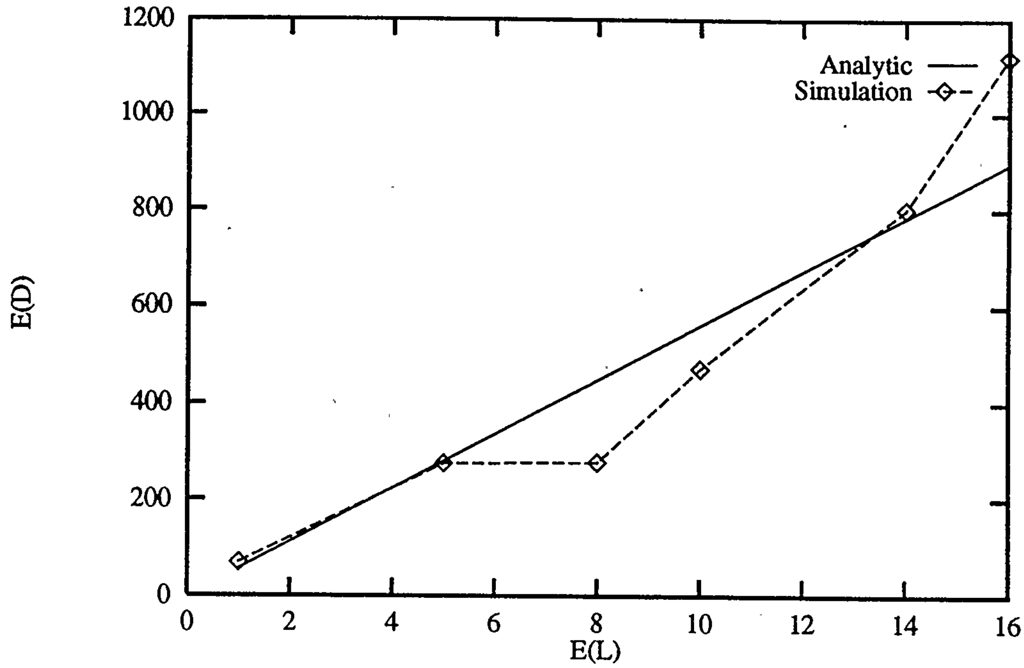

Figure 4.3. Mean Delay in the $M^X/M/1$ Queue, $t_p = 0, \rho = 0.875, \tau = 7ms$

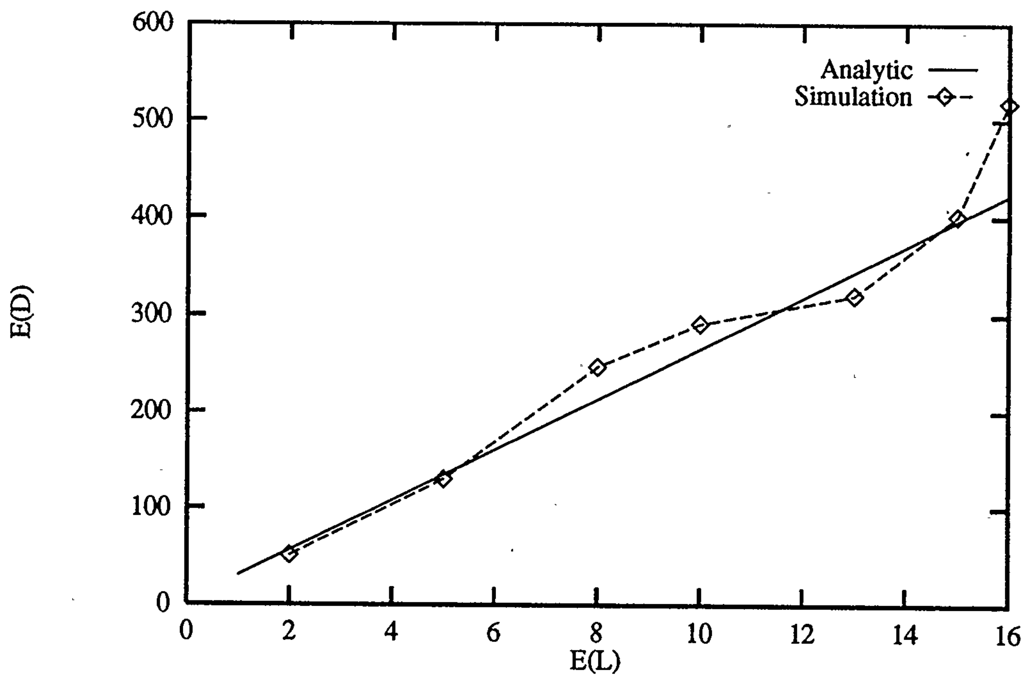

Figure 4.4. Mean Delay in the $M^X/M/1$ Queue, $t_p = 5ms, \rho = 0.875, \tau = 7ms$

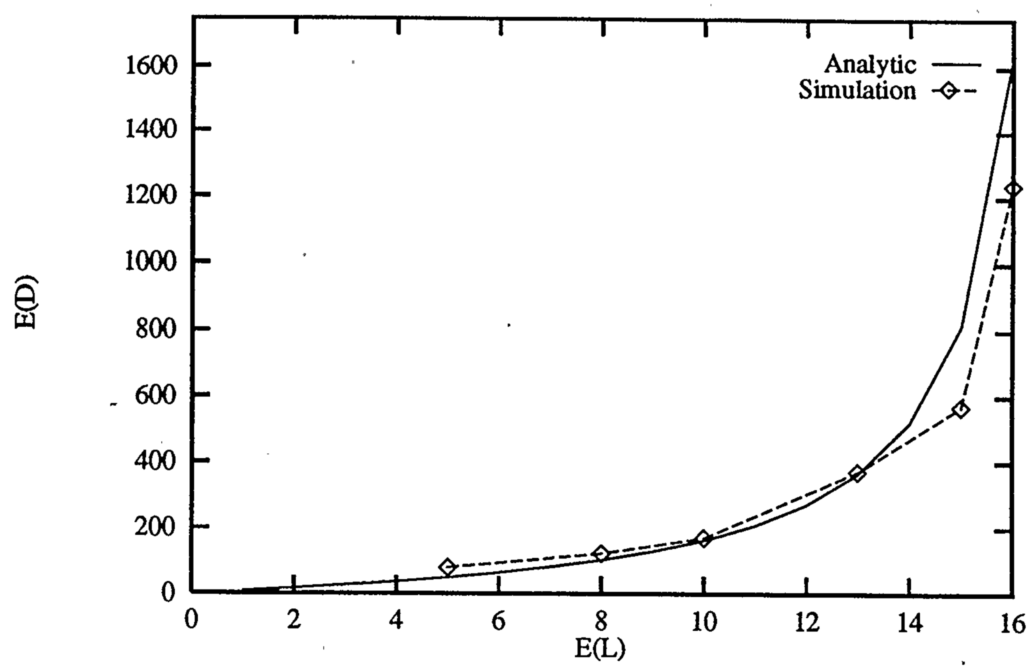

Figure 4.5. Mean Delay in the $M^X/D/1$ Queue, $t_p = 0, t_c = 120ms, \tau = 7ms$

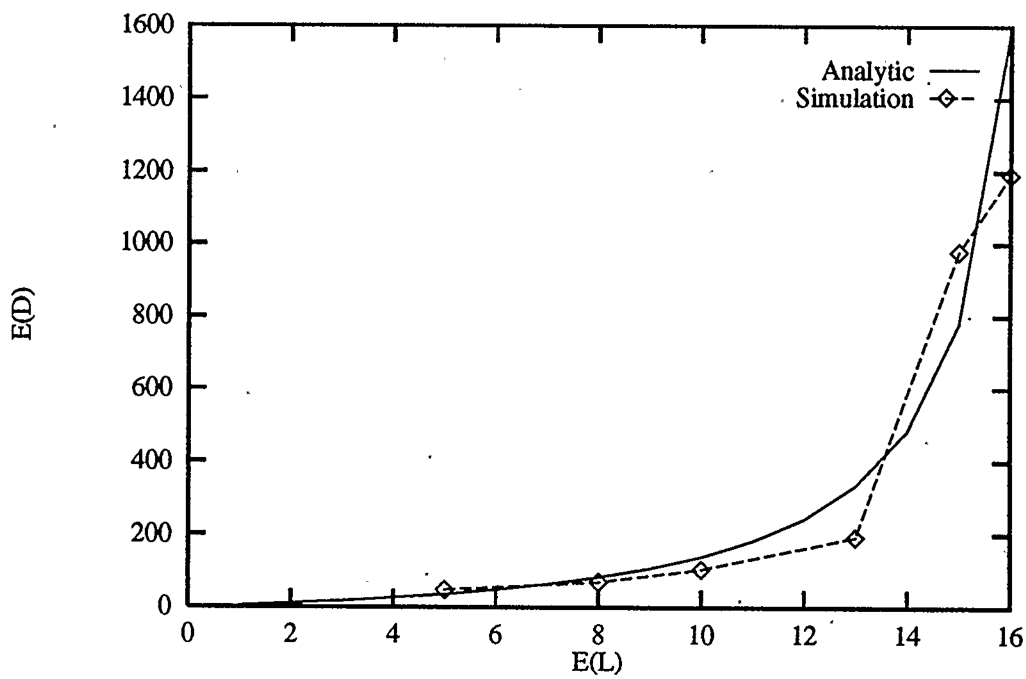

Figure 4.6. Mean Delay in the $M^X/D/1$ Queue, $t_p = 5ms, t_c = 120ms, \tau = 7ms$

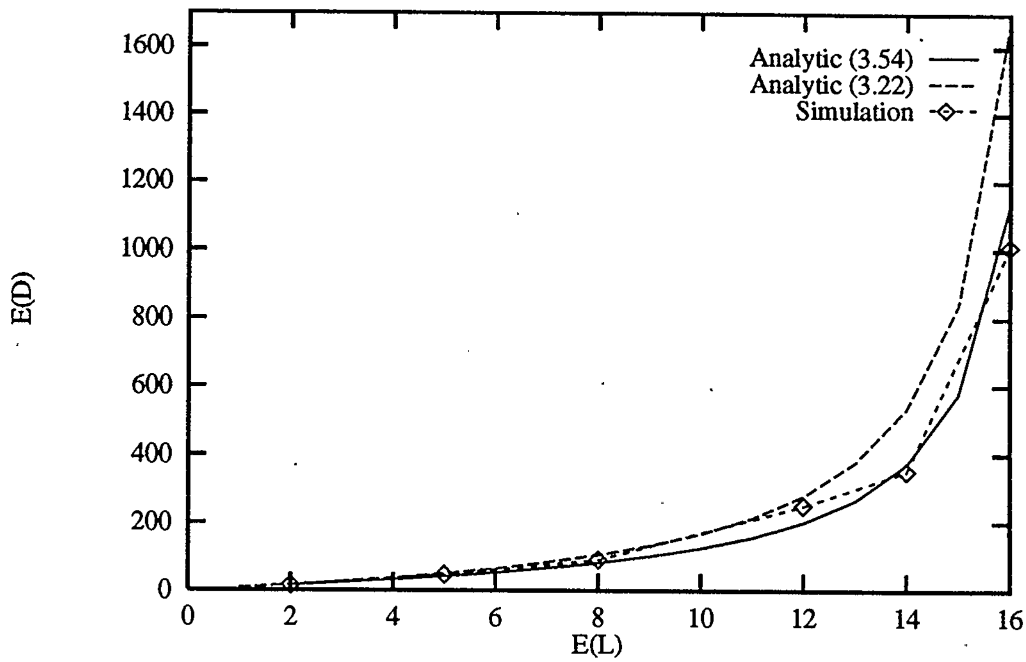

Figure 4.7. Mean Delay in the $G^X/M/1$ Queue, $t_p = 0, t_c = 120ms, \tau = 7ms$

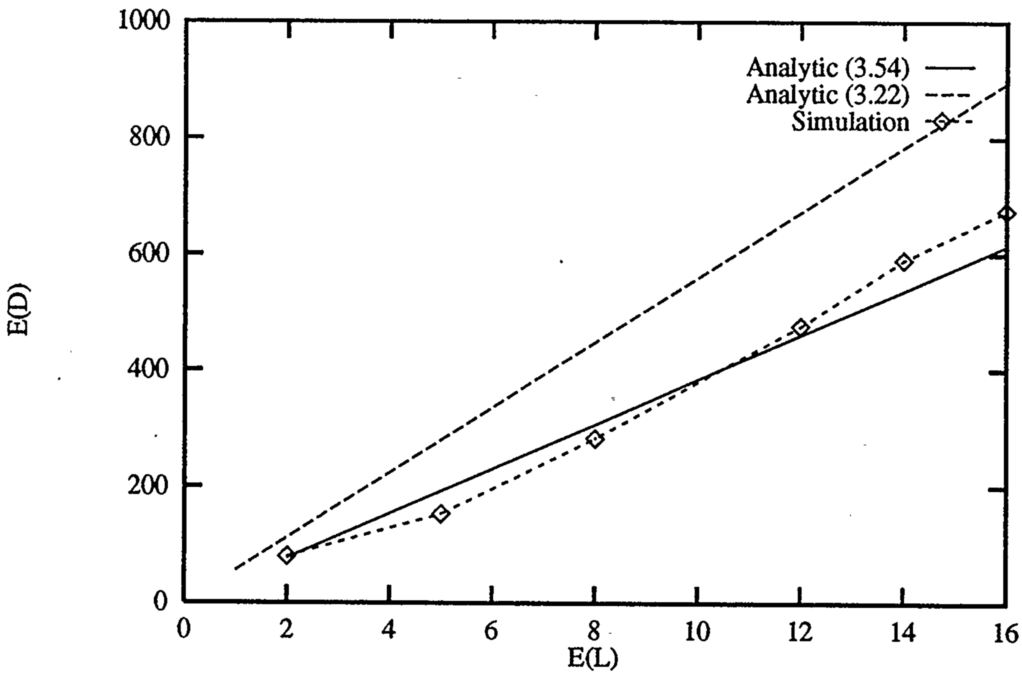

Figure 4.8. Mean Delay in the $G^X/M/1$ Queue, $t_p = 0, \rho = 0.875, \tau = 7ms$

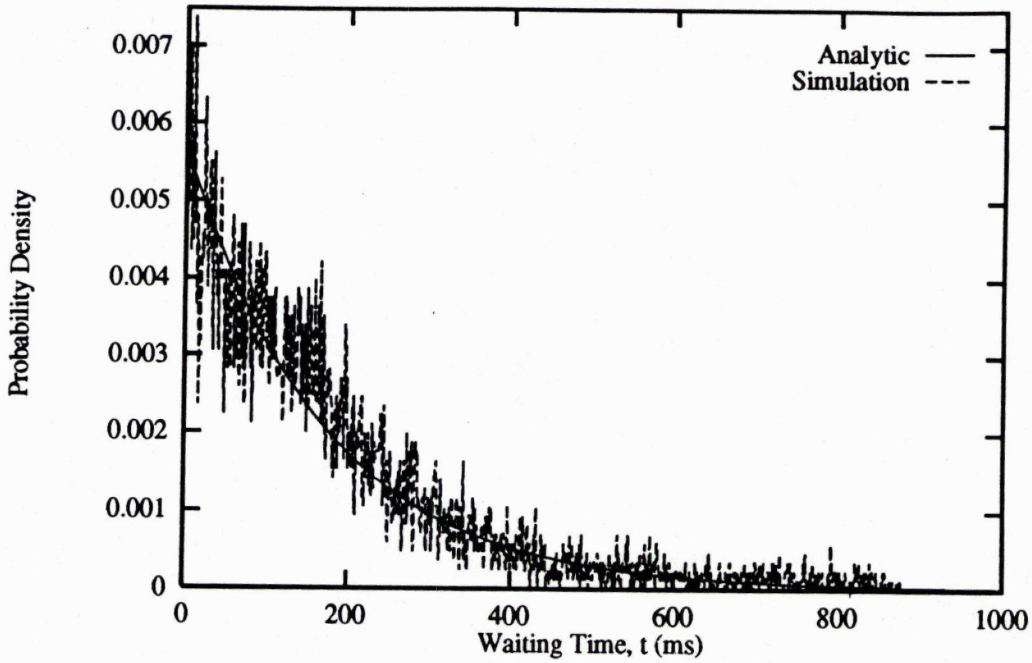

Figure 4.9. Waiting Time Probability Density for the $M^X/M/1$ Queue

$t_p = 0, t_c = 120ms, E(L) = 10, \tau = 7ms$

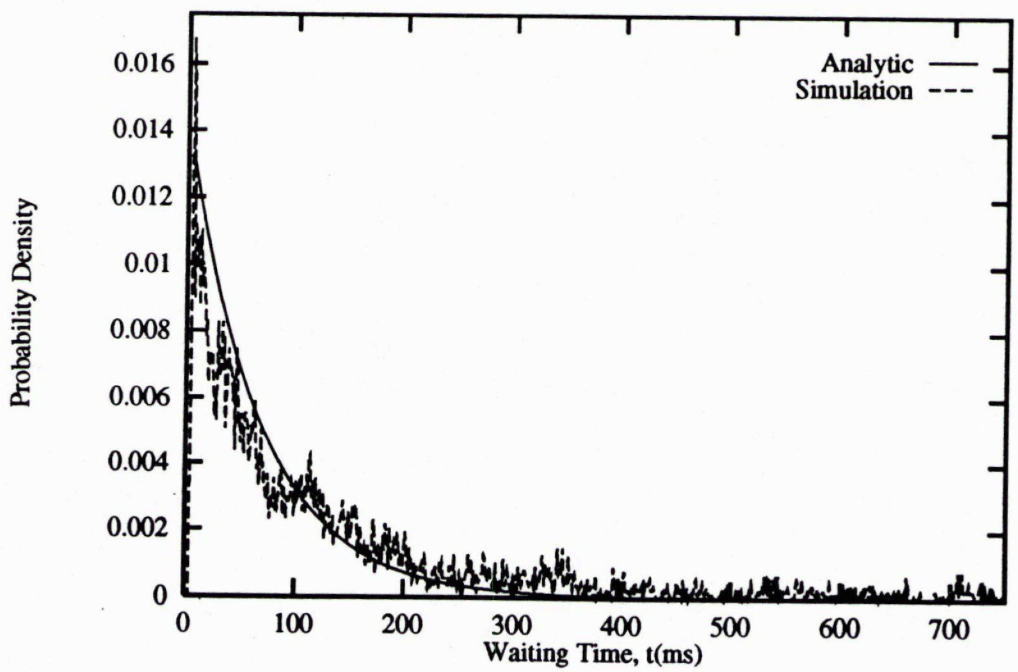

Figure 4.10. Waiting Time Probability Density for the $M^X/M/1$ Queue

$t_p = 5ms, t_c = 120ms, E(L) = 10, \tau = 7ms$

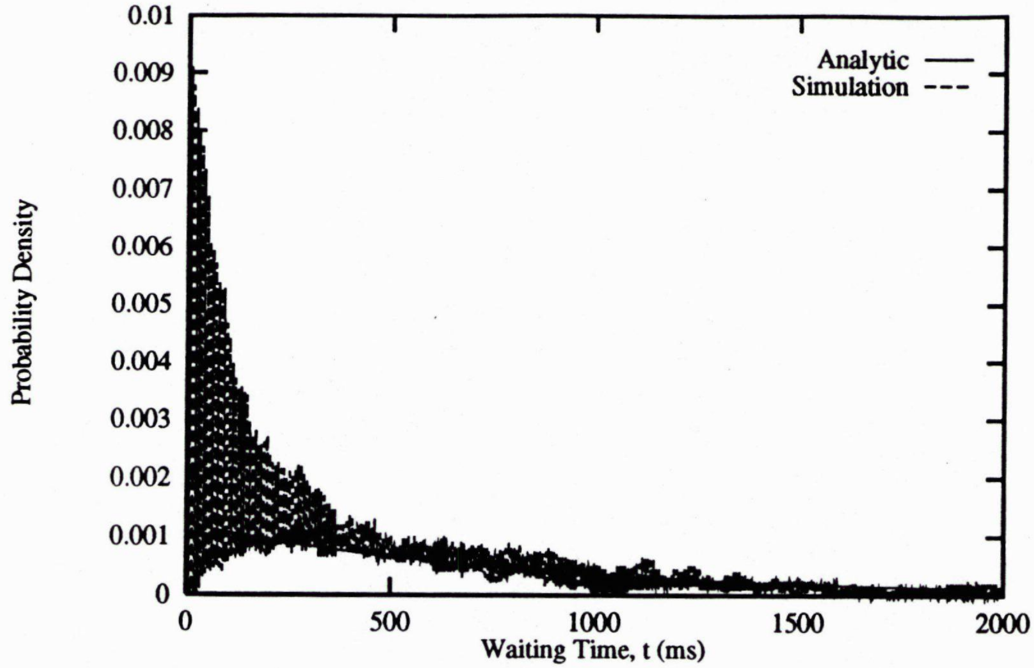

Figure 4.11. Waiting Time Probability Density for the $M^X/D/1$ Queue

$t_p = 0, t_c = 120ms, E(L) = 15, \tau = 7ms$

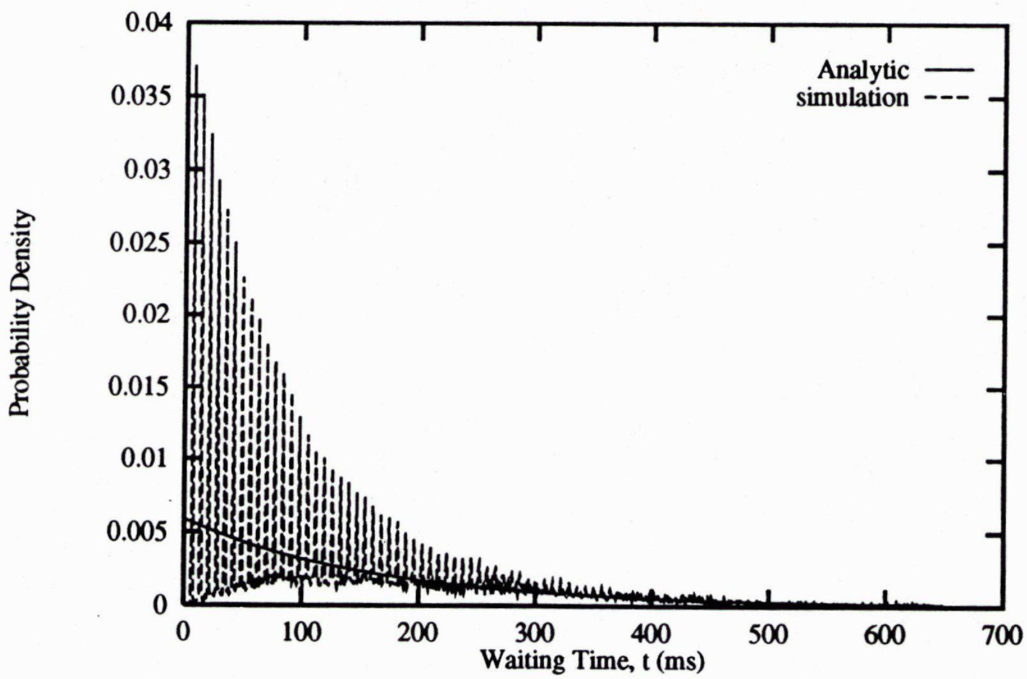

Figure 4.12. Waiting Time Probability Density for the $M^X/D/1$ Queue

$t_p = 0, t_c = 120ms, E(L) = 10, \tau = 7ms$

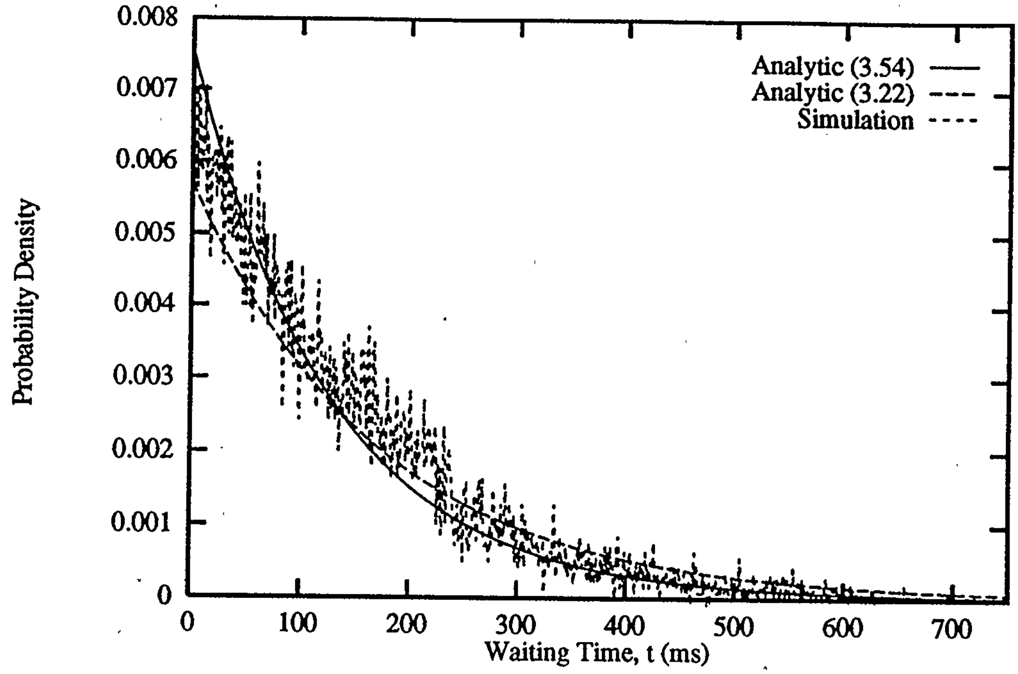

Figure 4.13. Waiting Time Probability Density for the $G^X/M/1$ Queue
$t_p = 0, t_c = 120ms, E(L) = 10, \tau = 7ms$

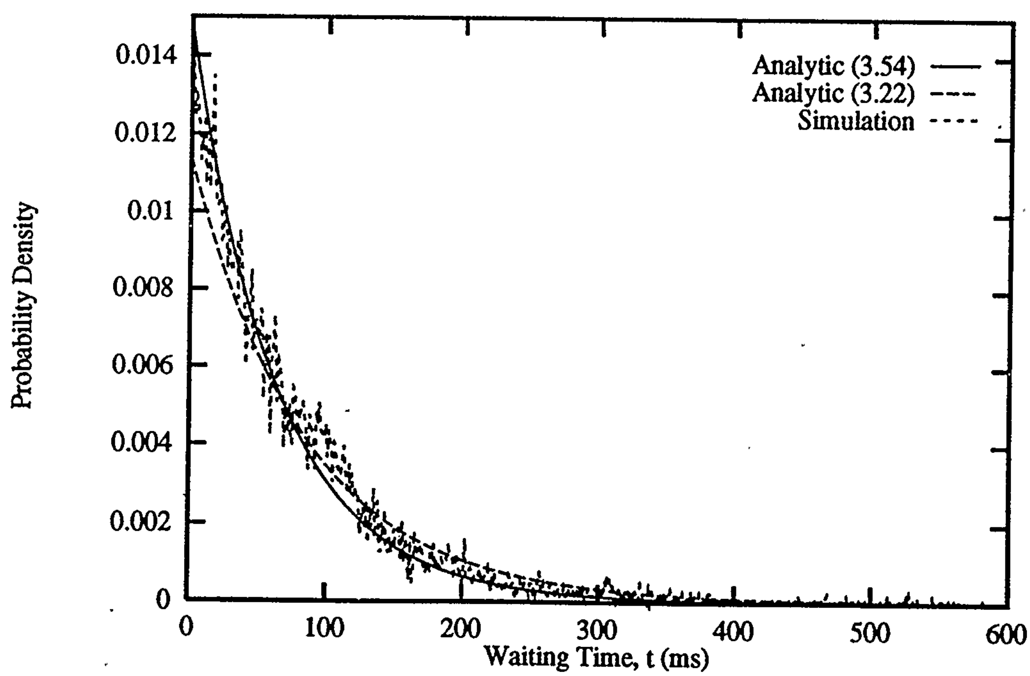

Figure 4.14. Waiting Time Probability Density for the $G^X/M/1$ Queue
$t_p = 0, t_c = 60ms, E(L) = 10, \tau = 7ms$

4.2 PERFORMANCE ANALYSIS OF THE MMPP/M/1 QUEUE

Queueing models that deal with the Markov-modulated Poisson process (MMPP) as an input have been studied by several authors. A single-server queue with general service time distribution and multilevel input has been studied by Neuts[71], where algorithmic results are presented, and in [72] results for the exponential service time case are also presented. In [73] Kuczura treats the superposition of a Poisson and interrupted Poisson process(IPP), which is equivalent to a two-state MMPP. The results in [73] are applied to the problem in [12] by Heffes. More recently in [21] Heffes and Lucantoni provide a method for the analysis of the performance of a statistical multiplexer with inputs consisting of the superposition of voice streams together with data streams, which is modeled as a MMPP/G/1 queue.

In this section we shall consider a single exponential server queueing system with a doubly stochastic Poisson process(DSPP) input. Since a DSPP may be approximated by a two-state MMPP, we shall model the queueing system as a MMPP/M/1 queue and use the G/M/1 model to analyze it.

4.2.1 Measurement Method For the MMPP Input

As we know, in order to apply the G/M/1 model to the performance analysis of the MMPP/M/1 queue, we have to determine the interarrival time distribution of the MMPP. In section 3.4.3, we have derived the interarrival time distribution for the MMPP given by (3.102). The distribution is a hyperexponential distribution characterizedd by parameters $\lambda_1, \lambda_2, r_1$ and r_2. Since $\lambda_1, \lambda_2, r_1$ and r_2 are determined by the original DSPP input process, the way we choose these parameters is very important to the use of (3.102) or (3.103) as the interarrival time distribution for the input process. In section 3.4.3 we use a method by Heffes for choosing those four parameters [12]. Here we introduce another method to obtain the parameters α, β_1, and β_2 for the distribution (3.102).

Let T be the interarrival time of the packet. Then the first three moments of T are, respectively, denoted by

$$m_1 = E(T) \tag{4.54}$$

$$m_2 = E(T^2) \tag{4.55}$$

and

$$m_3 = E(T^3) \tag{4.56}$$

Assume m_1, m_2 and m_3 can be obtained by measurement. When the input process is ergodic, we can use the time averages of the interarrival time as the measurement results of m_1, m_2 and m_3. Using (3.103) we establish the following equations

$$\frac{\alpha}{\beta_1} + \frac{1-\alpha}{\beta_2} = m_1 \tag{4.57}$$

$$\frac{2\alpha}{\beta_1^2} + \frac{2(1-\alpha)}{\beta_2^2} = m_2 \tag{4.58}$$

and

$$\frac{6\alpha}{\beta_1^3} + \frac{6(1-\alpha)}{\beta_2^3} = m_3 \tag{4.59}$$

By solving (4.57)-(4.59) for α, β_1 and β_2, we find

$$\alpha = \frac{1 - 2m_1 z + m_1^2 z^2}{1 - 2m_1 + m_2 z^2} \tag{4.60}$$

$$\beta_1 = \frac{m_1 z - 1}{m_2 z - m_1} \tag{4.61}$$

and

$$\beta_2 = z \tag{4.62}$$

where

$$z = \frac{(m_3 - m_2 m_1) + \sqrt{(m_3 - m_1 m_2)^2 + 4(m_3 m_1 - m_2^2)(m_1^2 - m_2)}}{2(m_3^2 - m_2^2)} \tag{4.63}$$

Thus, for given m_1, m_2 and m_3 we can determine the parameters α, β_1, and β_2 and the interarrival time distribution for the input process or the two-state MMPP by (4.60)-(4.63). Note that the method introduced in this section is a general way to determine the interarrival time distribution (3.102) for the two-state MMPP.

4.2.2 Performance Analysis of the MMPP/M/1 Queue

Let W be the waiting time, D the delay, N_q the queue length, μ_2 the mean service rate of the server, and $F_A^*(s)$ the Laplace-Stieltjes transform of the interarrival time distribution of the input process. By the G/M/1 model we have the mean waiting time

$$E(W) = \frac{\sigma}{\mu_2(1-\sigma)} \tag{4.64}$$

the mean delay

$$E(D) = \frac{1}{\mu_2(1-\sigma)} \tag{4.65}$$

the mean queue length

$$E(N_q) = \frac{\rho_2 \sigma}{1-\sigma} \tag{4.66}$$

the waiting-time probability distribution

$$F_W(t) = 1 - \sigma e^{-\mu_2(1-\sigma)t} \tag{4.67}$$

and the state probability distribution

$$p_k = \begin{cases} 1-\rho_2 & ,k=0 \\ \rho_2(1-\sigma)\sigma^{k-1} & ,k\geq 1 \end{cases} \tag{4.68}$$

where

$$\rho_2 = \frac{1}{m_1\mu_2} = \frac{\lambda_p\lambda_c}{\mu_1\mu_2} \tag{4.69}$$

$\lambda_p, \lambda_c, \mu_1$ are defined in section 3.4.1, and σ is the root of the equation

$$\sigma = F_A^*(\mu_2 - \mu_2\sigma) \quad , \quad 0 < \sigma < 1 \tag{4.70}$$

Using (3.102) we have

$$F_A^*(s) = \frac{\alpha\beta_1}{s+\beta_1} + \frac{(1-\alpha)\beta_2}{s+\beta_2} \tag{4.71}$$

Then σ satisfies (4.70), or

$$\sigma = \frac{\alpha\beta_1}{\mu_2 - \mu_2\sigma + \beta_1} + \frac{(1-\alpha)\beta_2}{\mu_2 - \mu_2\sigma + \beta_2} \tag{4.72}$$

or

$$a\sigma^3 + b\sigma^2 + c\sigma + d = 0 \tag{4.73}$$

where

$$a = \mu_2 \tag{4.74}$$

$$b = -(2\mu_2^2 + \mu_2\beta_1 + \mu_2\beta_2) \tag{4.75}$$

$$c = \mu_2^2 + 2\mu_2\beta_2 + \beta_1\mu_2 + \alpha\beta_1\mu_2 - \alpha\beta_2\mu_2 + \beta_1\beta_2 \tag{4.76}$$

and

$$d = -(\alpha\beta_1\mu_2 - \alpha\beta_2\mu_2 + \beta_2\mu_2 + \beta_1\beta_2) \tag{4.77}$$

4.2.3 Numerical Results

First, we show the packet interarrival time probability density of the two-state MMPP model with parameters α, β_1 and β_2 obtained by the measurement method and compare them with the simulation results.

In Fig. 4.15 and Fig. 4.16, s equals to 2 and s is the number of servers in the Erlang delay system which generates the DSPP traffic, see section 3.4.1. In Fig. 4.17 and Fig. 4.18 s equals to 4. We see that the measurement method presented in section 4.2.1 can also accurately determine the packet interarrival time distribution for the two-state MMPP model.

Then, we consider the mean delay of a packet in a MMPP/M/1 queue. In the following figures, for the curves of MMPP/M/1 I we use the method by Heffes [12] in section 3.4.3 to determine parameters $\lambda_1, \lambda_2, r_1$ and r_2 for the two-state MMPP model, and for the curves of MMPP/M/1 II we use the measurement method to obtain parameters α, β_1 and β_2 for the two-state MMPP model, and $\tau_2 = 1/\mu_2$ is equal to 2.5ms.

Fig. 4.19 and Fig. 4.20 show the mean delay as a function of the traffic intensity $\rho_2 = (\lambda_c \lambda_p)/(\mu_1 \mu_2)$ for s=2 and s=4 respectively. It can be seen that the results of MMPP/M/1 II are more closer to the simulation results than the results of MMPP/M/1 I.

The relation of the mean delay and the mean number of packets per call, λ_p/μ_1, is shown in Fig. 4.21 and Fig. 4.22 for s=2 and s=4, respectively. In both figures, $\rho_2 = 0.8$. We find that the mean delay increases slightly as λ_c/μ_1 increases even though ρ_2 is unchanged, and the simulation results are more sensitive to the value of λ_c/μ_1 than theoretical results. It indicates that the mean delay depends not only on

ρ_2 but also on λ_p/μ_1.

The compound effects of ρ_2 and λ_p/μ_1 on the mean delay are shown in Fig. 4.23 and Fig. 4.24. In both figures, ρ_2 increases as λ_p/μ_1 increases.

Last, we present the curves of the waiting time probability density of the MMPP/M/1 queue in Fig. 4.25 - Fig. 4.28. We see that the results of MMPP/M/1 queue can match the simulation results very well.

The results in this section show that our approximation of the DSPP traffic with the MMPP in section 3.4 and our analyses on the performances of MMPP/M/1 queues in section 4.2 are accurate. Comparisons of results of MMPP/M/1 I and MMPP/M/1 II show that the method to determine the parameters α, β_1 and β_2 in section 4.2.1 is more accurate than the method by Heffes. It means that if we choose the parameters for the MMPP model more accurately, we can obtain better results for the performance analysis of the queues.

4.3 SUMMARY

In this chapter we have obtained several performance analysis results for the $M^X/G/1$, $G^X/M/1$ and $MMPP/M/1$ queues. We derive the mean waiting time, mean delay and mean queue size for these queues. We use the maximum entropy method to obtain the ME solutions for the waiting time distribution and the state probability distribution for the $M^X/G/1$ queue. We apply the $G/M/1$ model to the $G^X/M/1$ queue and the MMPP/M/1 queue to obtain the waiting time distributions and the state probability distributions. We show that analytical and simulation results agree closely.

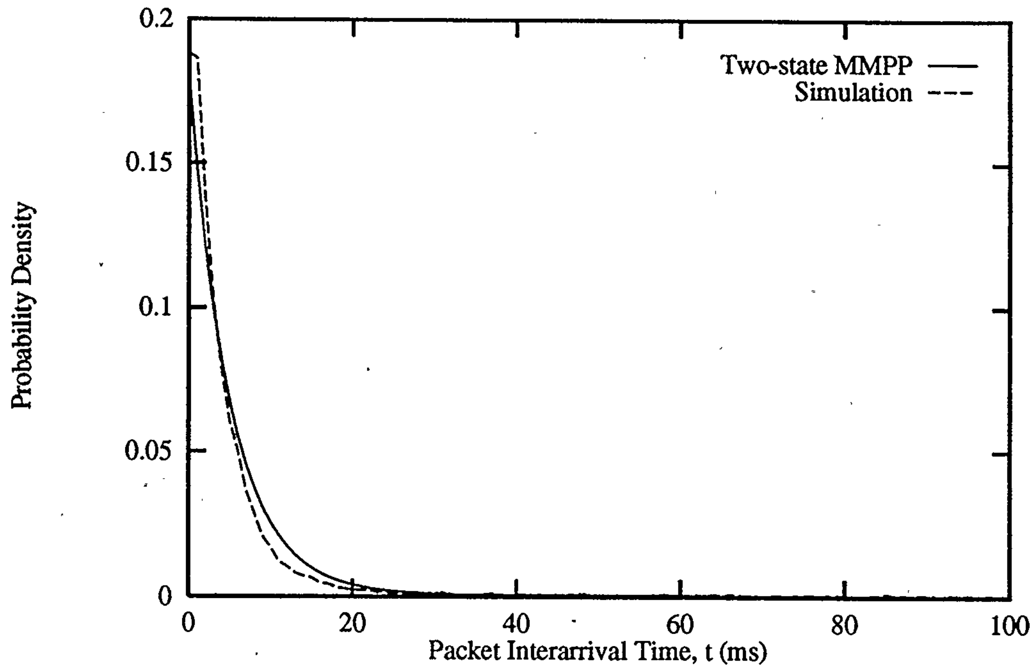

Figure 4.15. Packet Interarrival Time Probability Density, $s = 2, a_1 = 0.25, t_p = 5ms$

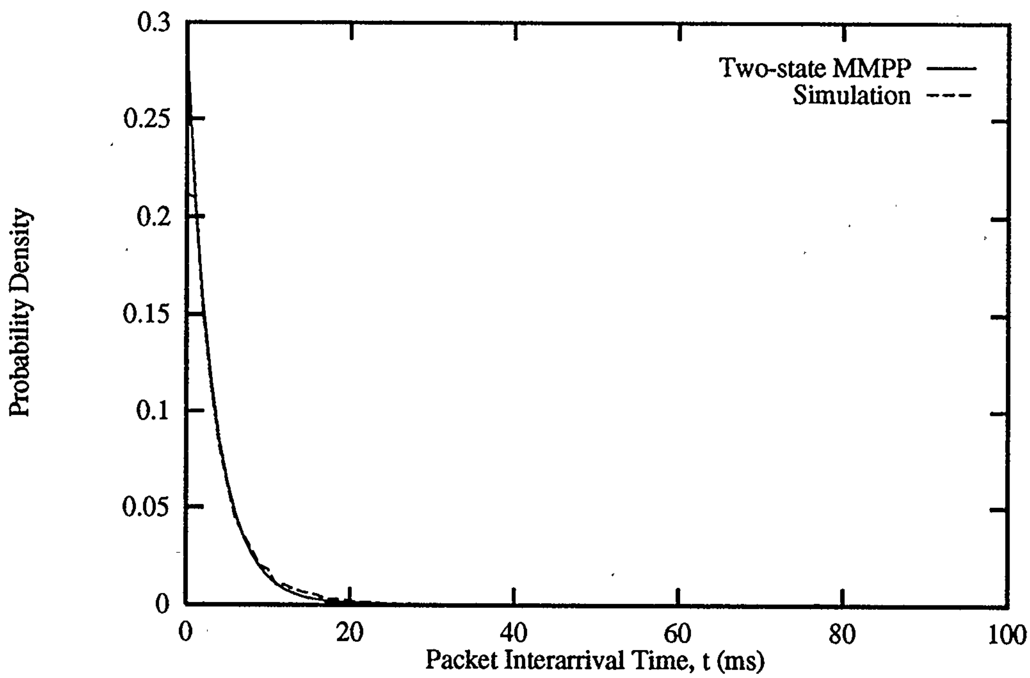

Figure 4.16. Packet Interarrival Time Probability Density, $s = 2, a_1 = 0.5, t_p = 5ms$

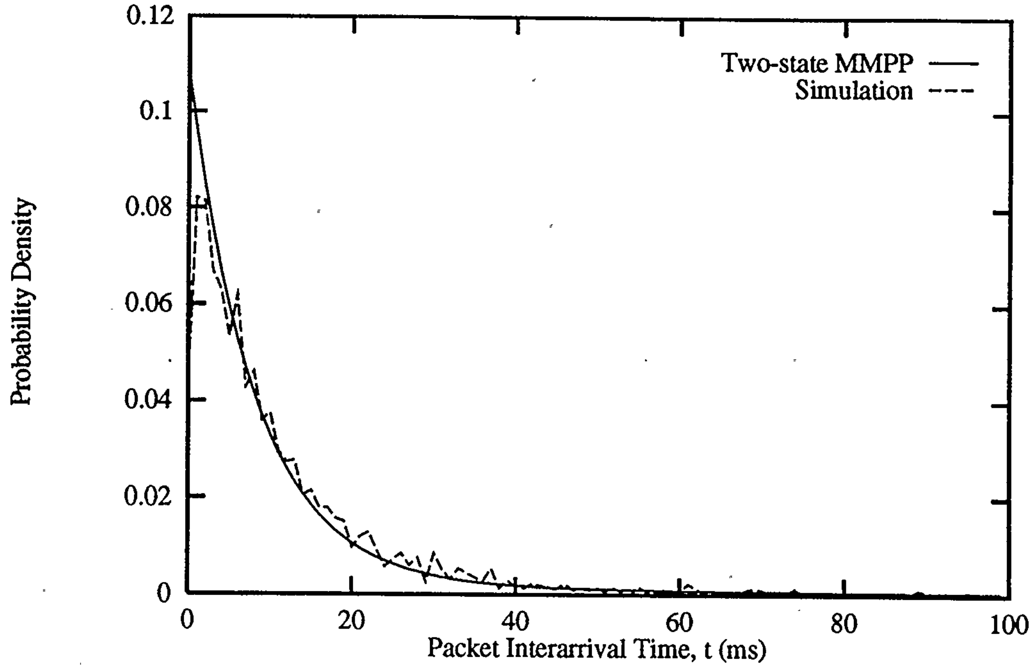

Figure 4.17. Packet Interarrival Time Probability Density, $s = 4, a_1 = 0.25, t_p = 5ms$

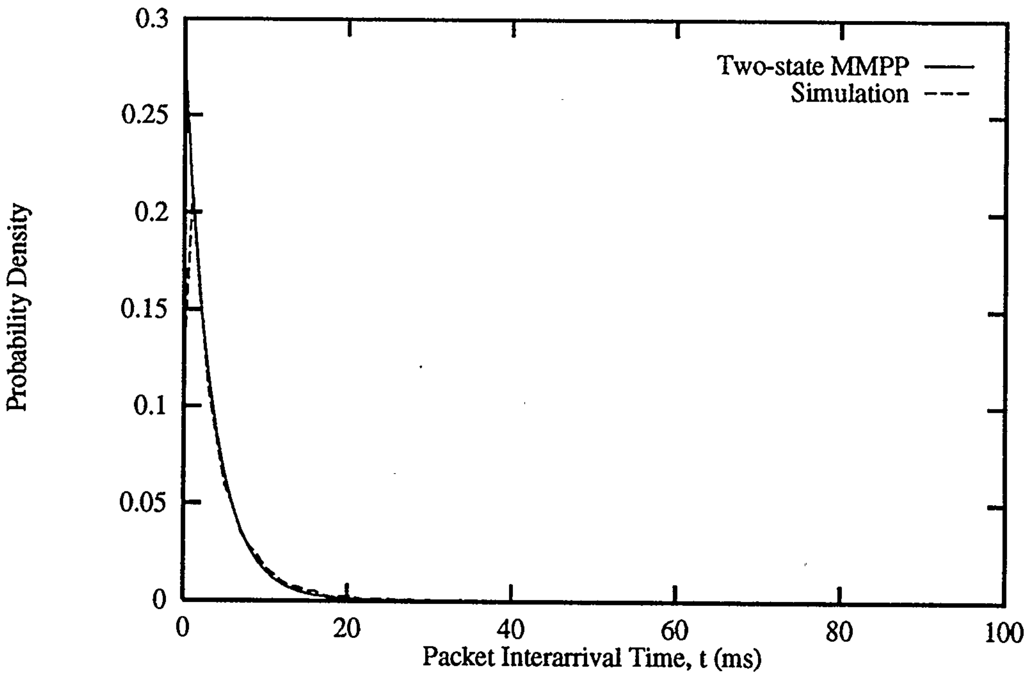

Figure 4.18. Packet Interarrival Time Probability Density, $s = 4, a_1 = 0.5, t_p = 5ms$

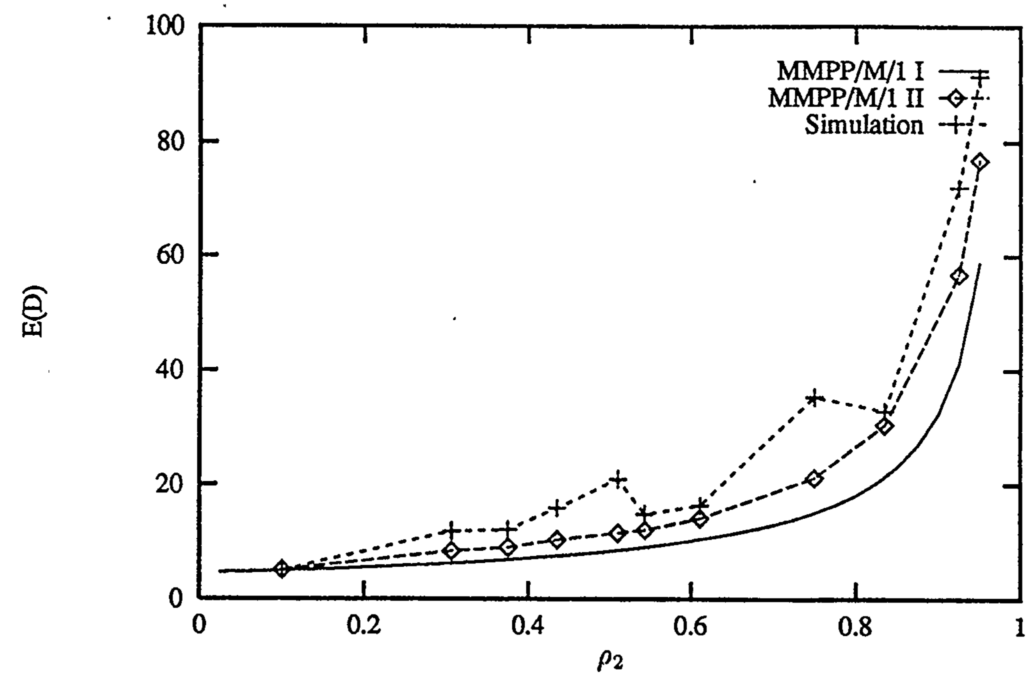

Figure 4.19. Mean Delay, $s = 2, \lambda_p/\mu_1 = 12$

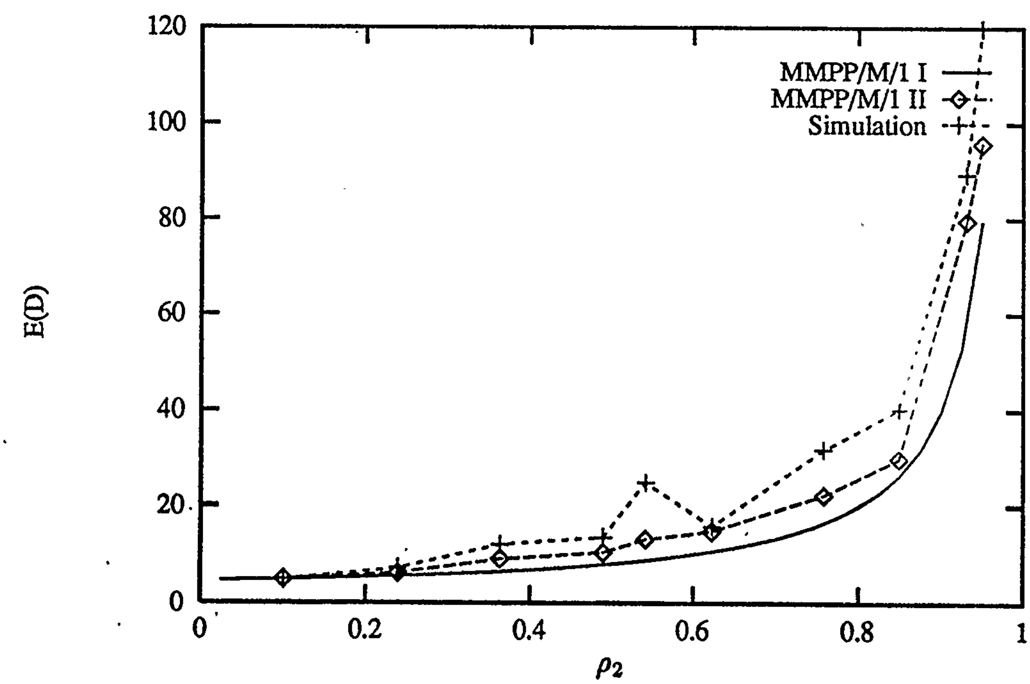

Figure 4.20. Mean Delay, $s = 4, \lambda_p/\mu_1 = 12$

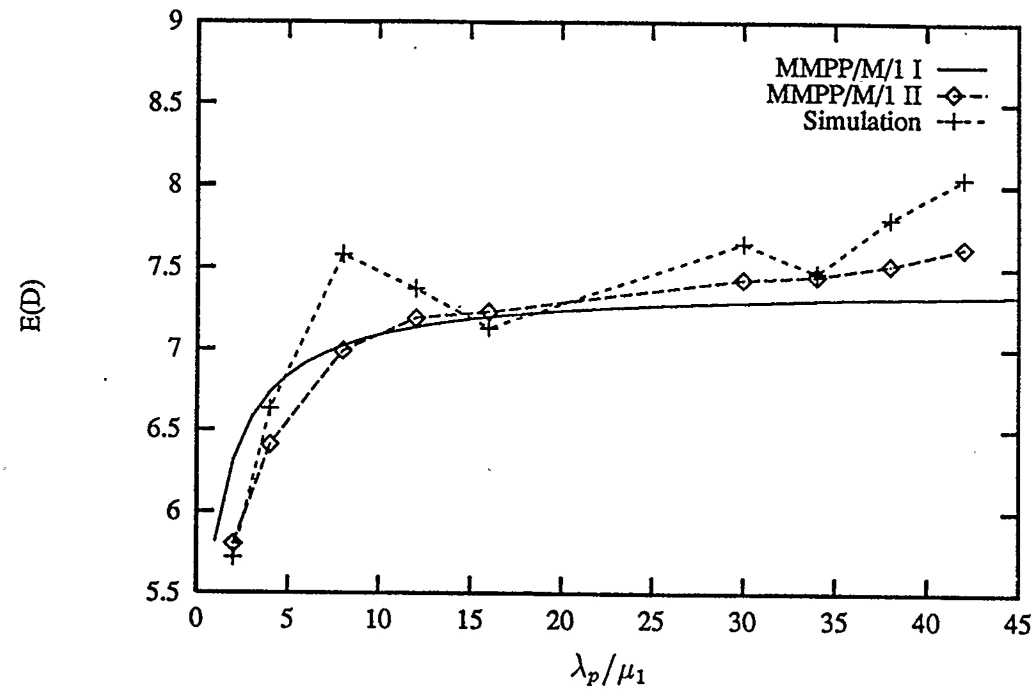

Figure 4.21. Mean Delay, $s = 2, a_1 = 0.8$

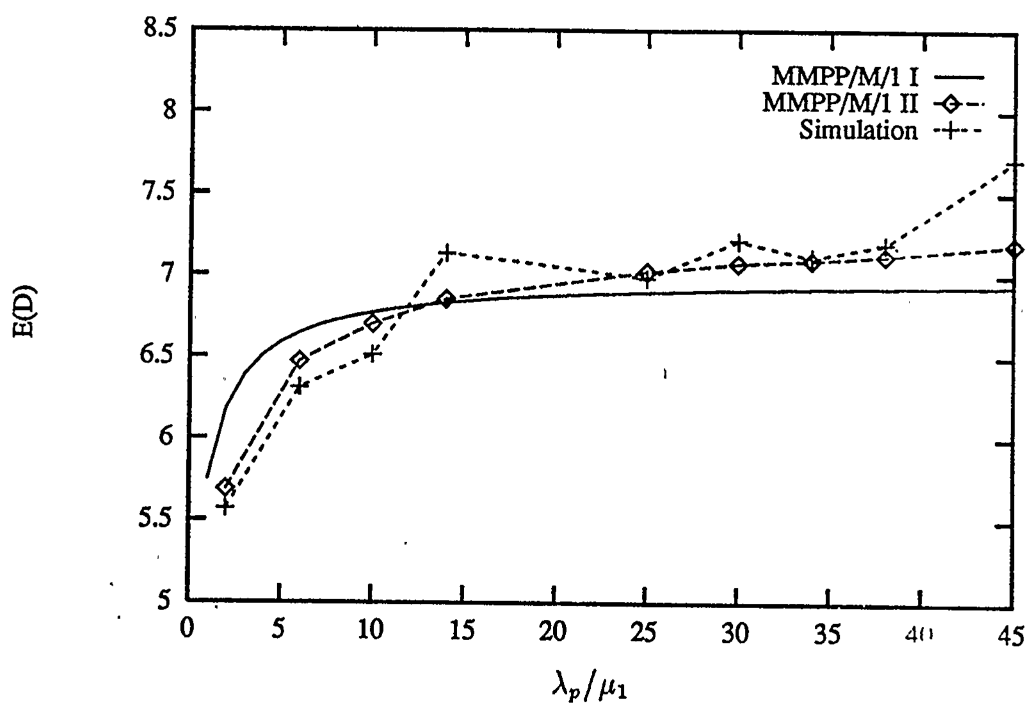

Figure 4.22. Mean Delay, $s = 4, a_1 = 0.8$

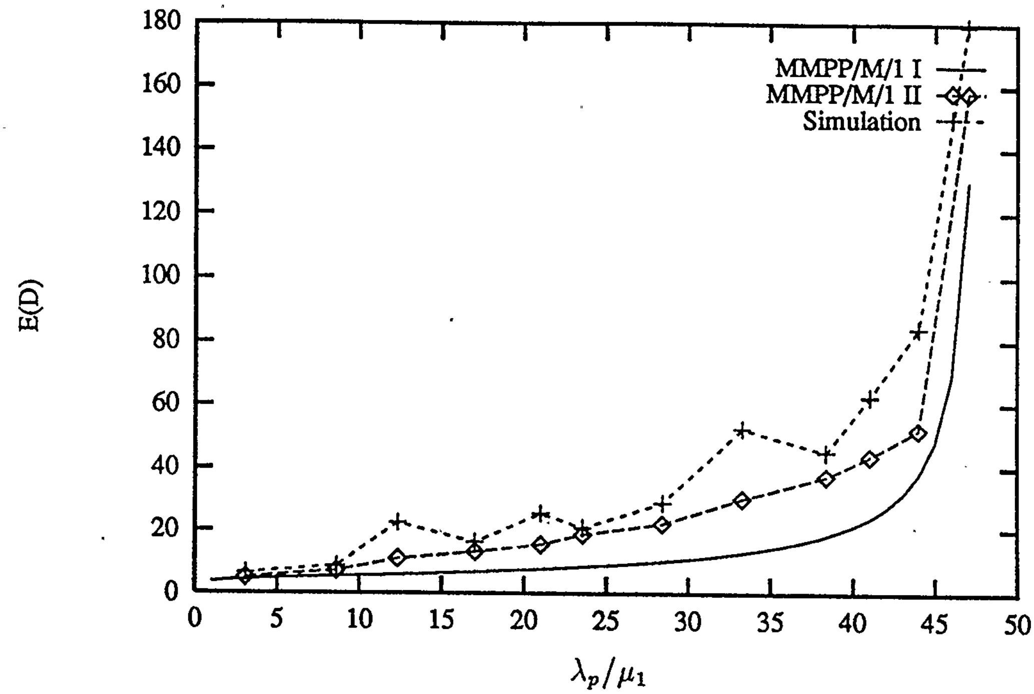

Figure 4.23. Mean Delay, $s = 2, t_c = 120ms, t_p = 5ms$

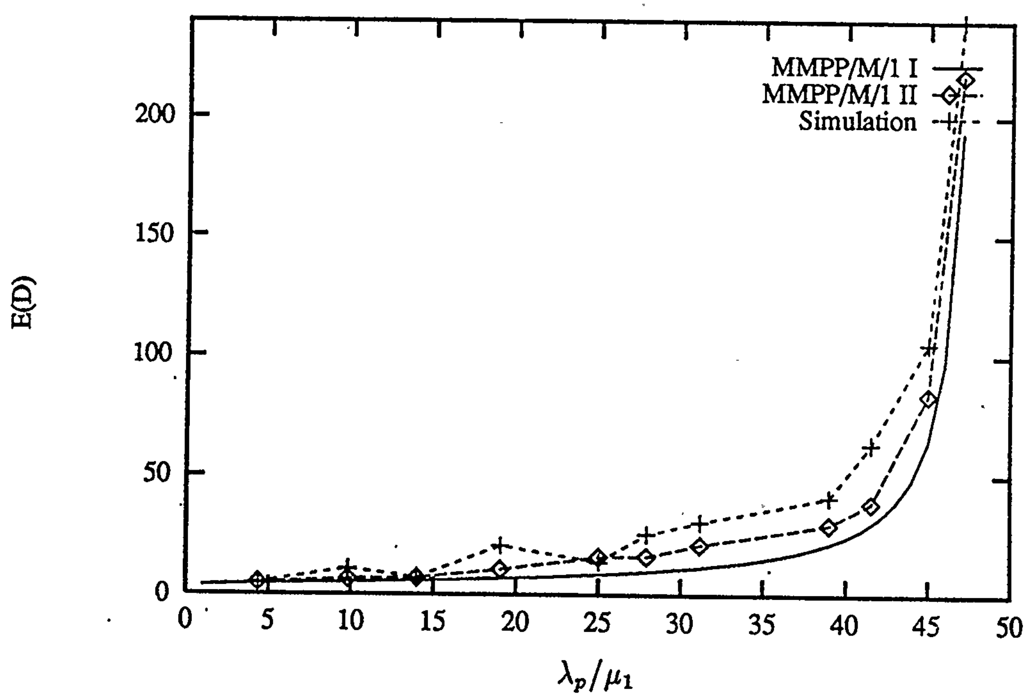

Figure 4.24. Mean Delay, $s = 4, t_c = 120ms, t_p = 5ms$

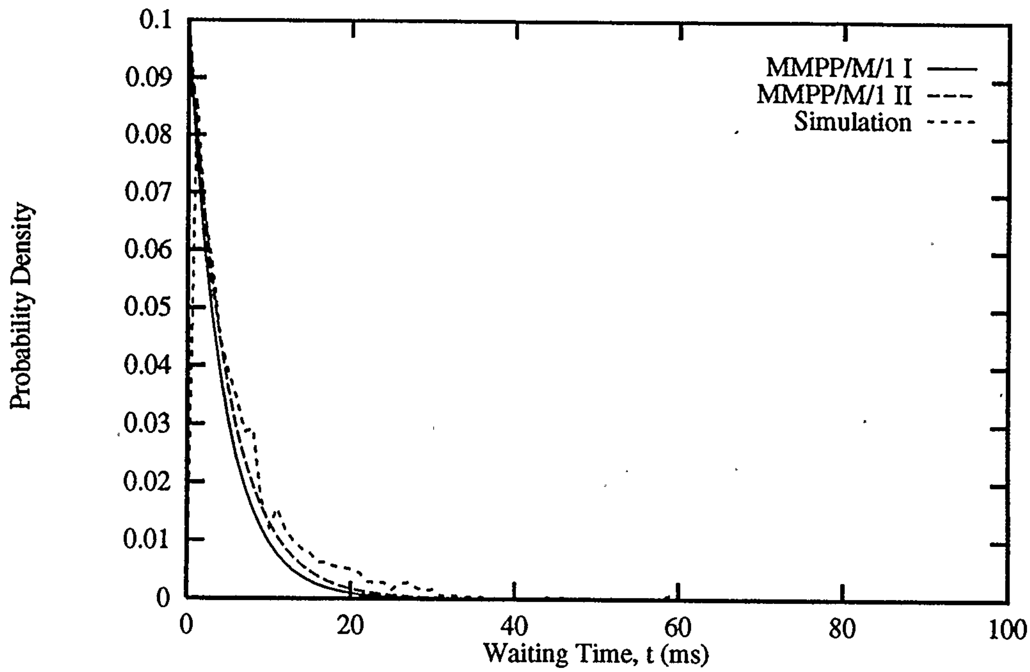

Figure 4.25. Waiting Time Probability Density, $s = 2, t_c = 120ms, \lambda_p/\mu_1 = 6$

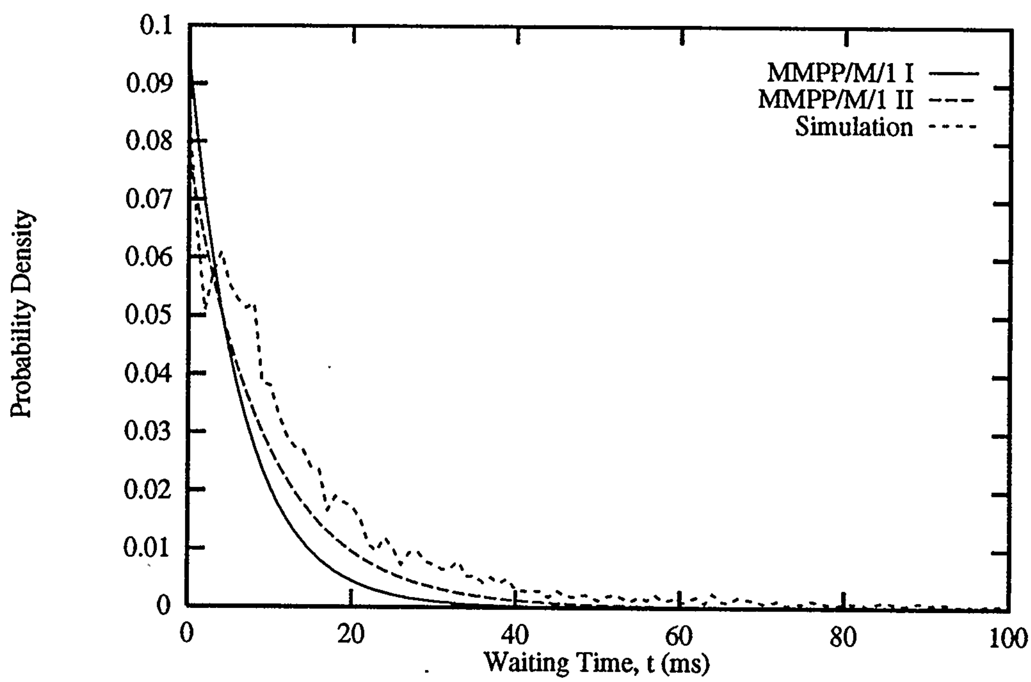

Figure 4.26. Waiting Time Probability Density, $s = 2, t_c = 120ms, \lambda_p/\mu_1 = 12$

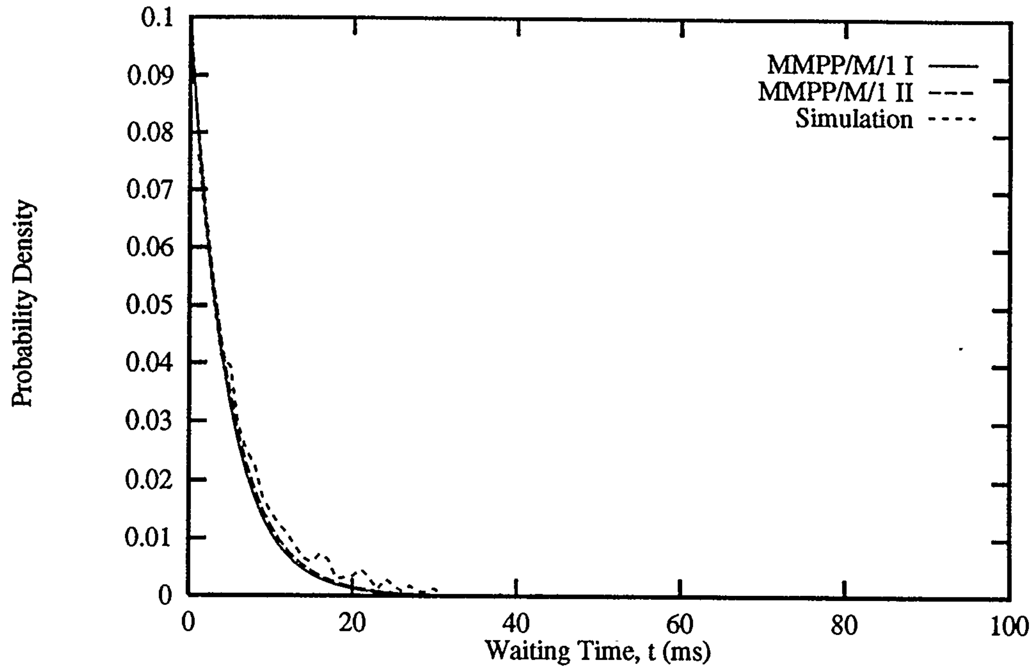

Figure 4.27. Waiting Time Probability Density, $s = 4, t_c = 120ms, \lambda_p/\mu_1 = 6$

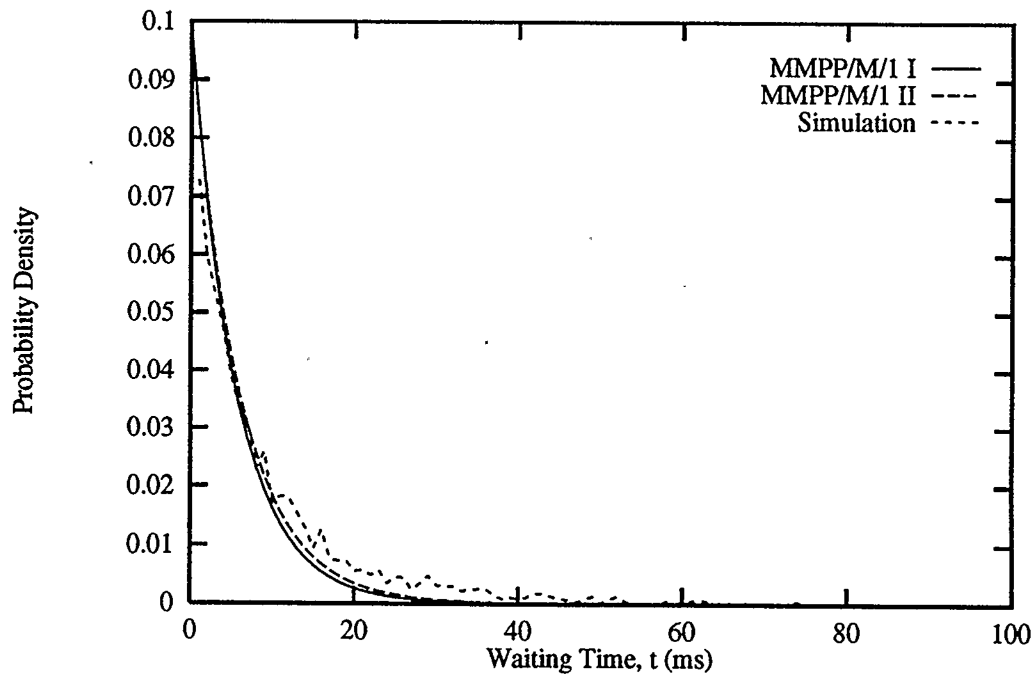

Figure 4.28. Waiting Time Probability Density, $s = 4, t_c = 120ms, \lambda_p/\mu_1 = 12$

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

In this thesis we have presented studies for modeling of input traffic and performance analysis of queueing systems with bursty input based on the principle of maximum entropy and on results of queueing theory.

We have applied the principles of maximum entropy and the minimum cross entropy to determine the equilibrium state probability distributions for several single server queues and multiserver queues. For single server queues, we use maximum entropy solutions of the G/G/1 queue obtained by Kouvatsos to calculate the distributions for the M/G/1 queue and G/M/1 queue. For multiserver queues, we employe the method of cross entropy minimization with the estimate factor of the distribution introduced to derive for the first time the state probability distributions for the Erlang loss system and the Erlang delay system, respectively. By showing that the well-known results from queueing theory can be obtained by the principle of maximum entropy we are led to the conclusion that the maximum entropy formalism can provide a framework for analysis of queueing systems.

We have investigated the characteristics of two typical inputs in packet networks and developed equivalent arrival processes for them. We establish mathematical models for their associated interarrival time distributions.

For the batch data arrival process, we derived two interarrival time distributions by using the method of entropy maximization. These distributions yield useful models for the batch input processes in which mean values of interarrival time and message length are taken into consideration in one distribution. In the first interarrival time distribution, we used the first moment of interarrival time as a constraint. The distribution turns out to be a generalized exponential type. In the second interarrival time distribution, we used both the first and second moments of the interarrival time as constraints. The result is a normal-like distribution. Comparisons of simulation and theoretical results indicate that the first distribution can yield exact results for the batch processes with Poisson message arrivals and, if the message process is not Poisson, the second distribution yields a better approximation.

For the doubly stochastic Poisson process, we used a two-state Markov modulated Poisson process as an approximation. We derive an interarrival time distribution for the two-state MMPP and found that the distribution is hyperexponential determined by the four parameters of the two-state MMPP. We also investigate the statistical properties of the traffic, such as the burstiness of the traffic characterized by the coefficient of variation of interarrival time. Numerical results show that the two-state MMPP is a good approximation for the doubly stochastic Poisson input process.

Following that part, we carry out performance analysis of queueing systems by applying the mathematical models established for the batch arrival processes and the doubly stochastic Poisson input processes and by using the method of entropy maximization and the G/M/1 queue results from queueing theory. We determine the mean delay, the mean queue length, the waiting time distribution and the state probability distribution for the queues. The results are compared favorably with simulations. These results show good accuracy of our approximations of the input traffic and the approaches we employed in the performance analysis of the queueing systems.

Before we finish the whole thesis, we shall make some suggestions for the problems remaining and for future studies.

When applying the method of entropy maximization to performance analysis of queueing systems, such as determining the system state probability distribution, the choice of proper prior information as constraints is of great significance. One of the problems is with what kind of minimum prior information can we determine the exact distribution for the queueing systems, and under what kind of constraints we can obtain good approximate results as required.

In the modeling of input traffic, the Markov modulated Poisson process is a useful model that can represent traffic with the bursty and correlated characteristics involved and which can represent particularly aggregate traffic generated by the superposition of several point processes. One problem existing in the applications of the MMPP is how to fit an MMPP model to the arrival processes, i.e. how to choose the parameters of the MMPP from the original arrival processes.

To obtain an analytical solution for performance analysis of a queueing system with non-Poisson input and general service time distribution is very difficult. In addition to the method of entropy maximization, other efficient approaches which can provide analytically practical solutions for such systems remain to be developed.

REFERENCES

[1] J. F. Shoch and J. A. Hupp, "Measured performance of an Ethernet local Network," *Communications of the ACM*, vol. 23, No. 12, pp.711-720, 1980.

[2] P. F. Pawlita, "Two decades of data traffic measurements: A survey of published results, experiences and applicability," in *Proc. 12th Int. Teletraffic Congress*, Torino, Italy, June 1988.

[3] R. Gusella, "A measurement study of diskless workstation traffic on an Ethernet," *IEEE Trans. Commun.*, vol. 38, no.9, pp.1557-1568, Sept. 1990.

[4] P. J. Burke, "Delays in single-server queues with batch input," *Opns. Res.*, vol. 23, pp. 830-833, 1975.

[5] J. W. Cohen, "On a single server queue with group arrivals," *J. Appl. Prob.*, vol. 13, pp. 619-622, 1976.

[6] M. L. Chaudhry and J. G. C. Templeton, *A first course in bulk queues*, John Wiley, New York, 1983.

[7] D. D. Yao, "Analyzing the steady-state $GI^X/G/1$ queue," *J. Opl. Res. Soc.,* vol. 35, no. 11, pp. 1027-1030, 1984.

[8] A. M. Eikeboom and H. C. Tijms, "Waiting time percentiles in the multiserver $M^X/G/1$ queue with batch arrivals," *Prob. Eng. Inform. Sci.*, vol. 1, pp. 75-96, 1987.

[9] J. C. W. Van Ommeren, "Exponential expansion for the tail of the waiting time probability in the single server queue with batch arrivals," *Adv. Appl. Prob.*, vol. 20, pp. 880-895, 1988.

[10] J.-S. Wu, "Maximum entropy analysis of queueing systems and networks," Ph.D dissertation, Univ. Calgary, Calgary, Nov. 1988.

[11] J. C. W. Van Ommeren, "Simple approximations for the batch-arrival $M^X/G/1$ queue," *Opns. Res.*, vol. 38, pp. 678-685, 1989.

[12] H. Heffes, "A class of data traffic processes–covariance function characterization and related queueing results," *Bell Syst. Tech. J.*, vol. 59, No. 6, pp. 897-929, 1980.

[13] D. Y. Burman and D. R. Smith, "Asymptotic Analysis of a queueing model with bursty traffic," *Bell Syst. Tech. J.*, vol. 62, No.6, pp. 1433-1453, 1982.

[14] S. Q. Li, "Study of information loss in packet voice systems," *IEEE Trans. Commun.*, vol. 37, pp. 1192-1202, Nov. 1989.

[15] J. N. Daigle and J. D. Langford, "Models for analysis of packet voice communications systems," *IEEE J. Select. Areas Commun.*, vol. SAC-4, pp. 847-855, Sept. 1986.

[16] I. Ide, "Superposition of interrupted Poisson process and its application to packetized voice multiplexers," in *Proc. 12th Int. Teletraffic Congress*, Torino, Italy, June 1988.

[17] C. J. Weinstein, "Fractional speech loss and talker activity model for TASI and for packet-switched speech," *IEEE Trans. Commun.*, vol. 26, pp. 1253-1257, Sept. 1978.

[18] T. E. Stern, "A queueing analysis of packet voice," in Conf. Rec. 1983 IEEE Global Telecommun. Conf., vol. 1, San Diego, CA, 1983, pp. 71-76.

[19] B. G. Kim, "Characterization of arrival statistics of Multiplexed Voice Packets," *IEEE J. Select. Areas Commun.*, vol. 1, pp. 1133-1139, Dec. 1983.

[20] J. N. Daigle and J. D. Langford, "Queueing analysis of a packet voice communication system," *Conf. Rec. IEEE INFOCOM'85*, Washington, DC, pp. 18-26, Mar. 1985.

[21] H. Heffes and D. M. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE J. Select. Areas Commun.*, vol. SAC-4, pp. 856-686, Sept. 1986.

[22] K. Q. Liao and L. G. Mason, "A discrete-time single server queue with a two-level modulated input and its applications," in *Proc. IEEE GLOBECOM '89*, pp. 26.1.1-26.1.6.

[23] M. Nomura, T. Fujii, and N. Otha, "Basic characteristics of variable rate video coding in ATM environment," *IEEE J. Select. Areas Commun.*, vol. 7, pp. 865-869, June 1989.

[24] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. D. Robbins, "Performance models of statistical multiplexing in packet video communications," *IEEE Trans. Commun.*, vol. 36, pp. 834-844, July 1988.

[25] P. Sen, B. Maglaris, N. E. Rikli, and D. Anastassiou, "Models for packet switching of variable-bit-rate video sources," *IEEE J. Select. Areas Commun.*, vol. 7, pp. 865-869, June 1989.

[26] Y. Yasuda, H. Yasuda, N. Otha, and F. Kishino, "Packet video transmission through ATM networks," in *Proc. IEEE GLOBECOM '89*, pp. 25.1.1-25.1.5.

[27] K. Sriram, P. K. Varshney and J. G. Shanthikumar, "Discrete-time analysis of integrated voice/data multiplexers with and without speech activity detectors," *IEEE J. Select. Areas Commun.*, vol. 1, No. 6, pp. 1124-1132, 1983.

[28] S. Q. Li and J. W. Mark, "Performance of voice/data integration on a TDM system," *IEEE Trans. Commun.*, vol. 33, No. 12, pp. 1265-1273, 1985.

[29] K. Sriram and W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data," *IEEE J. Select. Areas Commun.*, vol. 4, No. 6, pp. 833-846, 1986.

[30] B. N. W. Ma and J. W. Mark, "Performance analysis of burst switching for integrated voice/data service," *IEEE Trans. Commun.*, vol. 36, No. 3, 1988.

[31] S. Q. Li and J. W. Mark, "Traffic characterization for integrated services," *Proc. INFOCOM'88*, New Orleans, LA, Mar. 29-31, pp. 8A.3.1-8A.3.10, 1988.

[32] M. Reiser and H. Kobayashi, "Accuracy of the diffusion approximation for some queueing systems," *IBM J. Res. Devl.* 18, pp. 180-124, 1974.

[33] W. G. Marchal, "An approximation formula for waiting time in single server queues. *AIEE Trans.* 8, 473, 1976.

[34] W. Kramer and M. L-Belz, "Approximate formulae for the delay in the queueing system GI/G/1," in *Proc. 8th Int. Teletraffic Congress*, Melbourne, pp 235.1-235.8, 1976.

[35] E. Gelenbe and J. Mitrani, *Analysis and Synthesis of Computer Systems*, Academic Press, London, 1980.

[36] E. Koenigsberg, "Twenty five years of cyclic queues and closed queue networks: a review," *J. Opl Res. Soc.*, vol. 33, pp. 605-619, 1982.

[37] W. Whitt, "On approximations for queues, III: Mixtures of exponential distributions," *AT&T Bell Labs Tech. J.*, Vol. 63, pp. 163-175, 1984.

[38] C. H. Sauer and K. M. Chandy, *Computer systems performance modelling*, Prentice- Hall, Englewood Cliffs, 1981.

[39] J. E. Shore and R. W. Johnson, "Axiomatic Derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inf. Theory*, vol. 26, pp. 26-37, 1980.

[40] E. T. Jaynes, "Information theory and statistical mechanics I," *Phys. Rev.*, vol. 106, pp. 620-630, 1957.

[41] W. M. Elsasser, "On quantum measurements and the role of the uncertainty relations in statistical mechanics," *Phys. Rev.*, vol. 52, pp. 987-999, 1937.

[42] E. T. Jaynes, "Information theory and statistical mechanics," in *Statistical Physics*, vol. 3, Brandeis Lectures, K. W. Ford, Ed. New York, Benjamin, pp. 182-218, 1963.

[43] O. C. de Beauregard and M. Tribus, "Information theory and thermodynamics," *Helvetica Physica Acta*, vol. 47, pp. 238-247, 1974.

[44] M. Tribus, *Rational Descriptions, Decision, and Designs*, New York: Pergamon, 1969.

[45] A. E. Ferdinand, "A theory of general complexity," *Int. J. General Syst*, vol. 1, pp. 19-33, 1974.

[46] J. G. Albes, "Maximum entropy spectral analysis," *Astron. Astrophys. Suppl.*, vol. 15, pp. 383-393, 1974.

[47] E. T. Jaynes, "Prior Probabilities," *IEEE Trans. Syst. Sci. Cyb.*, vol. 4, pp. 227-241, 1968.

[48] M. Chan, "System simulation and maximum entropy," *Operations Research*, vol. 19, pp. 1751-1753, 1971.

[49] A. E. Ferdinand, "A statistical mechanics approach to systems analysis," *IBM J. Res. Develop.*, vol. 14, pp. 539-547, 1970.

[50] Y. Bard, "Estimation of state probabilities using the maximum entropy principle," *IBM J. Res. Develop.*, vol. 24, pp.563-569, 1980.

[51] J. E. Shore, "Derivation of equilibrium and time-dependent solutions to $M/M/\infty/N$ and $M/M/\infty$ queueing systems using entropy maximization," in *AFIPS Conf. Proc.*, vol. 47, pp. 483-487, 1978.

[52] J. E. Shore, "Information theoretic approximations for M/G/1 and G/G/1 queueing systems," *Acta Info.*, vol. 17, pp. 43-61, 1982.

[53] M. A. El-Affendi and D. D. Kouvatsos, "A maximum entropy analysis of the M/G/1 and G/M/1 queueing systems at equilibrium," *Acta Info.*, vol. 19, pp. 339-355, 1983.

[54] S. Guiasu, "Maximum entropy condition in queueing theory," *J. Opl Res. Soc.*, vol. 37, pp. 293-301, 1986.

[55] D. D. Kouvatsos, "A maximum entropy queue length distribution for the G/G/1 finite capacity queue," *Proc. of Perf. '86 and ACM Sigmetrics 1986*, pp. 224-236.

[56] D. D. Kouvatsos, " A maximum entropy analysis of the G/G/1 queue at equilibrum," *J. Opl Res. Soc.*, vol. 39, No. 2, pp. 183-200, 1988.

[57] D. D. Kouvatsos, "Maximum entropy methods for general queueing networks," In *Modelling Techniques and Tools for Performance Analysis*(D. Potier, Ed.), North-Holland, Amsterdam, pp. 589-608, 1985.

[58] D. D. Kouvatsos and J. Almond, "Maximum entropy two-station cyclic queues with multiple general servers," *Acta Info.*, vol. 26, pp. 241-267, 1988.

[59] D. D. Kouvatsos and A. T. Othman, "Optimal flow control of a G/G/c finite capacity queue," *J. Opl Res. Soc.*, vol. 40, No. 7, pp. 659-670, 1989.

[60] J. S. Wu and W. C. Chan, "Maximum entropy analysis of multiple-server queueing systems," *J. Opl Res. Soc.*, vol. 40, No. 9, pp. 815-825, 1989.

[61] L. Kleinrock, *Queueing Systems: Theory*, vol.1, Wiley, New York, 1975.

[62] S. S. Lavenberg, *Computer Performance Modeling Handbook*, Academic Press, 1983.

[63] J. L. Hammond and P. J. P. O'Reilly, *Performance analysis of local computer networks*, Addison-Wesley, 1986.

[64] Y. Takahashi, "Asymptotic exponentiality of the tail of the waiting-time distribution in a PH/PH/c queue," *Adv. Appl. Prob.*, vol. 13, pp. 619-630, 1981.

[65] R. Gusella, "Characterizing the variability of arrival processes with indexes of dispersion," *IEEE J. Select. Areas Commun.*, vol. 9, No. 2, pp.203-211, 1981.

[66] W. Whitt, "Approximating a point process by a renewal process: Two basic methods," *Oper. Res.*, vol. 30, no. 1, pp. 125-147, 1982.

[67] W. Whitt, "The queueing network analyzer," *Bell Syst. Tech. J.*, Part 1, vol. 62, no. 9, pp. 2779-2815, 1983.

[68] S. L. Albin, "Approximating a point process by a renewal process, II: Superposition arrival processes to queues," *Oper. Res.*, vol. 32, pp. 1133-1162, 1984.

[69] V. E. Benes, *Mathematical theory of connecting networks and telephone traffic,* Academic Press, New York, 1965.

[70] D. R. Cox and P. A. W. Lewis, *The statistical analysis of series of events,* London: Chapman and Hall, 1966.

[71] M. F. Neuts, "A versatile Markovian point process," *J. Appl. Prob.,* vol. 16, pp. 764-779, 1979.

[72] M. F. Neuts, "The M/M/1 queue with randomly varying rates," *Opsearch,* vol. 15, no. 4, pp. 158-168, 1978.

[73] A. Kuczura, "Queues with mixed renewal and Poisson inputs," *B.S.T.J.,* vol. 51, no. 6, pp. 1305-1326, 1972.