THE UNIVERSITY OF CALGARY

MAXIMUM ENTROPY ANALYSIS OF QUEUEING SYSTEMS AND NETWORKS

by

Jungshyr Wu

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF ELECTRICAL ENGINEERING

CALGARY, ALBERTA

November, 1988

THE UNIVERSITY OF CALGARY

MAXIMUM ENTROPY ANALYSIS OF QUEUEING SYSTEMS AND NETWORKS

by

Jungshyr Wu

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF ELECTRICAL ENGINEERING

CALGARY, ALBERTA

November, 1988

ISBN 0-315-50431-5

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, "Maximum Entropy Analysis of Queueing Systems and Networks", submitted by Jungshyr Wu in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dr. W. C. Chan, Supervisor
(Department of Electrical Engineering)

Dr. D. Irvine-Halliday
(Department of Electrical Engineering)

Dr. S. T. Nichols
(Department of Electrical Engineering)

Dr. S. C. Wirasinghe
(Department of Civil Engineering)

Dr. L. Lee, External Examiner
(Alberta Government Telephones)

Date: November 28, 1988

ABSTRACT

A study of queueing systems by means of entropy maximization method is presented. For general interarrival time and service time distributions, results for state probability distribution and waiting time distribution are obtained. For bulk-input queueing systems, equivalent arrival processes are obtained both in continuous time and in discrete time and then applied to study some single queueing systems. In extension to queueing networks, the networks with single-server nodes and general service times, multiple exponential-server nodes, and single-server nodes with finite buffer capacities are examined. Expressions for the determination of queue characteristics are derived. Numerical examples are included for illustration. In the application to computer communication networks, multistage interconnection networks are studied by taking the bulk-input into account. The queueing network models with window control are also investigated. Results for the throughput, utilizations, and mean number of packets at each node are obtained and compared with simulations. Finally, some problems for future studies are discussed.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to Dr. Wah-Chun Chan for his invaluable guidance and encouragement as well as precious suggestions during the course of this work.

I am also grateful to Dr. D. Irvine-Halliday, Dr. L. Lee, Dr. S. Nichols and Dr. S.C. Wirasinghe for their valuable advices and assistance.

Financial assistance received from the Department of Electrical Engineering in the form of graduate assistantship, and support from the Natural Sciences and Engineering Research Council of Canada are gratefully acknowledged.

To the Glory of

Lord Jesus

His guidance, wisdom, and grace

He abundantly gives.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF IMPORTANT SYMBOLS

N	number of customer in system
T	interarrival time
Y	service time
B	group size
W	waiting time
I	idle time
D	interdeparture time
C	number of servers
ρ	traffic intensity
H	entropy
r_{ij}	transition probability from node i to node j
E(X)	mean of random variable X
Var(X)	variance of X
c_X	coefficient of variation of X
$f_X(t)$	probability density function of X
$F_X(t)$	distribution function of X
$f_X^*(s)$	Laplace transform of $f_X(t)$
$F_X^*(s)$	Laplace-Stieltjes transform of $F_X(t)$

$G_X(z)$	generating function of X

QUEUEING NOTATIONS

A/B/C/m	C-server queue with interarrival time distribution and service time distribution identified by A and B, respectively, with capacity of queue m-1 (if the last descriptor is missing it is assumed to be infinite)
G	general interarrival time or service time distribution
M	negative exponential distribution
GE	generalized exponential distribution
E_2	2-stage Erlang distribution
D	constant distribution
Geom	geometric distribution
QGeom	quasi-geometric distribution
M^X	bulk-Poisson input

CHAPTER 1

INTRODUCTION

1.1 PROBLEMS OF COMPUTER COMMUNICATIONS NETWORK

The main function of computer communication networks is to provide a communication path between user devices connected to the network. The nature of the traffic offered by the user devices is a major factor in determining the performance of the network. Therefore, it is important to know whether the distributions of inter- message arrival process and the message lengths can be adequately modeled as exponential distributions. In the study of computer communications, recent results from experimental measurements of existing networks [1,2,3] show that the real network traffic is bursty. The interarrival time distribution of messages can be reasonably approximated by an exponential distribution. However, message length distribution appears to have a bimodal character with a few discrete message lengths in each portion of the distribution. 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 determining a more accurate input process for the performance analysis of computer communication networks.

A common approach used in previous analyses of queueing systems and queueing networks was based on independent message interarrival time and message length distribution. In the present work an equivalent packet interarrival

time distribution has been determined. The significances of the packet interarrival time distribution to the network input traffic problems are twofold : it gives a generalized model for the input traffic in which the mean values of message interarrival time and message length are taken into consideration in one distribution. Furthermore, it enables us to compare with the actual network input traffic.

Contention for resources in a communication network can be modeled as a network of queues, each consisting of a collection of service stations. Important quantities of analyses are the number of customers at each service station, the mean queue length, throughput rate, and the average waiting time. These quantities form a basis for assessing the functional effectiveness of the network.

The state of a queueing network can be modeled as a vector Markov process, each element representing the state of an individual queue. Under statistical equilibrium conditions the state of the queueing network is characterized by a unique probability distribution from which various performance measures of interest can be obtained.

In principle, the probability distribution of state of a queueing network can be obtained from a set of conditions which specify the equilibrium state of the system. These conditions are called the global balance conditions and take the form of a system of linear equations of the state probabilities. Numerical methods may be used to solve these global equations. However, as the size of the state space of the network increases, numerical methods for solving the global balance equations may

become formidable. To overcome this difficulty, the method of network decomposition has been used. Approximations to the distribution of customers at the service stations as well as the departure processes have been employed. These approximations have simplified the analysis and the results obtained have been compared favorably with simulation results.

1.2 RESEARCH APPROACH

The computer system analyst and the traffic engineer are often faced with the analysis of complex systems that may be best described by statistical methods. In statistical mechanics the physicist also deals with complex systems consisting of many particles In all these situations statistical techniques have led to very satisfactory results. In this thesis a statistical method using entropy maximization will be developed for the study of queueing problems. A function similar to the partition function of statistical mechanics will be employed under the condition of statistical equilibrium. A distribution of certain random variable throughout the range of permissible states will be determined. Under certain assumptions with regard to state independence a Markov process model of the queueing system considered will be established. The distribution is thus uniquely determined by means of entropy maximization. Results of the analysis of queueing problems by entropy maximization are twofold. First, we shall show that various well-known formulae of queueing theory can also 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 networks. Second, in the analysis of queueing networks , we shall also show that the principle of maximum entropy can be used to obtain systematically an estimate of the queue length distribution if the first few moments of the service-time distribution are given.

1.3 RELATED WORKS ON QUEUEING SYSTEMS USING ENTROPY MAXIMIZATION

Since the early developement of the theory of queues by A. K. Erlang to study the problems of telephone traffic, queueing theory has had extensive applications in a large number of practical situations. Computer systems often do not satisfy assumptions made by stochastic process models that are used in queueing theory. However, Kleinrock [4] has successfully studied the application of queueing theory for performance modeling and analysis of computer systems. This indicates that queueing theory equations have wider applicability than suggested by their classical derivations. Denning and Buzen [5] have provided one explanation in terms of "operational analysis". Shore and Johnson [6] have suggested another explanation based on maximium entropy and minimum cross entropy principles.

Maximum entropy analysis was first applied to networks of queues by Ferdinand [7], who showed that the solution of a network of queues obtained by entropy maximization subject to the mean queue lengths of the service stations corresponds to the network solution assuming exponential service times. Shore [8,9] also investigated single queueing systems of type M/M/∞/m, M/M/∞, M/G/1

and G/G/1 using entropy maximization. A further study of the M/G/1 queue, G/M/1 queue and the G/G/1/m queue was presented by Kouvatsos [10,11]. Bard [12,13] applied this method to the analysis of contention in I/O subsystems. Bobrowski [14] examined various applications of this method and argued that the maximum entropy formalism provides a framework for the analysis of networks of queues similar to *Operational Analysis* [5]. Recently Kouvatsos [15] also generalized Ferdinand's results and proposed an implementation of the entropy maximization formalism for the analysis of networks of queues.

1.4 THESIS OUTLINE

In Chapter 2 an equivalent packet arrival process for the bulk-Poisson process has been established and applied to several single-server queueing systems, for example, the $M^X/G/1$ queue, $M^X/M/1/m$ queue, $M^X/G/1$ queue with vacation, and the $M^X/M/C$ queue.

In Chapter 3 the principle of maximum entropy has been applied to study the G/G/C queueing system. State probability distribution and waiting-time distribution have been derived. The mean number of customers and the probability of delay are also calculated for some special systems. Moreover, the method of diffusion approximation has been modified and applied to study the G/G/C system and numerical examples are given for illustration.

Chapter 4 is mainly devoted to the network of queues with bulk arrivals. The following models have been examined : The network with single-server and

multiple-server stations in continuous time, the network with finite-capacity stations, and the network with window-based flow-controlled mechanism. Moreover, networks of stations with nonexponential multiple servers have also been studied.

In Chapter 5 numerical results of network models mentioned above are presented and compared with simulation results to verify the method proposed in Chapter 4.

Chapter 6 is devoted to the G/G/1 queue and queueing networks in discrete time. A framework for the analysis of queueing network based on the entropy maximization and network decomposition has been established. Numerical examples are presented for illustration.

Chapter 7 draws the main conclusions of the work and recommends some topics for further research.

CHAPTER 2

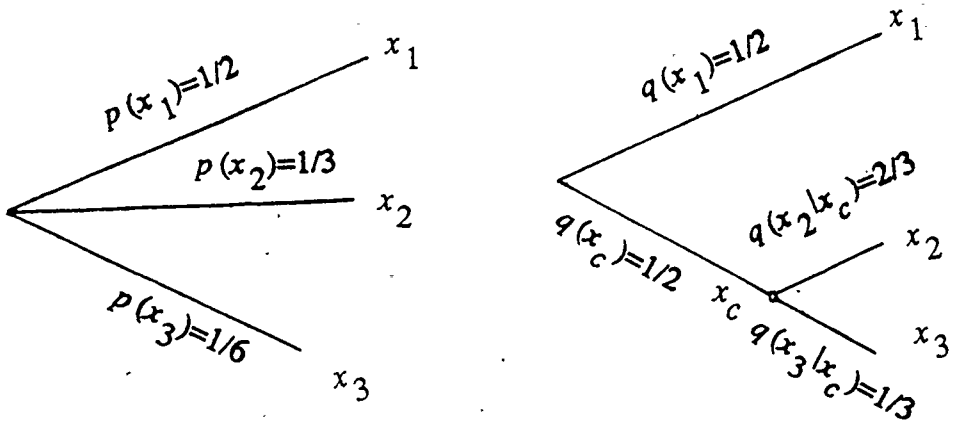
MAXIMUM ENTROPY ANALYSIS OF SIMPLE QUEUEING SYSTEMS

2.1 INTRODUCTION

Entropy maximization is a general approach to inferring a probability distribution from constraints which incompletely or partially characterize that distribution. Analysis by means of entropy maximization has been of interest since Shannon [16] showed, for discrete noiseless systems, that the best encoding of an information source, in the sense of enabling the highest information rate over a fixed capacity channel, is the one that maximizes the source entropy. In addition to continuing applications in communication theory, there has been a growing interest in the use of entropy maximization techniques for probabilistic analysis and problem solving in other fields. Much of this work has been stimulated by that of E.T. Jaynes [28]. Detailed discussions concerning the motivation, justification, and validity of various entropy maximization techniques are available elsewhere [17,19,20,21]. There have been applications in statistics mechanics [22], traffic networks [17], reliability estimation [21], production line decision making [23], system simulation [18], statistics [19,24], spectral analysis [25], image reconstruction [26], and general probabilistic problem solving [20,27].

2.2 THE PRINCIPLE OF MAXIMUM ENTROPY

In this section we shall introduce the principle of maximum entropy. Section 2.2.1 defines entropy as a measure of uncertainty or information. Section 2.2.2 presents the entropy maximization formalism for obtaining the distribution and shows that the analysis can be treated as a nonlinear optimization problem.



(a) (b)

Fig. 2.1 Decomposition of a choice from three possibilities

2.2.1 Definition of Entropy

It has been pointed out by Shannon [16] that there is more uncertainty about an experiment with different possible outcomes if its probability distribution is uniform than if it is sharply peaked. Suppose that an experiment consists of a set of events $\{x_i, i=1,...,k\}$ with probability distribution $\{ p(x_i), i=1,...,k\}$. Shannon's uniqueness [16] theorem shows that subject to requirements of consistency a

quantifiable measure H for the notion of "uncertainty" is the entropy

$$H(p(x_1),..., p(x_k)) = -\sum_{i=1}^{k} p(x_i) \ln p(x_i) \qquad (2.2.1)$$

The logarithm may be to the base 2 or to the base e [99]. When the logarithm is taken with the base 2, the information is measured in units of bits. When the logatithm is taken with the base e, the information is measured in units of nats (a shortened form of "natural units"). The conversion factor is that one nat equals 1.443 bits.

The consistency requirements are :

1. The function H that should be continuous in the variables $p(x_i)$. This means that small changes in the probability distribution can only produce small changes in its entropy function.
2. For equiprobable events, H should be a monotonically increasing function of the number k of these events.
3. If a choice is broken into two succesive choices, the original H should be the weighted sum of the individual values of H.

Fig. 2.1 illustrates the requirement 3 exemplified by Shannon [16]. In Fig. 2.1b, the events x_2 and x_3 are grouped into a composite event x_c first, where we let $q(x_1)=p(x_1)$, $q(x_2|x_c)q(x_c)= p(x_2)$ and $q(x_3|x_c)q(x_c)= p(x_3)$. In this case, it is required that

$$H\left[p(x_1),p(x_2),p(x_3)\right]=H\left[q(x_1),q(x_c)\right]+H\left[q(x_2|x_c),q(x_3|x_c)\right]q(x_c) \qquad (2.2.2)$$

With this definition, Jaynes [28] has justified the principle of maximum entropy as follows :

> Given the propositions of an event and any information relating to them, the most reasonable assignment for the corresponding probabilities is such that the entropy of the probability distribution is maximized subject to the available information, because if any other assignment were chosen, the amount of uncertainty of the probability distribution would not reflect adequately the uncertainty about it.

2.2.2 The Maximum Entropy Formulation

Suppose that the unknown state probability distribution $\{p(x_i),\ i=1,...,k\}$ is incompletely specified by the following mean values m_j of certain functions f_j defined on the system states $\{x_i,\ i=1,2,...,k\}$

$$\sum_{i=1}^{k} f_j(x_i)p(x_i) = m_j \quad , j= 1, 2,..., l \qquad (2.2.3)$$

To determine the distribution $\{p(x_i),\ i=1,2,...,k\}$ by entropy maximization for the constraints (2.2.3), we must solve the optimization problem :

$$\textit{Maximize} \qquad H\left[p(x_1),..., p(x_k)\right] = -\sum_{i=1}^{k} p(x_i)\ \ln p(x_i)$$

$$\textit{subject to} \quad \sum_{i=1}^{k} f_j(x_i)p(x_i) = m_j \ , \ j = 1, 2,..., l$$

where m_j and $f_j(x_i)$ are known quantities.

Applying the method of Lagrange multipliers we form the Lagrangian

$$L = H\left[p(x_1),..., p(x_k)\right] - \sum_{j=1}^{l} \beta_j \left[\sum_{i=1}^{k} f_j(x_i)p(x_i) - m_j\right]$$

where β_j are the Lagrange multipliers associated with the constraints.

Then the necessary conditions for a stationary point of H are

$$\frac{\partial L}{\partial p(x_i)} = 0 \ , i = 1, 2,..., k \tag{2.2.4}$$

$$\frac{\partial L}{\partial \beta_j} = 0 \quad , j = 1, 2,..., l \tag{2.2.5}$$

Performing the differentiations in (2.2.4) and (2.2.5), we obtain , respectively

$$1 + \ln p(x_i) + \sum_{j=1}^{l} \beta_j f_j(x_i) = 0 \ , i = 1, 2,..., k \tag{2.2.6}$$

$$\sum_{i=1}^{k} f_j(x_i)\, p(x_i) = m_j \quad , j = 1, 2,..., l \tag{2.2.7}$$

Solving (2.2.6) for $p(x_i)$ yields

$$p(x_i) = \exp\left[-1 - \sum_{j=1}^{l} \beta_j f_j(x_i)\right] \ , i = 1, 2,..., k \tag{2.2.8}$$

Substituting $p(x_i)$ of (2.2.8) into (2.2.7) results in the following system of nonlinear

equations

$$\sum_{i=1}^{k} f_j(x_i) \exp\left[-1-\sum_{j=1}^{l} \beta_j f_j x_i\right] = m_j \quad , j = 1, 2,...,l \tag{2.2.9}$$

From which we can determine the Lagrange multipliers β_j. Then $p(x_i)$, i=1,2,...,k are given by (2.2.8).

2.3 EQUILIBRIUM STATE PROBABILITY DISTRIBUTION OF THE $G^X/G/1$ QUEUE

In this section we shall apply the method of entropy maximization to study the $G^X/G/1$ queue. Closed-form expressions of state probability distribution will be derived and *generalized-exponential (GE)* distribution will be introduced.

To determine the mean waiting time and the state probability distribution , we shall make the following assumptions :

1. Groups of customers of random size B arrive according to a renewal process. The interarrival time T_g of successive groups has mean $E(T_g)=1/\lambda_g$, second moment $E(T_g^2)$ and coefficient of variation c_{T_g} which is the ratio of the standard deviation to the mean, i.e., $c_{T_g} = \sqrt{E(T_g^2)-E^2(T_g)}/E(T_g)$. This input process will be denoted by G^X.

2. A single server has a general service time distribution with mean $1/\mu$ and coefficient of variation c_Y.

3. The group size B has a general probability distribution P{B=i}, i=1,2,... with mean E(B) and coefficient of variation c_B.

4. The interarrival time T_g, the service time Y and the group size B are mutually independent random variables.

5. The queue discipline is *first in, first out (FIFO)* .

6. The traffic intensity $\rho=\lambda_g E(B)/\mu$ is less than unity and only the statistical equilibrium behavior of the system is considered. A system is said to be in statistical equilibrium if its state probabilities are constant in time. This does not mean that the system does not fluctuate from state to state, but rather that no trends result from the statistical fluctuations. From a practical point of view, a syetem is assumed to be in statistical equilibrium if it has been in operation long enough for the effect of the initial conditions to have worn off [98].

2.3.1 Derivation of the State Probability Distribution of the $G^X/G/1$ Queue

Let the random variable N be the number of customer in the system and $p_n=P\{N=n\}$, n=0,1,..., be the equilibrium state probability that there are n customers in the system. In order to find the *maximum entropy (ME)* state probability distribution for the $G^X/G/1$ queue, we must solve the optimization

problem :

$$\textit{Maximize} \quad H = -\sum_{n=0}^{\infty} p_n \ln p_n \tag{2.3.1}$$

subject to the constraints,

$$\textit{Normalization} \quad \sum_{n=0}^{\infty} p_n = 1 \tag{2.3.2a}$$

$$\textit{Utilization} \quad 1 - p_0 = \rho \ , \ \rho = \lambda_g E(B)/\mu < 1 \tag{2.3.2b}$$

$$\textit{Mean} \quad \sum_{n=0}^{\infty} np_n = E(N) \tag{2.3.2c}$$

For single server queues the utilization constraint is a well known result [4] which can be obtained for any queue with the mean group arrival rate λ_g and mean group size E(B) and mean service time $1/\mu$. Here utilization of the server is defined as the proportion of time that the server is busy.

The mean number of customers, E(N), in the system can be determined in terms of the mean group arrival rate λ_g, mean group size E(B), mean service time $1/\mu$ and the mean waiting time E(W) as shown in (2.3.8).

The corresponding Lagrangian is

$$L = -\sum_{n=0}^{\infty} p_n \ln p_n - \beta_1 \left[\sum_{n=0}^{\infty} p_n - 1 \right] - \beta_2 (1 - p_0 - \rho) - \beta_3 \left[\sum_{n=0}^{\infty} np_n - E(N) \right] \tag{2.3.3}$$

where β_j, j=1,2,3 are the Lagrange multipliers.

The necessary conditions for L to be stationary are

$$\frac{\partial L}{\partial p_0} = -ln\ p_0 - 1 - \beta_1 + \beta_2 = 0 \qquad (2.3.4a)$$

$$\frac{\partial L}{\partial p_n} = -ln\ p_n - 1 - \beta_1 - \beta_3 n = 0\ ,\ n = 1, 2,... \qquad (2.3.4b)$$

$$\frac{\partial L}{\partial \beta_j} = 0\ , \qquad j = 1, 2, 3 \qquad (2.3.4c)$$

Note that (2.3.4c) are restatements of the constraints in (2.3.2).

Solving (2.3.4) for p_n we find

$$p_n = a_1\ a_2^{h(n)} a_3^n\ ,\ n = 0, 1, 2,... \qquad (2.3.5)$$

where

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

and

$$a_1 = e^{-1-\beta_1}$$

$$a_2 = e^{\beta_2}$$

$$a_3 = e^{-\beta_3}$$

Using (2.3.2a) and (2.3.2c) and solving for a_2 and a_3, we find

$$a_2 = \frac{\rho^2}{(1-\rho)[E(N)-\rho]} \qquad (2.3.6a)$$

$$a_3 = \frac{E(N)-\rho}{E(N)} \qquad (2.3.6b)$$

2.3.2 The Mean Waiting Time

Now we consider a test customer in a group of the arrival process to the $G^X/G/1$ queue. We shall calculate the mean waiting time E(W) of this customer. Yao [29] proved that the mean waiting time E(W) can be divided into two parts:

1. The mean waiting time $E(W_1)$ of the group. This is the mean waiting time of the first customer in the group to receive service. Thus $E(W_1)$ is essentially the mean waiting time of a test group, where the interarrival time between two consecutive groups is assumed to have mean $1/\lambda_g$ and coefficient of variation c_{T_g} and $E(W_1)$ can be found by solving Lindley's equation using the spectral method [3].

2. The mean waiting time $E(W_2)$ of the test customer in that group. This mean time is the sum of the mean service times of customers who are ahead of him in the same group. Thus

$$E(W_2) = \frac{-1 + E(B)\,(1 + c_B{}^2)}{2\,\mu}$$

Therefore the mean waiting time E(W) is given by

$$E(W) = E(W_1) + \left[\frac{-1+E(B)\,(1+c_B^2)}{2\mu}\right] \tag{2.3.7}$$

By means of Little's formula [30], the average number of customers in the system is given by

$$E(N) = \lambda_g E(B)\left[E(W)+\frac{1}{\mu}\right] \tag{2.3.8}$$

Example The $M^X/G/1$ queue

Suppose that group arrivals follow a Poisson process with mean λ_g. Then the G^X input process reduces to the Poisson group arrival M^X input process.

Further, we assume that the group size B has a geometric distribution $P\{B=i\}=p(1-p)^{i-1}$, $i=1,2,...$, with mean E(B)=1/p and coefficient of variation $c_B = \sqrt{1-p}$. The generating function of P{B=i} is then given by

$$G_B(z) = \frac{pz}{1-(1-p)z}$$

Since $E(W_1)$ is known and given by [3]

$$E(W_1) = \frac{\rho(2/p-1+c_Y^2)}{2\mu(1-\rho)} \tag{2.3.9}$$

substituting $E(W_1)$, E(B) and c_B into (2.3.7), we find the mean waiting time

$$E(W) = \frac{\rho-2+\frac{2}{p}+\rho c_Y^2}{2\mu(1-\rho)}$$

Using (2.3.8), we obtain the average number of customers in the system

$$E(N) = \frac{\rho(2-\rho p+\rho p c_Y^2)}{2p(1-\rho)} \tag{2.3.10}$$

Furthermore, using (2.3.5),(2.3.6) and (2.3.10) we obtain the ME state probability distribution { p_k, k=0,1,...},

$$p_n = \begin{cases} 1-\rho \text{ , if } n = 0 \\ (1-\rho)\, a_2\, a_3^n \text{ , if } n \geq 1 \end{cases} \tag{2.3.11}$$

where

$$a_2 = \frac{2p\rho}{2+p\rho-2p+\rho pc_Y^2}$$

and

$$a_3 = \frac{2+p\rho-2p+\rho pc_Y^2}{2-\rho p+\rho pc_Y^2}$$

According to Kleinrock [3], the generating function of state probability distribution of the $M^X/G/1$ queue is given by

$$G_N(z) = \frac{(1-\rho)(1-z) f_Y^*\left[\lambda_g - \lambda_g G_B(z)\right]}{f_Y^*\left[\lambda_g - \lambda_g G_B(z)\right] - z} \tag{2.3.12}$$

where $f_Y^*(\bullet)$ is the Laplace transform of the probability density function of the service time. Now the generating function of the ME state probability distribution in (2.3.11) is given by

$$G_N(z) = \frac{(1-\rho)[1-a_3(1-a_2)z]}{1-a_3 z} \tag{2.3.13}$$

Suppose that the probability generating function $G_N(z)$ in (2.3.12) and (2.3.13) are equivalent. By equating the right hand sides of (2.3.12) and (2.3.13) and solving for $f_Y{}^*(s)$, we obtain

$$f_Y{}^*(s) = \frac{2\mu+s(c_Y^2-1)}{2\mu+s(c_Y^2+1)} \tag{2.3.14a}$$

Now taking the inverse Laplace transform results in the density function of the service time

$$f_Y(t) = \begin{cases} \left[\dfrac{c_Y^2-1}{c_Y^2+1}\right]\delta(t) \text{ , if } t = 0 \\ \mu\left[\dfrac{2}{1+c_Y^2}\right]^2 \exp\left[\dfrac{-2\mu t}{1+c_Y^2}\right] \text{ , if } t > 0 \end{cases} \tag{2.3.14b}$$

This probability density function may be regarded as the ME probability density function of the service time in the $M^X/G/1$ queue.

By integration, we obtain the corresponding distribution function

$$F_Y(t) = 1-\left[\frac{2}{1+c_Y^2}\right]\exp\left[\frac{-2\mu t}{1+c_Y^2}\right] \text{ , } t \geq 0 \tag{2.3.15}$$

In general, specifying only the mean λ and the coefficient of variation c_T of a random variable T, we can determine a distribution $F_T(t)$ expressed in the following form

$$F_T(t) = 1-be^{-at} \quad , t\geq 0 \tag{2.3.16}$$

with

$$b = \frac{2}{1+c_T^2}$$

and

$$a = \lambda b$$

This distribution (2.3.16) is known as a generalized exponential (GE) distribution [10].

We have shown that if the service time distribution of the M^X/G/1 queue is assumed to be that of (2.3.15), then the ME state probability distribution of the M^X/G/1 queue is given by (2.3.11).

2.4 CHARACTERIZATION OF STATISTICS OF PACKET ARRIVALS IN A MESSAGE ARRIVAL PROCESS

This section will establish an equivalence relationship between the message arrival process and the GE distribution. We shall first determine the interarrival time distribution for packets and then apply it to study single queues.

In computer communication networks, packet switching techniques are widely used, where the messages are divided into smaller pieces called packets, each of which has a maximum length. In this application the message arrival process corresponds to the group arrival process studied in the previous section and the packets correspond to customers.

Since the message length is a random quantity, a message may consist of a random number of packets. In this case we shall say that the arrival process of packets forms a bulk arrival process.

2.4.1 Equivalent Packet Arrival Process for a Message Arrival Process

Consider a message arrival process satisfying the following conditions:

1. Message arrivals follow a stationary and orderly input process with mean arrival rate λ_g (groups/second).
2. Each message consists of a random number B of packets with probability b_i=P {B=i} , i=1,2,... and mean E(B).
3. Let T be a random variable denoting the packet interarrival time with distribution function $F_T(t)$ and probability density function $f_T(t)$.

We note first that the probability of zero packet interarrival time is given by

$$P\{T=0\} = F_T(0) = 1 - \frac{1}{E(B)}$$

Second, since the message arrival rate is λ_g and the average message length is E(B), the mean packet arrival rate is simply equal to $\lambda_g E(B)$.

Now we shall derive the distribution function of the interarrival time T.

Since $F_T(t)$ has a jump of magnitude of 1- 1/E(B) at t=0, then the probability density function $f_T(t)$ can be written as

$$f_T(t) = (1-\frac{1}{E(B)})\delta(t) + f_c(t) \tag{2.4.1}$$

where $f_c(t)$ denotes the continuous part of the probability density function $f_T(t)$ and $\delta(t)$ is the Dirac delta function.

Let the entropy function H be defined as

$$H = -\int_{t=0^+}^{\infty} f_c(t) \ln f_c(t)\, dt \tag{2.4.2}$$

Using (2.4.1) we calculate the mean interarrival time

$$\int_{t=0}^{\infty} t f_T(t)\, dt = \int_{t=0^+}^{\infty} t f_c(t)\, dt = \frac{1}{\lambda_g E(B)} \tag{2.4.3}$$

And then the normalization condition is given by

$$F_T(0) + \int_{t=0^+}^{\infty} f_c(t) = 1$$

Since $F_T(0) = 1-1/E(B)$, it follows that

$$\int_{t=0^+}^{\infty} f_c(t) = \frac{1}{E(B)} \tag{2.4.4}$$

To find $f_c(t)$, we formulate the problem as follows:

Maximize H *in* (2.4.2)

subject to the constraints in (2.4.3) and (2.4.4)

This is an isoperimetric problem of calculus of variations [31]. We can solve the optimization problem by introducing the Lagrange multipliers β_k, k=1,2 and forming the Lagrangian (function)

$$L = -f_c(t)\ln f_c(t) + \beta_1 t f_c(t) + \beta_2 f_c(t) \tag{2.4.5}$$

Thus the unknown density function $f_c(t)$ must satisfy the Euler-Lagrange equation

$$\frac{\partial L}{\partial f_c(t)} = -\ln f_c(t) - 1 + \beta_1 t + \beta_2 = 0$$

Solving this equation for $f_c(t)$ leads to

$$f_c(t) = \exp(-1+\beta_2+\beta_1 t) \tag{2.4.6}$$

Now using (2.4.3) and (2.4.4), we find

$$\beta_1 = -\lambda_g$$

and

$$\beta_2 = 1 + \ln [\lambda_g / E(B)]$$

Substituting β_1 and β_2 into (2.4.6), we obtain

$$f_c(t) = \lambda_g e^{-\lambda_g t / E(B)} \tag{2.4.7}$$

Then it follows from (2.4.1) and (2.4.7) that the distribution function $F_T(t)$ is given by

$$F_T(t) = 1 - e^{-\lambda_g t / E(B)}, \; t \geq 0$$

It follows that the coefficient of variation c_T is

$$c_T = \sqrt{2E(B) - 1}$$

In general, we can express $F_T(t)$ as a GE distribution function in terms of the mean and the coefficient of variation in the form [10]

$$F_T(t) = 1 - \left[\frac{2}{1+c_T^2}\right] \exp\left[\frac{-2\lambda t}{1+c_T^2}\right]$$

We shall establish an equivalence relationship between the message arrival process and the associated GE interarrival time distribution $F_T(t)$ of packets as follows

Message Arrival Process		Equivalent Packet Arrival Process
Mean message arrival rate	λ_g	GE packet interarrival time distribution
Mean message size	$E(B)$	$1 - \frac{2}{1+c_T^2} \exp\left[\frac{-2\lambda t}{1+c_T^2}\right]$ with $\lambda = \lambda_g E(B)$ (2.4.8) and $c_T = \sqrt{2E(B) - 1}$

In particular, if message arrivals follow a Poisson process with mean arrival rate λ_g and message size has the geometric distribution, $P\{B=i\}=p(1-p)^{i-1}$, $i=1,2,...$, then

$$E(B) = \frac{1}{p} \tag{2.4.9}$$

and (2.4.7) becomes

$$f_c(t) = \lambda_g p e^{-\lambda_g t} \tag{2.4.10}$$

The coefficient of variation c_T is

$$c_T = \sqrt{2/p - 1} \tag{2.4.11}$$

This particular input process is called bulk Poisson process.

It is interesting to note that for the special case if message arrival process is Poisson and E(B)=1/p=1, then the equivalence relationship (2.4.8) reduces to a relationship between a Poisson arrival process and its associated exponential interarrival time distribution.

From (2.4.8), (2.4.9) and (2.4.11), we obtain

$$\lambda_g = \frac{2\lambda}{1+c_T{}^2} \tag{2.4.12a}$$

and

$$p = \frac{2}{1+c_T{}^2} \tag{2.4.12b}$$

Having established an equivalence between a bulk Poisson arrival process and its associated interarrival time distribution in the maximum-entropy sense, we shall consider the following queueing models : The $M^X/G/1$, the $M^X/M/1/m$, the $M^X/G/1$ with vacation, and the $M^X/M/C$ queue.

2.4.2 The $M^X/G/1$ Queue

Consider a bulk Poisson input, single-server queue with mean arrival rate λ_g and geometric batch size with mean 1/p. The general service time distribution has mean $1/\mu$, coefficient of variation c_Y and the Laplace-Stieltjes transform $F_Y{}^*(s)$. The traffic intensity $\rho=\lambda_g/\mu p$ is assumed to be less than one so that equilibrium state exists. The generating function of the queue length is given by [3]

$$G_N(z) = \frac{(1-\rho)\,(1-z)\,F_Y{}^*\left[\dfrac{\lambda_g-\lambda_g z}{1-(1-p)z}\right]}{F_Y{}^*\left[\dfrac{\lambda_g-\lambda_g z}{1-(1-p)z}\right]-z} \tag{2.4.13}$$

Following (2.4.12a) and (2.4.12b) and replacing p by $2/1+c_T^2$, λ_g by $2\lambda/1+c_T^2$ and keeping ρ unchanged, we obtain the generating function $G_{N_1}(z)$ of queue length for the equivalent GE/G/1 queue

$$G_{N_1}(z) = \frac{(1-\rho)\,(1-z)\,F_Y^*(\xi)}{F_Y^*(\xi)-z} \tag{2.4.14}$$

where

$$\xi = \frac{2\lambda(1-z)}{c_T^2+1-z\,(c_T^2-1)}$$

Suppose that the service time Y has the GE distribution $F_Y(t)$ as shown in (2.3.15). Since $F_Y^*(s)$ is given by (2.3.14a), by substituting $F_Y^*(\bullet)$ into (2.4.14), we get, after simplification, the generating function of the state probability distribution of the GE/GE/1 queue

$$G_{N_1}(z) = (1-\rho)\left[\frac{1-(1-a_1)\,a_2 z}{1-a_2 z}\right] \tag{2.4.15a}$$

where

$$a_1 = \frac{2\rho}{c_T^2+\rho c_Y^2+\rho-1} \tag{2.4.15b}$$

and

$$a_2 = \frac{c_T{}^2+\rho c_Y{}^2+\rho-1}{c_T{}^2+\rho c_Y{}^2-\rho+1} \tag{2.4.15c}$$

By taking the inverse z-transform of (2.4.15a), we obtain the state probability distribution $\{p_n\}$

$$p_n = \begin{cases} 1-\rho \ , \text{ if } \ n=0 \\ (1-\rho)\, a_1\, a_2{}^n \ , \text{ if } n \geq 1 \end{cases}$$

where a_1 and a_2 are given by (2.4.15b) and (2.4.15c), respectively. This result agrees with that obtained by Kouvatsos [11].

2.4.3 The $M^X/M/1/m$ Queue

Consider a $M^X/M/1$ queue with finite capacity of queue length m-1. From Gross and Harris [34], the state probability distribution is given by

$$p_n = \begin{cases} \dfrac{1-\rho}{1-\rho\alpha^m} \ , \text{ if } n=0 \\ p_0\, \rho p\, \alpha^{n-1} \ , \text{ if } 1 \leq n \leq m \end{cases} \tag{2.4.16a}$$

with

$$\rho = \frac{\lambda_g}{\mu p}$$

and

$$\alpha = 1-p\,(1-\rho)$$

Using the equivalence relationship of (2.4.12a) and (2.4.12b), we obtain the state probability distribution of the corresponding GE/M/1/m queue as

$$p_n = \begin{cases} \dfrac{1-\rho}{1-\rho\alpha^m}\,, \text{ if } n=0 \\ p_0 \left[\dfrac{2\rho}{1+{c_T}^2}\right] \alpha^{n-1}\,, \text{ if } 1 \le n \le m \end{cases} \qquad (2.4.16b)$$

with

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

and

$$\alpha = 1-\frac{2(1-\rho)}{1+{c_T}^2}$$

This result is identical with that obtained by Kouvatsos [11].

2.4.4 The $M^X/G/1$ Queue with Vacation

Consider the $M^X/G/1$ system in which the server takes vacation after completion of a service. When a vacation ends, the queue can be either empty or nonempty. If the queue is nonempty, service is resumed. If the queue is empty, a new vacation period begins. The $M^X/G/1$ vacation models are distinguished by the

rules which determine when a service stops and a vacation begins. We shall consider two cases:

a. The exhaustive service case in which the server works continuously until the system becomes empty.

b. The nonexhaustive service case where the server takes vacation after every service completion.

Case a. The Exhaustive Service Model

Let the Laplace-Stieltjes transforms of the waiting time distribution and vacation time distribution be $F_W{}^*(s)$ and $F_V{}^*(s)$, respectively. Baba [35] studied a $M^X/G/1$ queue with vacation and exhaustive service discipline and showed that the waiting time W is the sum of the following independent random variables, one of which is the waiting time of an arbitrary customer for the $M^X/G/1$ queue without vacation and the other one is the remaining vacation time for the server on vacation. He obtained the Laplace-Stieltjes transform $F_W{}^*(s)$ of the waiting time distribution as

$$F_W{}^*(s) = \frac{(1-\rho)sp}{s[1-(1-p)\,F_Y{}^*(s)]-\lambda_g[1-F_Y{}^*(s)]} \bullet \frac{1-F_V{}^*(s)}{sE(V)} \qquad (2.4.17)$$

By using the equivalence relationship of (2.4.12a) and (2.4.12b) for λ_g and p, the corresponding result for the GE/G/1 queue with vacation and exhaustive service

disciplicine becomes

$$F_W{}^*(s) = \frac{2s(1-\rho)}{s[1+c_T{}^2-(c_T{}^2-1)F_Y{}^*(s)]-2\lambda[1-F_Y{}^*(s)]} \bullet \frac{1-F_V{}^*(s)}{sE(V)} \quad (2.4.18)$$

where λ, c_T, $F_Y{}^*(s)$ and E(V) represent respectively the average arrival rate, the coefficient of variation of interarrival time, the Laplace-Stieltjes transform of the service time distribution, and the average vacation time.

Note that (2.4.18) agrees with that obtained by Keilson [36].

Case b. The Nonexhaustive Service Model

Keilson [36] also showed that the waiting time in the G/G/1 queue with vacation and nonexhaustive service discipline consists of two components :

1. The waiting time in the G/G/1 queue with the original interarrival time and the service time which is the sum of the original service time and the vacation time.

2. The remaining vacation time.

Note that the Laplace-Stieltjes transform $F_W{}^*(s)$ of the waiting time distribution of this model can be obtained from (2.4.18) by using a modified service time which is equal to the sum of the original service time and the vacation time. From (2.4.18) by replacing $F_Y{}^*(s)$ by $F_Y{}^*(s)F_V{}^*(s)$ we get the Laplace-Stieltjes transform of the waiting time distribution

$$F_W^*(s) = \frac{2s(1-\rho)}{s[1+c_T^2-(c_T^2-1)F_Y^*(s)\,F_V^*(s)]-2\lambda[1-F_Y^*(s)F_V^*(s)]} \bullet \frac{1-F_V^*(s)}{sE(V)} \tag{2.4.19}$$

By using the equivalence relationship of (2.4.12) for λ and c_T , we obtain the result of the $M^X/G/1$ vacation model

$$F_W^*(s) = \frac{(1-\rho)sp}{s[1-(1-p)F_Y^*(s)\,F_V^*(s)]-\lambda_g pF_Y^*(s)\,F_V^*(s)} \bullet \frac{1-F_V^*(s)}{sE(V)} \tag{2.4.20}$$

where $\rho=\lambda_g[\frac{1}{\mu}+E(V)]$

Since

$$E(W) = \frac{dF_W^*(s)}{ds}\Big|_{s=0},$$

using Little's formula, we obtain the average number of customers in the system as

$$E(N) = \lambda_g E(W)/p + \rho \tag{2.4.21}$$

$$= \frac{\lambda_g E(V^2)}{2pE(V)} + \frac{\rho^2(c_Y^2-1)+2\rho/p}{2(1-\rho)}$$

Note that this result agrees with that obtained by Kuehn [37].

2.4.5 The M^X/M/C Queue

We shall extend the study of single-server queues to multiserver queues with the following specifications :

1. The arrival process is a bulk Poisson process with arrival rate λ_g and geometric batch size with mean 1/p.
2. There are C exponential servers, each with the same mean service time $1/\mu$.

According to Chaudhry and Templeton [38], the equilibrium state probability distribution $\{ p_n, n = 0,1,..\}$ of the M^X/ M/C queue is given by

$$p_n = \begin{cases} \dfrac{p_0 \prod_{l=0}^{n-1} \alpha + lq}{n!} , \text{ if } 1 \le n \le C \\ p_C [\rho + q(1-\rho)]^{n-C} , \text{ if } n \ge C \end{cases} \qquad (2.4.22)$$

where

$$\alpha = \frac{\lambda_g}{\mu}$$

$$\rho = \frac{\lambda_g}{\mu C p}$$

$$q = 1-p$$

and p_0 is determined by the normalization condition

By using the equivalence relationship of (2.4.12a) and (2.4.12b), we find the corresponding state probability distribution for the GE/M/C queue

$$p_n = \begin{cases} \dfrac{p_0}{n!} \prod_{r=0}^{n-1} \dfrac{2\rho C + r(c_T^2 - 1)}{c_T^2 + 1}, & \text{if } 1 \le n \le C \\ p_C \left[\dfrac{c_T^2 - 1 + 2\rho}{1 + c_T^2} \right]^{n-C}, & \text{if } n \ge C \end{cases} \tag{2.4.23}$$

where $\rho = \dfrac{\lambda}{\mu C}$

This probability distribution is identical with those obtained by Takacs [32].

In particular, if we let $C = \infty$ in (2.4.22), then the equilibrium state probability distribution of the $M^X/M/\infty$ queue becomes

$$p_n = \frac{p_0}{n!} \prod_{r=0}^{n-1} \left[\frac{\lambda_g}{\mu} + r - rp \right], \quad n \ge 1$$

where

$$p_0 = 1/[1 + \sum_{n=1}^{\infty} \prod_{r=0}^{n-1} \frac{\lambda_g/\mu + r(1-p)}{n!}]$$

Similarly, let $C = \infty$ in (2.4.23). Then the corresponding state probability distribution of the GE/M/∞ queue is given by

$$p_n = \frac{p_0}{n!} \prod_{r=0}^{n-1} \frac{2\lambda/\mu + rc_T{}^2 - r}{c_T{}^2 + 1} \tag{2.4.24}$$

where

$$p_0 = 1/[1 + \sum_{n=1}^{\infty} \prod_{r=0}^{n-1} \frac{2\lambda/\mu + r(c_T^2 - 1)}{(c_T^2 + 1)\, n!}]$$

2.5 SUMMARY

In this chapter, a method based on the principle of maximum entropy is used to study queueing systems in continuous time. After the introduction of the basic concepts and solution methods, the $G^X/G/1$ queue is investigated and its equilibrium state probability distribution is determined. Then a GE (generalized exponential) distribution is introduced to show that the maximum entropy solution of the $M^X/GE/1$ queue is exact if the bulk size has a geometric distribution.

An equivalence relationship between the M^X and GE processes is established and is applied for converting queues with M^X input to queues with GE input. The significance of the generalized exponential distribution is that it combines two distributions into one and hence facilitates the analysis of many queueing systems with bulk Poisson input. In particular, the $M^X/G/1$ queue, the $M^X/G/1$ queue with vacation , the $M^X/M/1/m$ queue and the $M^X/M/C$ queue have been investigated.

CHAPTER 3

MULTISERVER QUEUES

3.1 INTRODUCTION

The increasing interest in multiprocessor computer systems stems from the rapid development of mini and microprocessor technology. The expanding applications of multiprocessor systems make it imperative to develop methods of evaluating operating quality figures for these systems. These systems can be modeled as G/G/C queueing systems with general interarrival time and service time distributions. However, exact analysis of the G/G/C queueing system is extremely difficult and formidable. Results [39] are often expressed in a complicated mathematical form and the formulas obtained are difficult to employ even for numerical computations.

In this chapter two methods will be developed for the analysis of G/G/C queues. One is based on the entropy maximization and the state probability distribution and waiting time distribution are considered. The other is based on diffusion approximation. Numerical examples are given for illustration.

3.2 THE G/G/C QUEUE - MAXIMUM ENTROPY FORMULATION I

Consider an arbitrary but stable G/G/C queue with C servers and *FIFO* queue discipline. Let $\{p_n\}$, n=0,1,... be the steady-state probability distribution of having n customers in the system at any moment. Further, let λ and μ be the mean arrival

rate and mean service rate, respectively.

Suppose that some information about the state probabilities are known and are expressed in the following constraints:

(i) The normalization condition

$$\sum_{n=0}^{\infty} p_n = 1 \tag{3.2.1}$$

(ii) The mean number of customers in the system

$$\sum_{n=0}^{\infty} n\, p_n = E(N) \tag{3.2.2}$$

Note that E(N) may be known numerically via system measurement for a finite observation period or can be determined via known analytic formulas based on operational or stochastic assumptions.

Moreover, for a G/G/C queue at equilibrium, the conservation law of traffic must hold, i.e., the arrival rate must be equal to the departure rate, so that we have the additional constraint,

(iii) The conservation law of traffic [92]

$$\lambda = \sum_{n=0}^{C-1} n\mu\, p_n + C\mu \sum_{n=C}^{\infty} p_n \tag{3.2.3}$$

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

$$\textit{Maximize} \quad H = -\sum_{n=0}^{\infty} p_n \ln p_n \tag{3.2.4}$$

subject to the constraints (3.2.1), (3.2.2) and (3.2.3).

The optimization problem can be carried out using Lagrange's method of undetermined multipliers leading to the solution

$$p_n = \begin{cases} \dfrac{x^n\, y^n}{z} & \text{if } 0 \le n \le C-1 \\ \dfrac{x^C\, y^n}{z} & \text{if } n \ge C \end{cases} \tag{3.2.5}$$

where

$$z = \exp(-1-\beta_1),$$

$$x = \exp(-\beta_3),$$

$$y = \exp(-\beta_2),$$

and β_1, β_2, β_3 are the Lagrange multipliers associated with the constraints (3.2.1), (3.2.2) and (3.2.3) respectively.

Substituting p_n of (3.2.5) into (3.2.1)-(3.2.3), we obtain respectively

$$p_0 = \frac{(1-y)(1-xy)}{1-y+x^C\, y^{C+1}-(xy)^{C+1}} \tag{3.2.6a}$$

$$E(N) = p_0\left[\frac{xy - C\,(xy)^C + (C-1)\,(xy)^{C+1}}{(1-xy)^2} + \Theta\right] \tag{3.2.6b}$$

where

$$\Theta = \frac{(xy)^C [C + (1 - C)y]}{(1-y)^2}$$

and

$$p_0 \left[\frac{xy - C(xy)^C - (1-C)(xy)^{C+1}}{(1 - xy)^2} + C\frac{(xy)^C}{1 - y} \right] = \frac{\lambda}{\mu} \qquad (3.2.6c)$$

Then from (3.2.5), we have the following "one-step" maximum entropy recursion formula :

$$p_n = \begin{cases} x\ y\ p_{n-1}\ , \text{ if } \ 1 \le n \le C-1 \\ y\ p_{n-1}\ , \ \text{ if } \ n \ge C \end{cases} \qquad (3.2.7)$$

In an operational context for a given estimated value for E(N), the probabilities in (3.2.5) can be evaluated by using numerical techniques with specified values of the Lagrange multipliers β_1, β_2 , and β_3. An APL program written by Johnson [40] may be used for this purpose. However,in a stochastic context, the quantities x and y can be determined approximately by using known asymptotic property of the state probability distribution of a G/G/C queue.

From Takahashi [41], for a given interarrival time distribution $F_T(\cdot)$ and service time distribution $F_Y(\cdot)$ with rational Laplace-Stieltjes transforms $F_T{}^*(s)$ and $F_Y{}^*(s)$, respectively, the limit

$$\lim_{n\to\infty} \frac{p_{n+1}}{p_n} = y < 1$$

exists, or

$$\frac{p_{n+1}}{p_n} = y < 1 \quad , \text{ if } n \textit{ is sufficiently large} \tag{3.2.8}$$

where

$$y = F_T{}^*(Ck) \tag{3.2.9}$$

and k is the unique positive root satisfying the characteristic equation [41]

$$F_T{}^*(Ck)\, F_Y{}^*(-k) = 1 \tag{3.2.10}$$

In particular, for the G/M/C queue, Takacs [32] showed that

$$\frac{p_{n+1}}{p_n} = 1 - \frac{k}{\mu}, \quad \text{if } n \geq C \tag{3.2.11}$$

Once y is determined, x and p_0 can be obtained by solving (3.2.6a) and (3.2.6c). Then $z=1/p_0$ can be determined. Note that the left-hand side of (3.2.6c) is a polynomial of x. By using the Routh-Hurwitz criterion [42], we find that (3.2.6c) has three positive roots where there exists a double root at $x = 1/y$ and a root less than $1/y$ which is desired for (3.2.5) and can be determined numerically by solving (3.2.6c).

Example 1. The G/G/1 queue

In the case of single-server queues, i.e. C=1, (3.2.3), (3.2.6a) and (3.2.6c) become, respectively,

$$\lambda = \mu(1 - p_0)$$

$$p_0 = \frac{1 - y}{1 - y + xy}$$

and

$$\frac{x\ y\ p_0}{1 - y} = \frac{\lambda}{\mu}$$

where

$$y = F_T{}^*(k)$$

and k is determined by (3.2.10) with C=1, i.e., $F_T{}^*(k)\ F_Y{}^*(-k) = 1$.

Furthermore, we get from (3.2.6b)

$$E(N) = \frac{\rho}{1 - y} \tag{3.2.12}$$

This result is identical to that obtained by Gaver [43] who analyzed the G/G/1 queue by using the method of diffusion approximation.

Example 2. The G/GE/1 queue

Suppose that the service time distribution is a generalized exponential distribution with service rate μ and coefficient of variation c_Y and its Laplace - Stieltjes transform $F_Y{}^*(s)$ is given by (2.3.14a). Then from (3.2.9) and (3.2.10),we obtain

$$y\ F_Y^{*}(-k) = 1$$

or

$$y = \frac{2\mu - k(1 + c_Y^2)}{2\mu - k(c_Y^2 - 1)}$$

Substituting this y into (3.2.12), we have

$$E(N) = \frac{\lambda}{k} + \frac{\rho(1 - c_Y^2)}{2} \tag{3.2.13}$$

where k is the positive root of equation (3.2.10).

It should be pointed out that E(N) is exact and agrees with the result obtained by Kouvatsos [11].

3.3 THE G/G/C QUEUE - MAXIMUM ENTROPY FORMULATION II

Now suppose that instead of the conservation law of traffic in (3.2.3) we have the constraints,

$$p_n = a_n \quad , n = 0, 1, ..., C-1$$

where a_n are supposed to be given.

The entropy maximization problem of the G/G/C queue can be formulated as follows:

Maximize $$H(p) = -\sum_{n=0}^{\infty} p_n \ln p_n$$

subject to the constraints,

$$\sum_{n=0}^{\infty} p_n = 1 \tag{3.3.1a}$$

$$\sum_{n=0}^{\infty} n\, p_n = E(N) \tag{3.3.1b}$$

and

$$p_n = a_n \quad , \text{if } n = 0, 1, \ldots, C-1 \tag{3.3.1c}$$

By using the solution method described in Section 2.3, we find

$$p_n = \begin{cases} \exp(-1-\beta_1-\alpha_n-\beta_2 n) \text{ , if } 0 \le n \le C-1 \\ \exp(-1-\beta_1-\beta_2 n) \text{ , if } n \ge C \end{cases} \tag{3.3.2}$$

where β_i, i=1,2 and α_j, j=0,1,...,C-1 are the Lagrange multipliers associated with the constraints in (3.3.1a)-(3.3.1c), respectively.

Substituting p_n of (3.3.2) into (3.3.1a)-(3.3.1c), we obtain, after some algebraic manipulations,

$$e^{-\beta_2} = \frac{E(N) - C(1-a_0) + \sum_{i=1}^{C-1}(C-i)a_i}{E(N) - (C-1)(1-a_0) + \sum_{i=1}^{C-1}(C-i-1)a_i} \tag{3.3.3a}$$

$$e^{-1-\beta_1} = \frac{(1 - \sum_{i=0}^{C-1} a_i)^2 \Psi}{\left[E(N) - C(1-a_0) + \sum_{i=1}^{C-1}(C-i)a_i\right]^C} \tag{3.3.3b}$$

where

$$\Psi = \left[E(N) - (C-1)(1-a_0) + \sum_{i=1}^{C-1}(C-i-1)a_i\right]^{C-1}$$

and

$$e^{-\alpha_j} = a_j \, e^{\beta_1 + j\beta_2 + 1} \quad , j=0,1,\cdots C-1 \tag{3.3.3c}$$

Introducing (3.2.8) into (3.3.2), we obtain

$$e^{-\beta_2} = y \tag{3.3.4a}$$

where y is given by (3.2.9).

And then substituting for $e^{-\beta_2}$ into (3.3.3a), we obtain the mean number of customers in the system

$$E(N) = \frac{(C - Cy + y)(1 - a_0) + \sum_{i=1}^{C-1} [(C-i)(y-1) - y]a_i}{1 - y} \tag{3.3.4b}$$

By using E(N) in (3:3.3a)-(3.3.3c), β_1 and α_i can be determined.

3.3.1 The G/M/C Queue

For the G/M/C queue, we let $F_T(t)$ be the interarrival time distribution and $\rho = \frac{\lambda}{\mu C} < 1$. The service time distribution is now given by

$$F_Y(t) = 1 - e^{-\mu t} \tag{3.3.5}$$

where μ is the mean service rate.

From (3.2.10), we have

$$F_T^*(Ck)\,\frac{\mu}{\mu - k} = 1 \tag{3.3.6a}$$

or

$$k = \mu - \mu\, F_T^*(Ck) \tag{3.3.6b}$$

where $F_T^*(\cdot)$ is the Laplace - Stieltjes transform of $F_T(t)$.

Since $y = F_T^*(Ck)$, (3.3.6b) can be written as

$$k = \mu(1 - y)$$

or

$$Ck = C\mu(1-y)$$

Therefore, we obtain the functional equation

$$y = F_T{}^*[C\ \mu(1-y)] \tag{3.3.6c}$$

Takacs [32] studied the G/M/C queue and obtained the state probabilities $\{p_n\}$, for $0 \le n \le C-1$ as

$$p_0 = 1 - \rho - C\ \rho \sum_{j=1}^{C-1} p^*{}_{j-1}\ (\frac{1}{j} - \frac{1}{C})$$

$$p_n = \frac{C\rho}{n}\ p^*{}_{n-1}\ ,\ 1 \le n \le C-1 \tag{3.3.7}$$

where the probability distribution of the number of customers found at the arrival instant is

$$p_n^* = \sum_{r=n}^{C-1} (-1)^{r-n} \binom{r}{n} u_r\ ,\ 0 \le n \le C-1$$

with

$$u_r = \alpha\, C_r \sum_{j=r+1}^{C} \frac{\binom{C}{j}}{C_j(1-\phi_j)} f_j$$

$$\alpha = \frac{1}{\frac{1}{1-y} + \sum_{j=1}^{C} \frac{\binom{C}{j}}{c_j\ (1-\phi_j)} f_j}$$

$$C_j = \prod_{l=1}^{j} \frac{\phi_l}{1-\phi_l}$$

and

$$\phi_l = F_T^*(l\mu)$$

$$f_j = \frac{C\ (1-\phi_j) - j}{C\ (1-y) - j}$$

Substituting p_n, n=0,..., C-1 of (3.3.7) for a_n into (3.3.4b), we find the mean number of customers

$$E(N) = \frac{\alpha\,\rho + C\ \rho(1-y)^2}{(1-y)^2} \tag{3.3.8}$$

Then applying $p_n = a_n$, $0 \le n \le C-1$ and E(N) into (3.3.3b), we obtain

$$e^{-1-\beta_1} = \alpha\,\rho\,y^{-1-C}$$

Finally introducing $e^{-1-\beta_1}$ and $e^{-\beta_2} = y$ into (3.3.2), we obtain

$$p_n = \alpha\,\rho\,y^{n-C-1} \quad , \text{if } n \ge C \tag{3.3.9}$$

It is interesting to note that results obtained in (3.3.8) and (3.3.9) are identical with those obtained by Takacs [32]. In other words, the state probabilities $\{p_n\}$, $n \ge C$ for the G/M/C queue are uniquely determined for a given set of state probabilities $\{p_n\}$, $0 \le n \le C-1$ by using entropy maximization.

3.3.2 The M/G/C Queue

From the literature [44-46], it is known that the state probabilities $\{p_n\}$, $0 \leq n \leq C-1$ of the M/M/C queue may be used as an approximation to those of the M/G/C queue, where for the M/M/C queue

$$p_n = p_0 \frac{(C\rho)^n}{n!} \quad \text{, for } n = 0,1,...,C-1 \tag{3.3.10}$$

and

$$p_0^{-1} = \sum_{j=1}^{C-1} \frac{(C\rho)^j}{j!} + \frac{(C\rho)^C}{(1-\rho)C!}$$

with

$$\rho = \frac{\lambda}{\mu C} < 1$$

Since the interarrival time distribution for a Poisson input with mean arrival rate λ is given by

$$F_T(t) = 1 - e^{-\lambda t} \tag{3.3.11}$$

it follows from (3.2.9) that

$$y = \frac{\lambda}{\lambda + Ck} \tag{3.3.12}$$

and k must satisfy the characteristic equation

$$\lambda F_Y^*(-k) - \lambda = Ck \tag{3.3.13}$$

Solving for k and substituting for y and $\{p_n\}$, $0 \leq n \leq C-1$ into (3.3.3a)-(3.3.3c),

we can find the β_i, i=1,2 and α_j, j=0,1,..., C-1. Then the state probabilities $\{p_n\}$, $n \geq C$ can be determined from (3.3.2).

Example 3. The M/GE/C queue

For the M/GE/C queue, we know that the Laplace - Stieltjes transforms of the interarrival time and the service time distributions are respectively given by

$$F_T^*(s) = \frac{\lambda}{\lambda + s}$$

and

$$F_Y^*(s) = \frac{2\mu + s(c_Y^2 - 1)}{2\mu + s(c_Y^2 + 1)}$$

From (3.3.13), we find

$$k = \frac{2\mu(1-\rho)}{1 + c_Y^2} > 0$$

Substituting this k into (3.3.12), we find

$$y = \frac{\rho(1 + c_Y^2)}{2(1-\rho) + \rho(1 + c_Y^2)}$$

Then substituting this y and $\{p_n\}$, $0 \leq n \leq C-1$ of (3.3.10) into (3.3.3a)-(3.3.3c) and (3.3.4), from (3.3.2) we obtain the state probabilities

$$p_n = \begin{cases} p_0 \dfrac{(C\rho)^n}{n!}, \text{ if } 0 \le n \le C-1 \\ \left[1 - p_0 \sum_{n=0}^{C-1} \dfrac{(C\rho)^n}{n!}\right]\left(\dfrac{2-2\rho}{2+\rho c_Y^2 - \rho}\right)\left(\dfrac{\rho + \rho c_Y^2}{2+\rho c_Y^2 - \rho}\right)^{n-C} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{, if } n \ge C \end{cases} \tag{3.3.14}$$

where

$$p_0 = \left[\sum_{n=0}^{C-1} \frac{(C\rho)^n}{n!} + \frac{(C\rho)^C}{C!}\left(\frac{1}{1-\rho}\right)\right]^{-1}$$

and the mean number of customers

$$E(N) = \frac{\rho^2 (1 + c_Y^2)}{2(1-\rho)^2} p_{C-1} + C\ \rho$$

For the case of the M/M/C queue where $c_Y = 1$, (3.3.14) gives the state probability distribution

$$p_n = \begin{cases} p_0 \dfrac{(C\rho)^n}{n!}, & n \le C-1 \\ p_0 \dfrac{C^C}{C!}\rho^n, & n \ge C \end{cases}$$

Thus, this result is exact.

3.4 THE WAITING TIME DISTRIBUTION

Let W be a continuous random variable denoting the waiting time with $F_W(t)$ and $f_W(t)$ being its distribution function and density function, respectively. Since the distribution function $F_W(t)$ has a jump of $F_W(0)$ at t=0, thus the probability density function $f_W(t)$ can be written as

$$f_W(t) = F_W(0)\ \delta(t) + f_{W_c}(t) \tag{3.4.1}$$

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

Let the entropy function of $f_W(t)$ be defined as

$$H = -\int_{0^+}^{\infty} f_{W_c}(t) \ln f_{W_c}(t)\, dt \tag{3.4.2}$$

The normalization condition is given by

$$F_W(0) + \int_{0^+}^{\infty} f_{W_c}(t)\, dt = 1 \tag{3.4.3}$$

and the mean of W is given by

$$E(W) = \int_{-\infty}^{\infty} t\, f_W(t)\, dt = \int_{0^+}^{\infty} t\, f_{W_c}(t)\, dt \tag{3.4.4}$$

We shall formulate the problem of finding $f_{W_c}(t)$ as follows :

Maximize H *in* (3.4.2) *with respect* to $f_{W_c}(t)$

subject the constraints (3.4.3) and (3.4.4)

Solving this optimization problem for $f_{W_c}(t)$ leads to

$$f_{W_c}(t) = e^{-1 + \beta_1 + t\beta_2} \tag{3.4.5}$$

where β_1 and β_2 are the Lagrange multipliers associated with (3.4.3) and (3.4.4), respectively.

Then substituting $f_{W_c}(t)$ into (3.4.3) and (3.4.4), we find

$$\beta_1 = 1 + \ln \frac{[1 - F_W(0)]^2}{E(W)}$$

and

$$\beta_2 = \frac{F_W(0) - 1}{E(W)}$$

Substituting β_1 and β_2 into (3.4.5), we obtain

$$f_{W_c}(t) = \frac{[1 - F_W(0)]^2}{E(W)} \exp\left[\frac{-[1 - F_W(0)]t}{E(W)}\right] \tag{3.4.6}$$

Takahashi [41] studied the G/G/C queue and showed that the waiting time density function has an asymptotic exponential tail, i.e.,

$$\frac{f_{W_c}(t+u)}{f_{W_c}(t)} = e^{-Cku} \quad , \text{if } t \text{ is sufficiently large} \tag{3.4.7}$$

where k is the positive root of the characteristic equation (3.2.10)

From (3.4.6) and (3.4.7), we find

$$\frac{1 - F_W(0)}{E(W)} = Ck$$

or

$$F_W(0) = 1 - CkE(W) \tag{3.4.8}$$

where k is determined by (3.2.10).

Applying Little's formula, we have

$$E(W) = \frac{E(N)}{\lambda} - \frac{1}{\mu C} \tag{3.4.9}$$

where E(N) is given in (3.2.6b)

Having found E(W) and $F_W(0)$, (3.4.6) yields the solution for $f_{W_c}(t)$.

In the heavy traffic case, since the probability of zero waiting time $F_W(0)$ is very small, (3.4.6) can be written as

$$f_{W_c}(t) = Ck\ e^{-Ckt} \tag{3.4.10}$$

Then expanding the left-hand side of (3.2.10) in a Taylor series in k about the origin

$$F_T^*(Ck)\,F_Y^*(-k) = 1 - (\frac{C}{\lambda} - \frac{1}{\mu})k + \frac{C^2\,\sigma_T^2\,k^2}{2} + R(k) \tag{3.4.11}$$

where the remainder R(k) is $0(k^2)$ representing the higher order terms of k and σ_T^2, σ_Y^2 are the variances of the interarrival time and service time, respectively.

Using the characteristic equation (3.2.10), we obtain

$$\frac{-2(\frac{C}{\lambda} - \frac{1}{\mu})}{k(\sigma_T^2C^2 + \sigma_Y^2)} + 1 + \left[\frac{2}{\sigma_T^2C^2 + \sigma_Y^2}\right]\frac{R(k)}{k^2} = 0 \tag{3.4.12}$$

Let us consider the situation in which $k \to 0$. Since R(k) equals $0(k^2)$, it follows that as $k \to 0$

$$\frac{2(\frac{C}{\lambda} - \frac{1}{\mu})}{k(\sigma_T^2C^2 + \sigma_Y^2)} = 1$$

or

$$Ck = \frac{2(1-\rho)}{\lambda\left[\sigma_T^2 + \sigma_Y^2/C^2\right]} \tag{3.4.13}$$

Introducing Ck into (3.4.10) for $f_{W_c}(t)$, we obtain

$$f_{W_c}(t) = \frac{2(1-\rho)}{\lambda(\sigma_T^2 + \sigma_Y^2/C^2)}\exp\left[\frac{-2(1-\rho)t}{\lambda(\sigma_T^2 + \sigma_Y^2/C^2)}\right] \tag{3.4.14}$$

and it should be pointed out that this result agrees with Kingman [9] and Kollerstrom [47].

Example 4. Waiting time distribution of the G/GE/1 queue

In order to determine the waiting time density function $f_W(t)$, we first calculate the mean waiting time E(W). Substituting E(N) of (3.2.13) into the Little's formula (3.4.9), we have

$$E(W) = \frac{1}{k} - \frac{(1 + c_Y^2)}{2\mu}$$

Then applying E(W) into (3.4.8) , we find

$$F_W(0) = \frac{k(1 + c_Y^2)}{2\mu}$$

Finally substituting E(W) and $F_W(0)$ into (3.4.6), we obtain

$$f_{W_c}(t) = \frac{(2\mu - k - kc_Y^2)}{2\mu} k\ e^{-k\,t}$$

Note that this result is exact and agrees with that of Kouvatsos [48].

Example 5. Waiting time distribution of the G/M/C queue

To determine the waiting time density function $f_W(t)$, we first find the mean waiting time E(W). Substituting E(N) of (3.3.8) into (3.4.9), we get

$$E(W) = \frac{\alpha}{\mu\, C(1-y)^2}$$

where y is the root of the functional equation $y = F_T{}^*[C\ \mu(1-y)]$.

Then substituting E(W) and $Ck = C\mu(1-y)$ into (3.4.8), we find

$$F_W(0) = 1 - \frac{\alpha}{1-y}$$

Employing the calculated E(W) and $F_W(0)$ into (3.4.6), we get

$$f_{W_c}(t) = \alpha\mu C\ \exp(-\mu C(1-y)t)$$

Finally from (3.4.1), we obtain

$$f_W(t) = \left[\frac{1-y-\alpha}{1-y}\right]\delta(t) + \alpha\mu C\ \exp(-\mu C(1-y)t)$$

It should be pointed that this result is exact and agrees with that of Takacs [32].

Example 6. Waiting time distribution of the M/G/C queue

Let N_q = max(0, n-C) be the steady-state queue length and its generating function be

$$G_{N_q}(z) = \sum_{n=0}^{C-1} p_n + \sum_{n=C}^{\infty} p_n\ z^{n-C} \qquad (3.4.15)$$

Since

$$\sum_{n=0}^{C-1} p_n = F_W(0),$$

Substituting p_n of (3.3.2) into (3.4.15), after simplification, we obtain

$$G_{N_q}(z) = F_W(0) + \frac{Ck[1-F_W(0)]}{\lambda - \lambda z + Ck}$$

$$= F_W{}^*(\lambda(1-z))$$

This agrees with the well-known result for the M/G/C queue [34] and it means that the queue length N_q has the same distribution of the number of customers who arrive during the waiting time W.

3.5 NUMERICAL EXAMPLES

Up to now no explicit exact analytic solution for the G/G/C system has been obtained. It is, however, possible to obtain numerical solutions for a class of interarrival time and service time distributions. Hillier and Yu [49] provided numerical tables and graphs for the state probability distribution of the multiserver queue. Seleen et al [50] recently gave more comprehensive queueing tables for the G/G/C system.

Using these tables, we can numerically examine the accuracy of our approximate formulas for E(N) and the delay probability $1-F_W(0)$. Moreover, we can calculate the coefficient of variation for the number of customers and the waiting time. We shall then examine several combinations of the coefficients of

variation for the service time and the interarrival time; but we present in Figs. 3.1-3.7 only some typical results. The others have a similar degree of accuracy. Fig. 3.1-3.4 show the E(N) and the delay probabilities respectively for (a) the $GE/E_2/C$ queue(C=2,3,5) with c_T^2=2 (b) the E_2/E_2/C queue (C=2,3,5) (c) the GE/GE/C queue (C=2,3,5) with $c_Y^2 = c_T^2$ = 1.5 and (d) the M/GE/C queue (C=2,5) and c_Y^2 = 4 and Figs. 3.5-3.7 display the delay probabilities $1-F_W(0)$ for systems (b), (c) and (d) with various traffic intensities. In these figures, ME1 and ME2 stand for the results obtained from formulation I and II respectively and Q.table means those obtained from queueing tables.

From these figures, the results indicate that the accuracy of the performance measures improves as the number C of servers and the coefficients of variation of service time and interarrival time decrease. From our studies [93], it shows the accuracy of ME1 is acceptable for C ≤ 5 and c_Y^2 and c_T^2 < 10 while in the cases of M/G/C queues, the accuracy of ME2 is higher than that of ME1 especially as C, c_T^2 and c_Y^2 are large. For the G/M/C and the G/GE/1 queue, ME2 is correct as we mentioned before.

3.6 DIFFUSION APPROXIMATION OF MULTISERVER QUEUES

Under heavy traffic conditions, Kingman [51] studied the multiserver queue using approximation methods. He conjectured that the waiting time should be approximately exponentially distributed with a mean depending only on the first two moments of the interarrival time and service time distributions. This conjecture was verified later by Kollerstrom [47]. Iglehart and Whitt [52] obtained several heavy traffic limit theorems for the G/G/C queue. Their results indicated that the basic queueing processes associated with the G/G/C queue can be approximated by Brownian motion processes. On the basis of heavy traffic limit theorems, Halachmi et al [53] and Sunsga et al [54] investigated the G/G/C queue by approximating the discrete valued process of the number of customers as a diffusion process and obtained the state probability distribution. This method is called diffusion approximation and has been widely used for the study of more complicated queueing systems [55-56].

This section deals with the diffusion approximations for the G/G/C queue. We shall incorporate well-known results for the G/G/C queue into our model. Introducing new diffusion parameters, we shall show the accuracy of the approximation can be improved. This is particularly true under light traffic conditions.

3.6.1 The G/G/C Queue - Formulation I

In order to study the G/G/C queue we shall make the following assumptions :

1. The interarrival times of customers are identically distributed, mutually independent random variables with mean $1/\lambda$ and variance σ_T^2.
2. The service times of customers are also identically distributed, mutually independent random variables with mean $1/\mu$ and variance σ_Y^2.
3. The service facilities have infinite waiting rooms and the queue discipline is *FIFO*.

Let N(t) denote the number of customers in the system at time t and $p_k = \lim_{t\to\infty} P\{N(t) = k\}$, k=0, 1,..., be the steady-state probability. The essence of this method is to approximate the discrete random process {N(t)} by a continuous diffusion process {X(t)}.

Condider the one-dimensional homogeneous diffusion process {X(t)}, $t \geq 0$. Let the conditional probability density function of X(t) be defined by

$$p(x,t \mid x_0)\, dx = P\left[x \leq X(t) < x + dx \mid X(0) = x_0\right]$$

Then $p(x,t \mid x_0)$ satisfies the forward Kolmogórov-Fokker- Planck equation [57].

$$\frac{1}{2}\frac{\partial^2}{\partial x^2}\left[a(x)\cdot p(x,t \mid x_0)\right] - \frac{\partial}{\partial x}\left[b(x)p(x,t \mid x_0)\right] = \frac{\partial p}{\partial t} \qquad (3.6.1)$$

where b(x) and a(x) are, respectively, the infinitesimal mean and variance defined by

$$b(x) = \lim_{\Delta t \to 0} \frac{E\left[X(t+\Delta t) - X(t) | X(t) = x\right]}{\Delta t}$$

and

$$a(x) = \lim_{\Delta t \to 0} \frac{Var\left[X(t+\Delta t) - X(t) | X(t) = x\right]}{\Delta t} .$$

Since the primary interest is in the steady-state solution, i.e. $p(x) = \lim_{t \to \infty} p(x,t|x_0)$, the problem is to investigate the equation

$$\frac{1}{2} \frac{d^2}{dx^2} \left\{ a(x)\, p(x) \right\} - \frac{d}{dx} \left\{ b(x)\, p(x) \right\} = 0 \qquad (3.6.2)$$

In order to use (3.6.2) as an approximate model for the G/G/C queue, we shall first study the problem of determining the parameters a(x) and b(x). Next, we shall establish the boundary conditions for the diffusion equation (3.6.2). Finally we shall find the steady-state probabilities $\{p_k\}$ by performing a discretization of the state space.

1. The diffusion parameters

To obtain a diffusion approximation model for the queueing system, the parameters a(x) and b(x) must be specified. Kimura [58] proposed that a(x) and b(x) be piecewise continuous functions.

For k-1 < x < k, k=1, 2,..., C,

$$b(x) = \lambda - \min([x], C)\, \mu \qquad (3.6.3a)$$

$$a(x) = \lambda c_T^2 + \min([x], C)\mu\, c_Y^2 \tag{3.6.3b}$$

where [x] denotes the smallest integer not less than x.

However, since his formulation is not consistent with the law of conservation of traffic given in (3.2.3), we shall introduce the modified parameters

$$b(x) = \left[\lambda - \min([x], C)\,\mu\right] y_0 \tag{3.6.4a}$$

$$a(x) = \left[\lambda\, c_T^2 + \min([x], C)\,\mu c_Y^2\right] y_0 \tag{3.6.4b}$$

and define the parameters

$$b_k = b(x) = (\lambda - k\mu)\, y_0 \quad , k-1 < x < k, k = 1,..., C$$

$$a_k = a(x) = (\lambda c_T^2 + k\mu c_Y^2)\, y_0 \quad , k-1 < x < k \quad , k = 1,..., C$$

$$b_C = b(x) = (\lambda - C\mu)\, y_0 \quad , x > C$$

$$a_C = a(x) = (\lambda c_T^2 + C\mu c_Y^2)\, y_0 \quad , \quad x > C$$

where the parameter y_0 shall be determined later to ensure that the traffic conservation law holds.

2. The boundaries

The boundary condition considered is the absorbing barrier with elementary return [59]. Its property can be explained by the trajectory of X(t). As the trajectory of X(t) reaches the boundary at the origin, it remains there for a random

interval of time. This holding time at the origin also represents the idle period of the G/G/C queue. In the steady state, it has been shown [57] that the behavior of the diffusion process is insensitive to the actual distribution of the holding time in the sense that only its mean value appears in the diffusion equation. Let h_0 denote the mean holding time of the diffusion process {X(t)} at the boundary. Since for general G/G/C queues, there is no exact way to obtain this measure, we shall approximate the mean holding time by the average residual interarrival time which is expressed as [57]

$$h_0 = \frac{2\lambda}{1 + c_T^2}$$

Thereafter, the trajectory jumps to an interior point x of the state space with the probability density function $f_0(x)$. Since for the G/G/C queue, this must happen after an arrival to an empty system, the number of customers in the system increases instantaneously by one. Thus we define

$$f_0(x) = \delta(x - 1) \tag{3.6.5}$$

From Feller [59], the steady state probability density function p(x) satisfies the equation

$$\frac{1}{2}\frac{d^2}{dx^2}[a(x)\,p(x)] - \frac{d}{dx}[b(x)\,p(x)] = -h_0\,p_0\,\delta(x-1) \quad , x>0 \tag{3.6.6}$$

where p_0 denotes the probability mass at the origin and satisfies the equation

$$\frac{1}{2}\frac{d}{dx}[a(x)\,p(x)] - b(x)\,p(x)|_{x=0} = h_0 p_0 \tag{3.6.7}$$

It is well known that conservation of probability leads to the condition

$$p_0 + \int_{0^+}^{\infty} p(x)\,dx = 1 \tag{3.6.8}$$

Further the following conditions are required

$$\lim_{x\to 0} p(x) = 0 \tag{3.6.9}$$

$$\lim_{x\to\infty} p(x) = 0 \tag{3.6.10}$$

and

$$\lim_{x\to\infty} \frac{dp(x)}{dx} = 0$$

3. Discretization

In order to obtain the probabilities p_k, k=1,2,..., there exist some variants for such a discretization [14], we shall adopt the following one

$$p_k = P\{N(t) \le k\} - P\{N(t) \le k-1\}$$

$$= \int_{k-1}^{k} p(x)\,dx \ , \ k=1,2,... \tag{3.6.11}$$

We shall proceed to solve (3.6.6) subject to the constraints (3.6.7)-(3.6.10). It should be pointed out that there exists a continuous solution of (3.6.6) even if the functions a(x) and b(x) are piecewise continuous with a finite number of

discontinuities [22].

Integrating (3.6.6) and using (3.6.7), we obtain

$$\frac{1}{2}\frac{d}{dx}[a(x)\,p(x)] - b(x)\,p(x) = h_0\,p_0[1 - U(x-1)] \quad , x \geq 0 \tag{3.6.12}$$

where $u(\bullet)$ denotes the unit step function. By introducing

$$p(x) = \begin{cases} p_k(x) & \text{if } k-1 < x \leq k,\ k = 1,\ldots,\ C \\ p_{C+1}(x) & \text{if } x > C \end{cases} \tag{3.6.13}$$

and using (3.6.4), (3.6.12) becomes

$$\frac{1}{2}\,a_1\,y_0\,\frac{d}{dx}\,p_1(x) - b_1\,y_0\,p_1(x) = h_0\,p_0 \quad , \text{if } 0 < x \leq 1 \tag{3.6.14a}$$

$$\frac{1}{2}\,a_k\,y_0\,\frac{d}{dx}\,p_k(x) - b_k\,p_k(x)\,y_0 = 0 \ , \text{if } k-1<x\leq k,\ k=2,\ldots,C \tag{3.6.14b}$$

$$\frac{1}{2}\,a_C\,y_0\,\frac{d}{dx}\,p_{C+1}(x) - b_C\,y_0\,p_{C+1}(x) = 0 \quad , \text{if } x > C \tag{3.6.14c}$$

Solving (3.6.14a) with the boundary condition (3.6.9), we have

$$p_1(x) = \begin{cases} \dfrac{p_0\,h_0}{y_0\,b_1}(e^{2b_1 x/a_1} - 1), \text{ if } b_1 \neq 0 \\ \dfrac{2p_0\,h_0 x}{a_1\,y_0}, \text{ if } b_1 = 0 \end{cases} \tag{3.6.15}$$

To solve (3.6.14b) and (3.6.14c), it is necessary to impose C additional conditions for determining the unknown integrating constants. Because of the

continuity of p(x), we impose the following smoothing conditions

$$\lim_{x \to k-1} p_k(x) = \lim_{x \to k-1} p_{k-1}(x) \quad , \text{if } k = 2, 3,..., C+1 \tag{3.6.16}$$

Solving (3.6.14) with (3.6.16), we obtain the solution

$$p_k(x) = p_1(1) \prod_{j=2}^{k} e^{2b_j/a_j} e^{2b_k (x-k)/a_k} \quad , k = 2,...,C+1 \tag{3.6.17}$$

where we set $a_{C+1} = a_C$ and $b_{C+1} = b_C$, and $p_1(1)$ is obtained from (3.6.15).

Also from (3.6.8) the state probability p_0 is given by

$$p_0 = \left[1 - \frac{h_0}{y_0 b_1} + \sum_{k=1}^{C-1} \left(\frac{a_k}{2b_k} - \frac{a_{k+1}}{2b_{k+1}} \right) q_k \right]^{-1} \tag{3.6.18}$$

where

$$q_1 = \frac{h_0}{y_0 b_1} (e^{2b_1/a_1} - 1)$$

$$q_k = q_1 \prod_{j=2}^{k} e^{2b_j/a_j} \quad , k = 2, 3,..., C+1$$

Applying $p_k(x)$ of (3.6.15) and (3.6.17) into (3.6.11) yields

$$p_1 = \begin{cases} \dfrac{p_0 h_0}{y_0 b_1} \left[\dfrac{a_1 (e^{2b_1/a_1} - 1)}{2b_1} - 1 \right], & \text{if } b_1 \neq 0 \\ \dfrac{p_0 h_0}{y_0 a_1}, & \text{if } b_1 = 0 \end{cases} \tag{3.6.19}$$

For k = 2,..., C,

$$p_k = \begin{cases} \dfrac{p_0 a_k (q_k - q_{k-1})}{2b_k}, & \text{if } b_k \neq 0 \\ p_0 q_k, & \text{if } b_k = 0 \end{cases} \tag{3.6.20}$$

For k = C+1, C+2,...,

$$p_k = p_0 a_C \frac{(q_{C+1} - q_C)}{2b_C} \exp[2b_C (k - C - 1)/a_C] \tag{3.6.21}$$

The next step is to determine the unknown factor y_0 from the equation of traffic conservation which is given by

$$\sum_{n=0}^{C-1} n \mu p_n + \sum_{n=C}^{\infty} C \mu p_n = \lambda \tag{3.6.22}$$

Introducing p_k of (3.6.19) and (3.6.20) into (3.6.22) yields

$$y_0 = \frac{z}{\rho} + \frac{h_0}{b_1} [1 - (e^{2b_1/a_1} - 1) \Xi] \tag{3.6.23}$$

where

$$\Xi = \sum_{k=1}^{C-1} \left[\left(\frac{a_k}{2b_k} - \frac{a_{k+1}}{2b_{k+1}} \right) \frac{q_k}{q_1} \right]$$

$$z = \frac{h_0}{C\, b_1} \left[\frac{a_1}{2b_1} (e^{2b_1/a_1} - 1) - 1 \right] + \frac{h_0}{C\, b_1} (e^{2b_1/a_1} - 1) \sum_{k=2}^{C} k \, \frac{a_k}{2b_k} \left[\frac{q_k}{q_1} - \frac{q_{k-1}}{q_1} \right] + \frac{\frac{a_C\, h_0}{2b_C\, b_1} (e^{2b_1/a_1} - 1) \left[\frac{q_{C+1}}{q_1} - \frac{q_C}{q_1} \right]}{1 - e^{2b_C/a_C}}$$

Furthermore, the average number of customers is obtained as

$$E(N) = \sum_{k=0}^{\infty} k\, p_k = \sum_{k=0}^{C} k\, p_k - p_0 \frac{a_C}{2b_C} q_C \left[C + \frac{1}{1 - e^{2b_C/a_C}} \right] \qquad (3.6.24)$$

For the single server case, i.e. C=1, (3.6.23) reduces to

$$y_0 = \frac{2}{1 + c_T^2}$$

and (3.6.18) reduces to

$$p_0 = 1 - \rho$$

3.6.2 The G/G/C/ Queue - Formulation II

It is seen from (3.2.8) that the ratio of two neighbouring state probabilities of the G/G/C system asymptotically tends to a constant y given in (3.2.9). However, from the solution (3.6.21) of the diffusion equation, we find

$$\frac{p_{n+1}}{p_n} = \exp(2b_C/a_C) \quad , \text{if } n \geq C \tag{3.6.25}$$

In order to make the result (3.6.25) consistent with (3.2.8), we shall let $\exp(2b_C/a_C)$ equal to y or

$$\frac{2b_C}{a_C} = \ln y \tag{3.6.26}$$

while parameters a_k, b_k, k=1,..., C-1 remain unchanged.

Then by substituting these new parameters into (3.6.17) and (3.6.18), the state probability distribution can be calculated from (3.6.20) and (3.6.21).

The relationship between the quantity $\frac{2b_C}{a_C}$ and $\ln y$ is now investigated. From (3.2.9), we know

$$y = F_T{}^*(Ck)$$

where k must satisfy (3.2.10). And it is seen from (3.4.13) that as $k \to 0$ we have

$$k = \frac{2(\frac{C}{\lambda} - \frac{1}{\mu})}{\sigma_T^2 C^2 + \sigma_Y^2} \tag{3.6.27}$$

We shall show that as $k \to 0$, $\lim_{k \to 0} \ln y = \frac{2b_C}{a_C}$. From (3.6.27), we see that as

$k \to 0$, $\lambda \to C\mu$ *or* $\rho \to 1$. Further, since

$$\ln y = \ln F_T^*(Ck)$$

expanding $\ln F_T^*(Ck)$ as a Taylor series in k about the origin, we obtain

$$\ln y = \ln F_T^*(Ck) = \frac{Ck}{\lambda} + 0(k)$$

Under heavy traffic condition, as $k \to 0$, this means

$$\frac{Ck}{\lambda} = \frac{-2C(\frac{1}{\mu} - \frac{C}{\lambda})}{\lambda(\sigma_T^2 C^2 + \sigma_Y^2)} = \frac{2b_C}{a_C}$$

3.6.3 Numerical Results

Consider a C-server queueing system with various interarrival and service time distributions. The quanty of interest is the mean number of customers in the system. Tables 3.1-3.3 show the E(N) for (i) the M/M/C queue (ii) the M/GE/C queue with $c_Y^2 = 2.5$ and (iii) the GE/GE/C queue with $c_T^2 = 2$ and $c_Y^2 = 4$. In these tables, approx. I and approx. II stand for the results obtained from the formulation I and II using the method of diffusion approximation respectively.

From these tables, the results indicate that the accuracy of the performance measure using approx. II is better than that of using approx. I, especially as the coefficients of variation of the interarrival time and service time are large. The accuracy of approx. I and II is much better than the results obtained without

considering the conservation of traffic and asymptotic property of G/G/C queues in all cases.

3.7 SUMMARY

The G/G/C queue has been investigated by means of entropy maximization. The steady-state probability distribution and the waiting time distribution have been obtained. They are exact for the G/GE/1 queue and G/M/C queue for the ME2 formulation. In comparison of analytic and simulation studies, the results showed that the ME2 is better than the ME1 for the M/G/C queue, especially when the number of servers and the coefficients of variation of the service time and the interarrival time are large. Moreover, by using the method of diffusion approximation incorporating the law of traffic conservation and Takahashi's identity, the accuracy of estimating the mean number of customers in the G/G/C queue has been greatly improved.

From the results of Table 3.1 to Table 3.3, it can be seen that at light utilizations improvement in the mean number of customers, E(N), by approx. II can reduce the error from 35% (without flow balance) down to 6% for the M/M/C queue and from over 100% down to less than 50% for the GE/GE/C queue. Furthermore, at heavy utilizations approx. II yields better improvements and the error in E(N) can be reduced to about 15% only.

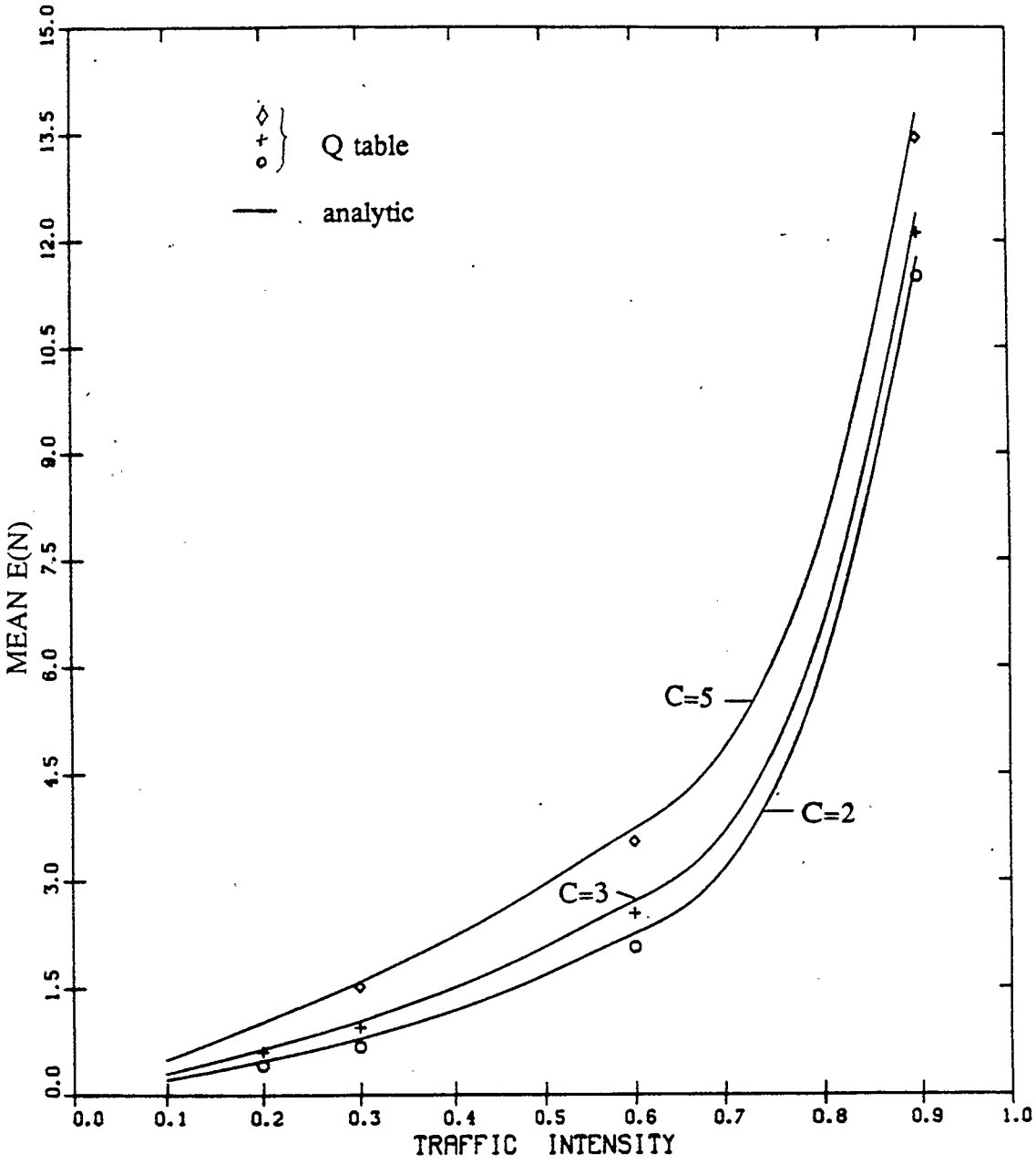


Fig. 3.1 Mean E(N) of the GE/E_2/C queue (c_T^2=2, c_Y^2=0.5)

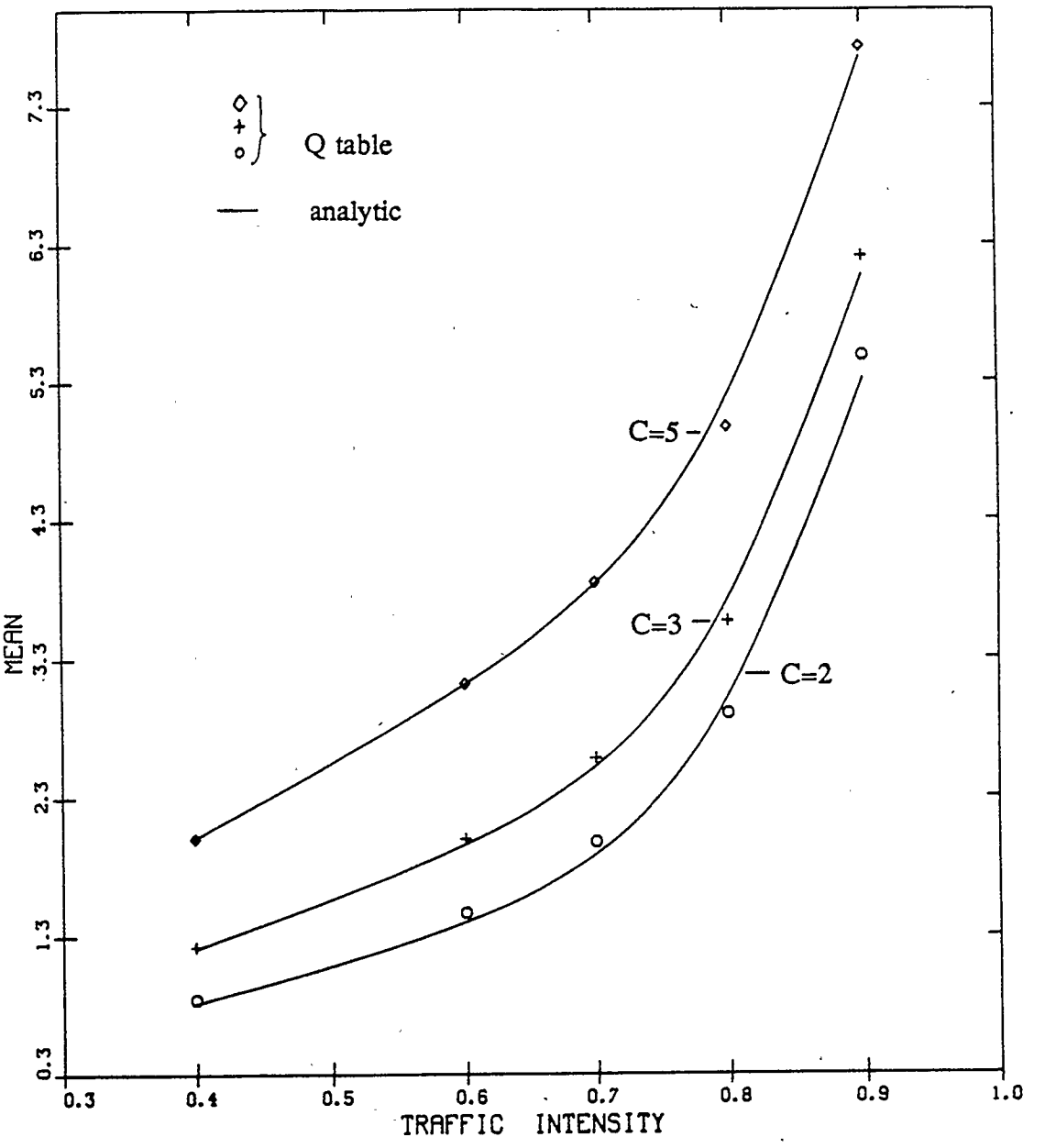


Fig. 3.2 Mean E(N) of the E_2/E_2/C queue

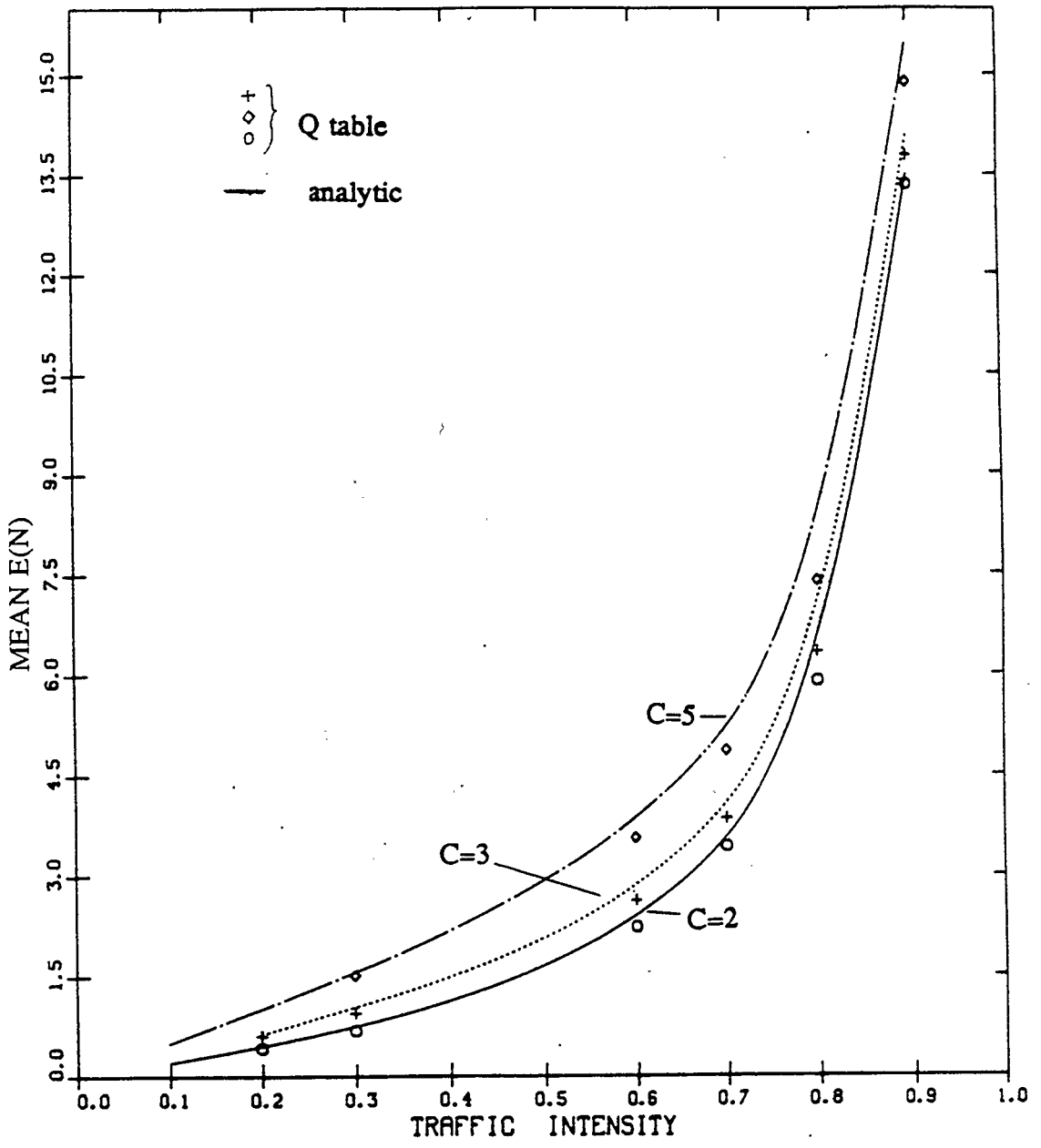


Fig. 3.3 Mean E(N) of the GE/GE/C queue ($c_T^2=c_Y^2=1.5$)

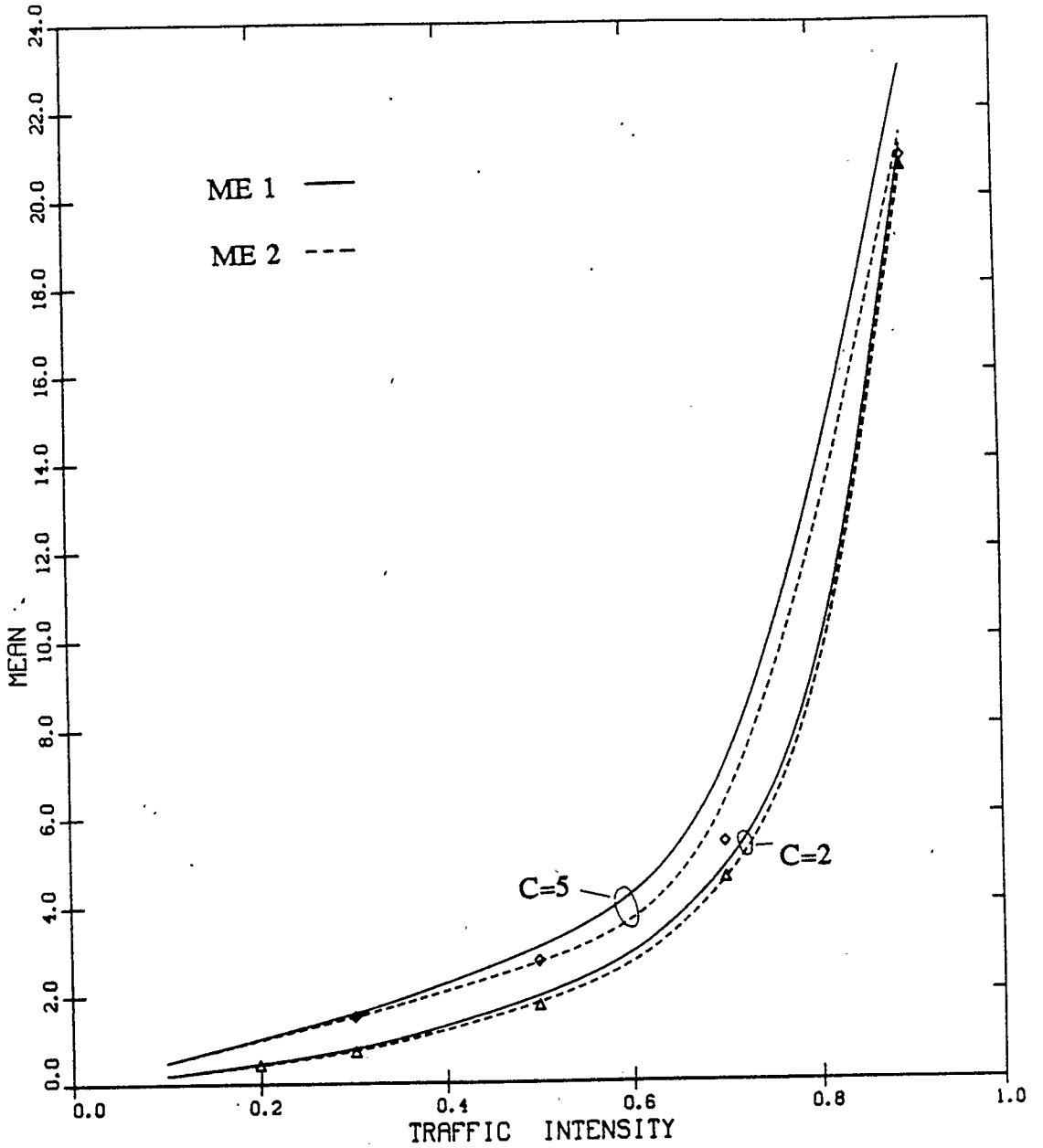


Fig. 3.4 Mean E(N) of the M/GE/C queue ($c_Y{}^2$=4)

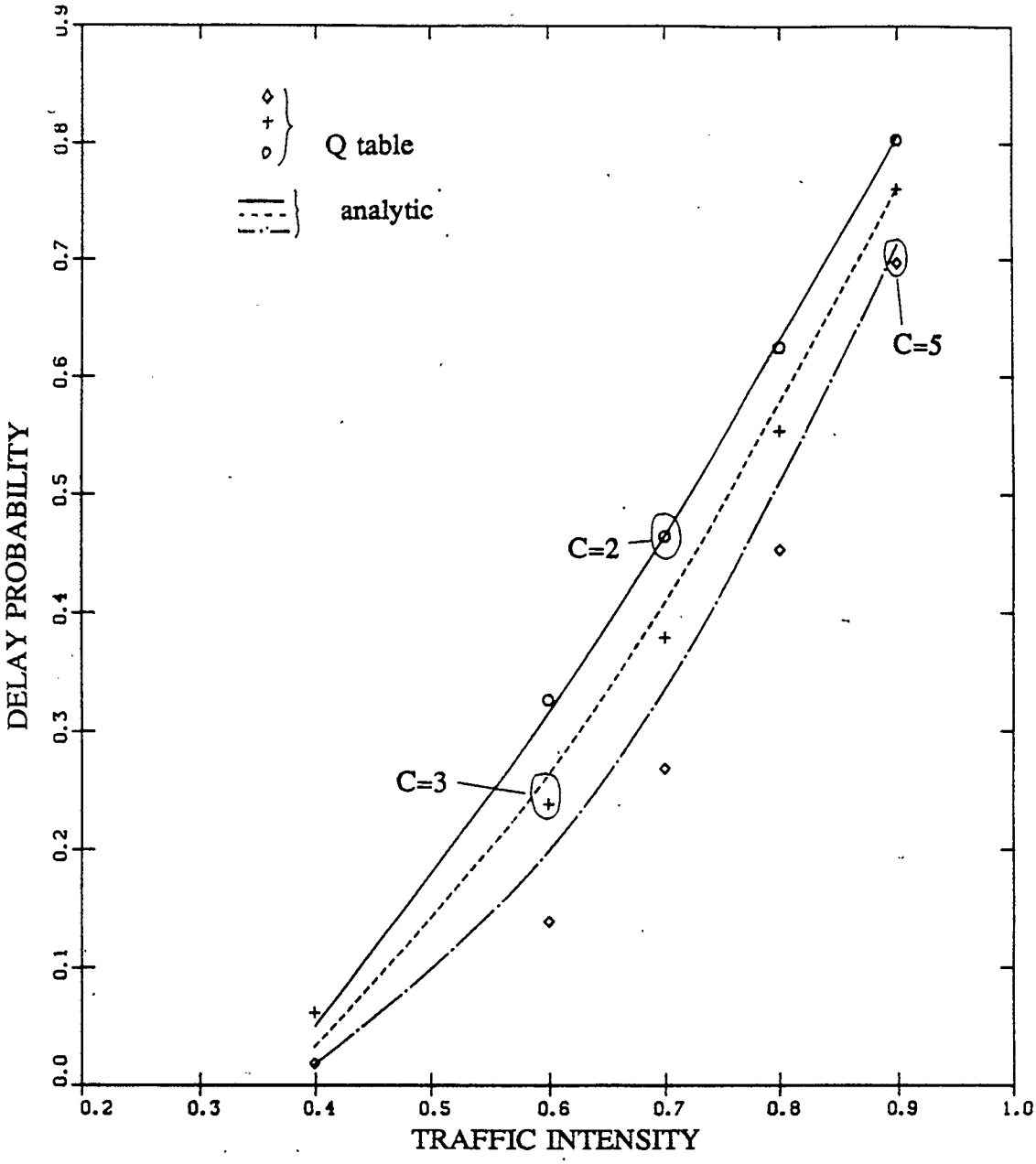


Fig. 3.5 Delay probability of the E_2/E_2/C queue

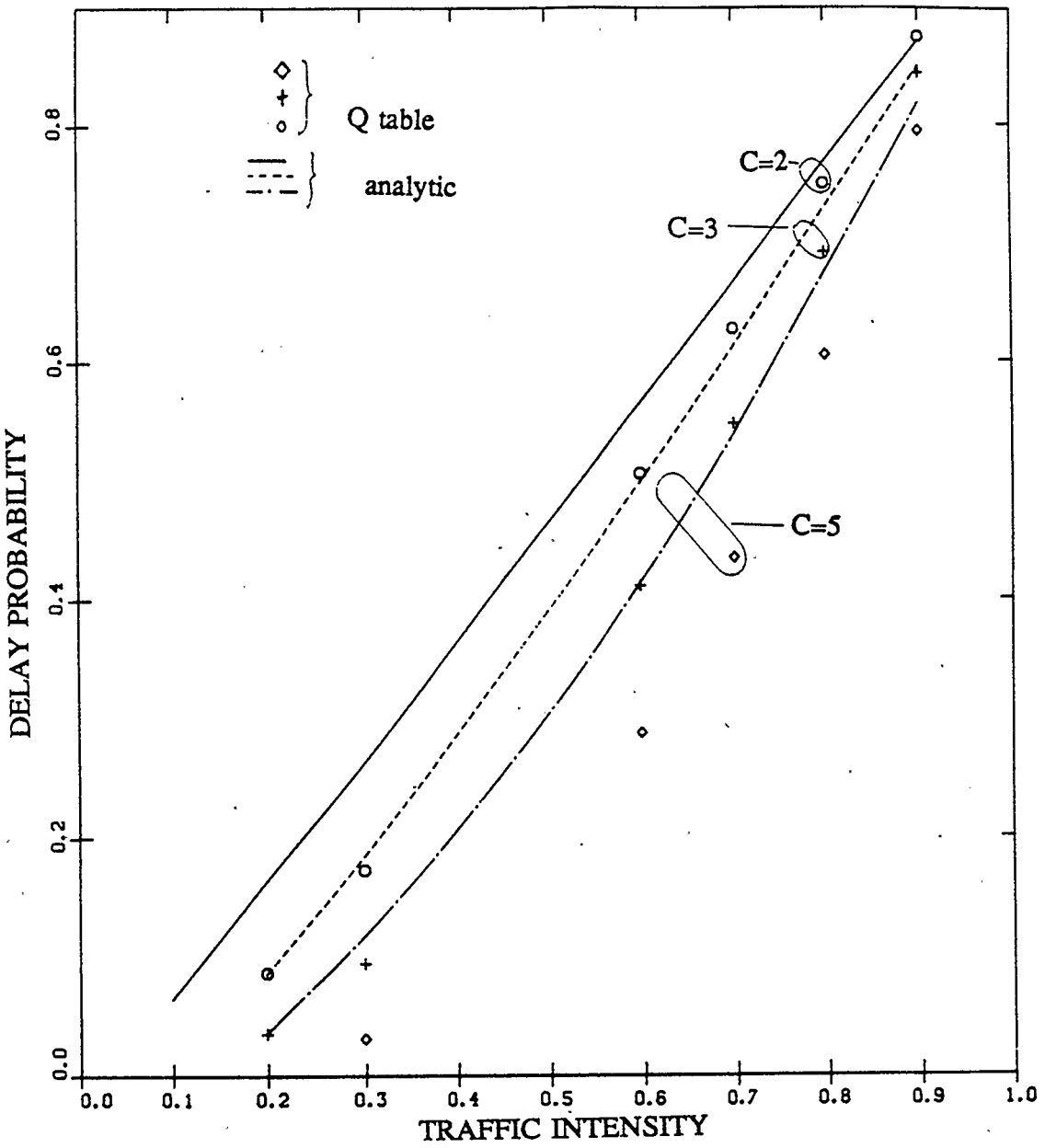


Fig. 3.6 Delay probability of the GE/GE/C queue ($c_T^2 = c_Y^2 = 1.5$)

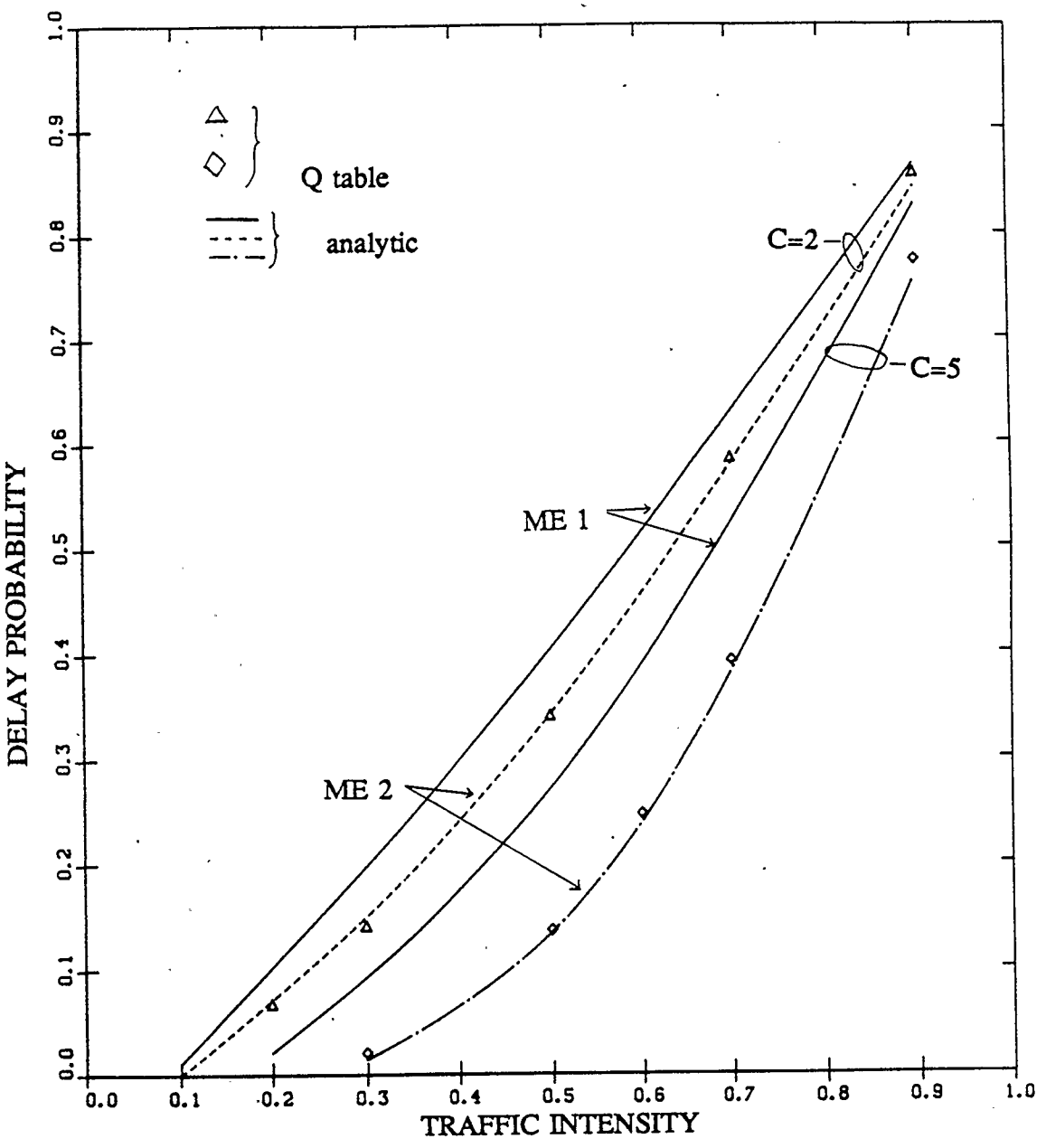


Fig. 3.7 Delay probability of the M/GE/C queue ($c_Y{}^2$=4)

Table 3.1 Mean number of customers in the M/M/C queue

(a) C = 2

ρ	without flow balance	approx. I	approx. II	Q table
0.3	0.909	0.719	0.696	0.659
0.4	1.262	1.041	1.015	-
0.6	2.278	2.024	2.001	1.875
0.8	4.917	4.657	4.643	-
0.9	9.975	9.719	9.711	9.474

(b) C = 3

ρ	without flow balance	approx. I	approx. II	Q table
0.2	0.845	0.621	0.613	0.606
0.4	1.671	1.349	1.330	1.294
0.6	2.786	2.444	2.422	2.321
0.8	5.487	5.166	5.151	4.989
0.9	10.56	10.26	10.25	10.05

(c) C = 5

ρ	without flow balance	approx. I	approx. II	Q table
0.3	1.912	1.520	1.515	1.509
0.5	3.108	2.681	2.667	2.631
0.7	4.889	4.503	4.486	4.382
0.9	11.88	11.58	11.57	11.36

Table 3.2 Mean number of customers in the M/GE/C queue ($c_Y{}^2$=2.5)

(a) C = 2

ρ	without flow balance	approx. I	approx. II	Q table
0.2	0.902	0.614	0.466	0.425
0.3	1.395	1.009	0.784	0.691
0.7	5.234	4.647	4.114	3.665
0.9	17.07	16.45	15.77	15.09

(b) C = 3

ρ	without flow balance	approx. I	approx. II	Q table
0.2	1.222	0.747	0.636	0.608
0.3	1.808	1.203	1.018	0.943
0.5	3.213	2.470	2.120	1.864
0.7	5.828	5.048	4.529	3.984
0.9	17.81	16.94	16.26	15.34

(c) C = 5

ρ	without flow balance	approx. I	approx. II	Q table
0.3	2.546	1.675	1.559	1.511
0.5	4.166	3.210	2.924	2.687
0.7	6.930	6.024	5.537	4.889
0.9	18.92	18.13	17.45	16.16

Table 3.3 Mean number of customers in the GE/GE/C queue (c_T^2=2, c_Y^2=4)

(a) C = 2

ρ	without flow balance	approx. I	approx. II	Q table
0.2	0.914	0.817	0.616	0.448
0.3	1.509	1.382	1.080	0.772
0.6	4.836	4.675	4.070	3.195
0.8	12.37	12.22	11.41	10.10
0.9	27.38	27.24	26.33	24.78

(b) C = 3

ρ	without flow balance	approx. I	approx. II	Q table
0.2	1.243	0.933	0.756	0.617
0.3	1.962	1.553	1.275	0.985
0.6	5.575	4.992	4.399	3.349
0.7	8.173	7.560	6.862	5.477
0.8	13.26	12.62	11.82	10.08

(c) C = 5

ρ	without flow balance	approx. I	approx. II	Q table
0.3	2.757	1.969	1.747	1.525
0.5	5.049	4.079	3.635	2.872
0.7	9.469	8.435	7.757	6.062
0.9	29.789	28.758	27.851	24.896

CHAPTER 4

QUEUEING NETWORKS

4.1 INTRODUCTION

Mathematical modeling using the theory of queues provides a very important means for investigation of computer network performance, whether for predicting the behavior of new designs or for studying proposed changes to existing systems. However, despite the advances made in the analysis of queueing models, no exact solution has been obtained for queueing networks with FIFO discipline and group arrivals. Computer network traffics are usually bursty [60] and may not be adequately modeled by a Poisson process. However, from the measurements reported in the literature on the message characteristics of computer communication networks , it has been found that Poisson message arrivals with geometric length represent a fairly realistic model for the message statistics [61].

Previous studies mainly focused on single queueing system with group arrivals [38,62,63]. However, in this chapter we shall be concerned with the analysis of networks of queues with group arrivals by using the entropy maximization method. In Section 2 we shall address the analysis of queueing networks of queue with single and multiple-server stations and bulk Poisson input. An expression for the state probability distribution of networks subject to constraints on the mean number of packets and utilization factors will be established. In Section 3 we deal with a queueing network model with window-

based flow control mechanism. A solution technique shall be presented. In Section 4, a solution for queueing networks with G/G/C stations and arbitrary interconnections is provided.

4.2 BURSTY TRAFFIC MODELING OF OPEN NETWORKS

In the literature it is usually assumed that the network traffic for analytic studies is completely characterized by a message interarrival time distribution and a message length distribution [3]. However, the message consists only of a single packet. In the following we shall deal with the message consisting of a random number of packets. The main objective of our study is the application of the entropy maximization method for deriving the packet interarrival time distribution. In Chapter 2 we have shown that the packet interarrival time distribution of a bulk Poisson process is specified by a GE distribution. In this section we shall extend the results from the single queue case to open queueing networks, where arrival processes which are initially bulk Poisson become less and less clustered from node to node in the network.

4.2.1 Open Networks with Single - Server Nodes

Consider an arbitrary open network at equilibrium with L nodes each containing a single server with infinite capacity and satisfying the following conditions :

1. External arrivals constitute a bulk Poisson process , where the group arrival is Poisson-distributed with arrival rate λ_{g_i} and group size is geometrically distributed with mean $1/p_i$;

2. Service times at node i are mutually independent and identically distributed with mean $1/\mu_i$ and coefficient of variation c_{Y_i};

3. Upon service completion at node i, a packet requests service from node j with probability r_{ij}, i, j = 1,..., L or leaves the system with probability $r_{i,L+1} = 1 - \sum_{j=1}^{L} r_{ij}$.

4. The queue discipline is first-come, first-served (FIFO) at all nodes.

Let λ_i be the arrival rate of packets entering node i. Then λ_i satisfies the flow balance equations of traffic

$$\lambda_i = \frac{\lambda_{g_i}}{p_i} + \sum_{j=1}^{L} \lambda_j \, r_{ji} \qquad (4.2.1)$$

The utilization factor ρ_i of the server at node i is given by

$$\rho_i = \frac{\lambda_i}{\mu_i} < 1 \qquad (4.2.2)$$

The state of the network can be described by the integer valued vector $\mathbf{n}=(n_1, n_2, \ldots, n_L)$, where n_i is the number of packets at node i. Let p(**n**) be the steady-state probability that the system is in state **n**, and $p_i(n_i)$ be the marginal

equilibrium probability that node i is in state n_i, i.e.,

$$p_i(n_i) = \sum_{\substack{all\ n_l \\ 1 \le l \le L\ ,\ l \ne i}} p(\mathbf{n}) \tag{4.2.3}$$

Let c_{T_i} be the coefficient of variation of the packet interarrival time for node i and let node i be modeled as a GE/G/1 queue. Then from (2.3.10) by substituting $p=2/(1+c_T^2)$, the average number of packets at node i, $E(N_i)$, is given by

$$E(N_i) = \frac{\rho_i\ (1 + c_{T_i}{}^2 - \rho_i + \rho_i c_{Y_i}{}^2)}{2(1-\rho_i)} \tag{4.2.4}$$

To determine the state probability distribution $p(\mathbf{n})$ of the network, we formulate the optimization problem as follows.

$$\textit{Maximize}\quad H = -\sum_{\mathbf{n}} p(\mathbf{n}) \ln p(\mathbf{n})$$

subject to the constraints:

$$\textit{Means}\qquad \sum_{\mathbf{n}} n_i\ p(\mathbf{n}) = E(N_i)$$

$$\textit{Utilizations}\quad 1 - p_i(n_i = 0) = \rho_i\ ,\ i=1, 2,..., L$$

and

$$\textit{Normalization}\ \sum_{\mathbf{n}} p(\mathbf{n}) = 1$$

It follows from (2.2.8) and (2.2.9) that the solution is given by

$$p(\mathbf{n}) = \prod_{i=1}^{L} p_i(n_i) \tag{4.2.5}$$

where

$$p_i(n_i) = \begin{cases} 1 - \rho_i \text{ , if } n_i = 0 \\ (1 - \rho_i) a_{i_1} a_{i_2}^{n_i} \text{ , if } n_i \geq 1 \end{cases}$$

$$a_{i_1} = \frac{\rho_i^2}{(1-\rho_i)[E(N_i)-\rho_i]}$$

and

$$a_{i_2} = \frac{E(N_i) - \rho_i}{E(N_i)}$$

The significance of the above solution (4.2.5) is that each node of the network can be modeled as a G/G/1 queue in a manner similar to Jackson's theorem [64] for exponential queueing networks, in which the interarrival times and the service times have exponential distributions. However, performance measures of each node depend on the parameters of arrival and departure processes such as their means and variances. These parameters shall be determined as follows.

(a) The output process

For a GE/G/1 queue suppose that the means and coefficients of variation of the interarrival time and the service time are specified respectively by $1/\lambda_i$, $c_{T_i}^2$

and $1/\mu_i$, $c_{Y_i}^2$. Then the coefficient of variation of the interdeparture time is given by [48]

$$c_{D_i} = \left[\rho_i + (1 - \rho_i)c_{T_i}^2 + (c_{Y_i}^2 - 1)\rho_i^2 \right]^{1/2} \tag{4.2.6}$$

(b) The decomposition process

Consider a traffic stream i which is split into L components according to probabilities r_{ij}, j=1,...,L. Sevcik et al [65] showed that the means and coefficients of variation of the interarrival time of the traffic stream j are given by

$$\frac{1}{\lambda_{i\,j}} = \frac{1}{\lambda_i\, r_{i\,j}} \tag{4.2.7}$$

and

$$c_{T_{ij}} = \sqrt{1 + r_{i\,j}(c_{T_i}^2 - 1)} \quad , j = 1,2,..., L \tag{4.2.8}$$

(c) The superposition process

Consider L traffic streams combined together to form a single stream j. Let the interarrival time distribution of component i be a GE distribution of the form

$$F_{T_{ij}}(t) = 1 - \frac{2}{1 + c_{T_{ij}}^2} \exp\left[\frac{-2\lambda_{ij}t}{1 + c_{T_{ij}}^2} \right] \quad , i = 1,2,..., L \tag{4.2.9}$$

Kouvatsos [15] showed that the overall interarrival time distribution $F_{T_j}(t)$ is still a GE distribution with

$$F_{T_j}(t) = 1 - \frac{2}{1 + c_{T_j}^2} \exp\left[\frac{-2\lambda_j t}{1 + c_{T_j}^2}\right] \tag{4.2.10}$$

where

$$\lambda_j = \sum_{i=1}^{L} \lambda_{ij} \quad , j=1,2,\ldots L \tag{4.2.11}$$

and

$$c_{T_j} = \left[\frac{\lambda_j}{\sum_{i=1}^{L} \frac{\lambda_{ij}}{1 + c_{T_{ij}}^2}} - 1\right]^{1/2} \tag{4.2.12}$$

Example 1. Consider a queueing model with the following specifications.

- The input process consists of L identical bulk Poisson streams with message arrival rate λ_g message/sec and geometrically distributed message length with mean $1/p$ packets/message;
- Service time distributions are exponential with mean $1/\mu$;
- Queue discipline is FIFO.

The quantity of interest is the average number of packets in the system.

Using the equivalence equation (2.4.9) we have an input of L GE streams, each with packet arrival rate λ_g/p and variation coefficient of the interarrival time $\sqrt{2/p - 1}$. Then from (4.2.10) the overall interarrival time distribution $F_T(t)$ is a

GE distribution,

$$F_T(t) = 1 - \frac{2}{1 + {c_T}^2} \exp\left[\frac{-2\lambda t}{1 + {c_T}^2}\right]$$

where

$$\lambda = \frac{\lambda_g L}{p}$$

$${c_T}^2 = \frac{2}{p} - 1$$

From (4.2.4) with $c_Y = 1$, the average number of packets in the system is given by

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

where $\rho = \lambda_g \frac{L}{\mu p}$.

This result is correct and agrees with the result obtained by Spirn [91].

4.2.2 Open Networks with Multiserver Nodes

Consider open queueing networks at equilibrium with assumptions 1, 3, 4 in Section 4.2.1 and an additional assumption

5. Node i consists of C_i servers. The service times at node i are exponentially and identically distributed with mean $1/\mu_i$, i=1,..., L.

To find the joint state probability distribution of the network, we formulate the optimization problem as follows.

$$\textit{Maximize} \quad H = -\sum_{\mathbf{n}} p(\mathbf{n}) \ln p(\mathbf{n})$$

subject to constraints :

$$\textit{Means} \quad \sum_{\mathbf{n}} n_i \, p(\mathbf{n}) = E(N_i)$$

$$\textit{State dependent utilizations} \quad \sum_{\substack{all\ n_l,\ l \neq i \\ l=1,\ldots,L}} \sum_{n_i=C_i}^{\infty} p(\mathbf{n}) = p_i(n_i), \; 0 \leq n_i \leq C_i - 1$$

$$\textit{Normalization} \quad \sum_{\mathbf{n}} p(\mathbf{n}) = 1$$

From (3.3.8) and (3.3.7), the means and the state probabilities are given by

$$E(N_i) = \frac{\alpha_i \rho_i + C_i \rho_i (1 - y_i)^2}{(1 - y_i)^2} \tag{4.2.13a}$$

$$\textit{with} \quad \rho_i = \frac{\lambda_i}{\mu_i \, C_i}$$

and

$$p_i(n_i) = \begin{cases} p_i(0) \text{ , if } n_i = 0 \\ \dfrac{p_i(0)}{n_i \, !} \displaystyle\prod_{r=0}^{n_i - 1} \frac{2\rho_i C_i + r(c_{T_i}{}^2 - 1)}{c_{T_i}{}^2 + 1} \text{ , if } 1 \leq n_i \leq C_i - 1 \end{cases} \tag{4.2.13b}$$

with

$$p_i(0) = \left[1 + \sum_{k=1}^{C_i-1} \prod_{r=0}^{k-1} \frac{2\rho_i C_i + r(c_{T_i}^2 - 1)}{k!\,(1+c_{T_i}^2)} + \prod_{r=0}^{C_i-1} \frac{2\rho_i C_i + r(c_{T_i}^2 - 1)}{2(1-\rho_i)C_i!} \right]^{-1}$$

$$y_i = \frac{c_{T_i}^2 - 1 + 2\rho_i}{1 + c_{T_i}^2}$$

$$\alpha_i = \frac{p_i(0)}{C_i!} \prod_{r=0}^{C_i-1} \frac{y_i[2\rho_i C_i + r(c_{T_i}^2 - 1)]}{\rho_i(1 + c_{T_i}^2)}$$

$$\rho_i = \frac{\lambda_i}{\mu_i C_i}$$

It follows from (2.2.8) and (2.2.9) that the solution is given by

$$p(\mathbf{n}) = \prod_{i=1}^{L} p_i(n_i) \qquad (4.2.14)$$

where

$$p_i(n_i) = \begin{cases} p_i(0) \text{ , if } n_i = 0 \\ \dfrac{p_i(0)}{n_i!} \prod_{r=0}^{n_i-1} \dfrac{2\rho_i C_i + r(c_{T_i}{}^2 - 1)}{1 + c_{T_i}{}^2} \text{ , if } 1 \le n_i \le C_i \\ p_i(C_i) \left[\dfrac{c_{T_i}{}^2 - 1 + 2\rho_i}{1 + c_{T_i}{}^2} \right]^{n_i - C_i} \text{ , if } n_i \ge C_i \end{cases}$$

To implement the solution (4.2.14) , we shall determine the interdeparture time distribution of the GE/M/C queue.

Heffes [66] obtained the interdeparture time distribution of the H_2/M/C queue. The H_2 stands for a two-phase hyperexponential distribution which is given by

$$F_T(t) = 1 - a_1 e^{-\lambda_1 t} - a_2 e^{-\lambda_2 t} \qquad (4.2.15)$$

where

$$a_2 = \frac{\lambda_2(\lambda_1 - \lambda)}{\lambda(\lambda_1 - \lambda_2)}$$

$$a_1 = 1 - a_2$$

and λ is the average arrival rate. Since the GE distribution is a limiting case of H_2 if the parameter λ_1 approaches infinity, it is possible to obtain the result of the GE/M/C queue in terms of the Laplace transform of the interdeparture time

distribution.

Adapted from Heffes [66], the Laplace-Stieltjes transform of the interdeparture time distribution $F_{D_i}{}^*(s)$ of queue i is given by

$$F_{D_i}{}^*(s) = 1 - s \sum_{k_i=0}^{C_i} H_{k_i}(s) - \frac{sA_i\, y_i}{(1-y_i)(s + C_i\mu_i)} \tag{4.2.16}$$

where

$$y_i = \frac{c_{T_i}{}^2 - 1 + 2\rho_i}{1 + c_{T_i}{}^2}$$

$$A_i = \frac{y_i\, p_i(C_i)}{\rho_i}$$

$$H_{k_i}(s) = \left(1+\left[\frac{c_{T_i}{}^{2}-1}{c_{T_i}{}^{2}+1}\right]^2\right) \sum_{n_i=0}^{k_i} p_i{}^+(n_i) \prod_{j=n_i}^{k_i} \frac{\lambda_1^{\,k_i-j}}{\left(s+\frac{2\lambda_i}{1+c_{T_i}{}^2}+j\mu_i\right)}, \quad k_i=0,\ldots, C_i-1$$

$$H_{C_i}(s) = \left[1+\frac{\lambda_i}{s+C_i\mu_i}\right]\left(1+\left[\frac{c_{T_i}{}^{2}-1}{c_{T_i}{}^{2}+1}\right]^2\right)$$

$$\times \sum_{n_i=0}^{C_i} \frac{p_i^{+}(n_i)\,\lambda_i^{c_i-n_i}}{\prod_{j=n_i}^{C_i}(s+\frac{2\lambda_i}{1+c_{T_i}^{2}}+j\mu_i)}$$

and the probability of a departure leaving n_i packets at node i is

$$p_i^{+}(n_i)=\begin{cases}\mu_i\dfrac{(n_i+1)}{\lambda_i}\,p_i(n_i+1)\,, & \text{if } n_i=0,\cdots,C_i-2\\ \dfrac{C_i\,\mu_i}{\lambda_i}\,p_i(n_i+1)\,, & \text{if } n_i=C_i-1,\,C_i\end{cases} \tag{4.2.17}$$

From (4.2.16), we have

$$E(D_i)=\frac{1}{\lambda_i} \tag{4.2.18}$$

and

$$E(D_i^{2})=\frac{2A_i\,y_i}{\mu_i^{2}\,C_i^{2}\,(1-y_i)}-2\sum_{k_i=0}^{C_i}\frac{dH_{k_i}(s)}{d\,s}\Big|_{s=0}$$

Then the coefficient of variation of the interdeparture time c_{D_i} is given by

$$c_{D_i}=\left[\frac{E(D_i^{2})}{E^{2}(D_i)}-1\right]^{1/2} \tag{4.2.19}$$

$$= \left[\frac{2\rho_i^2 A_i y_i}{1-y_i} - 2\lambda_i^2 \sum_{k_i=0}^{C_i} \frac{dH_{k_i}(s)}{d\,s} \Big|_{s=0} - 1 \right]^{1/2}$$

Finally, from Section 4.2.1 we see that the decomposition and the superposition processes are the same as those of (4.2.7), (4.2.8) and (4.2.10).

4.2.3 Open Networks with Finite - Capacity Nodes

In this section, we propose a queueing network model and a method for its approximate solution. This model is open, implying no population limit except for buffer size limits. This model is an extention of the model studied in Section 4.2.1 except that assumption 2 is replaced by

2^*. Each node is modeled as a GE/M/1/m queue. The service times at node i are exponentially and identically distributed with mean $1/\mu_i$ and the capacity of node i is m_i packets.

To find the state probability distribution of the network, the problem is formulated as follows.

Maximize $$H = -\sum_{\mathbf{n}} p(\mathbf{n}) \ln p(\mathbf{n})$$

subject to the constraints:

Means $$\sum_{\mathbf{n}} n_i \, p(\mathbf{n}) = E(N_i)$$

State dependent utilization $\sum_{\substack{all\ n_j \\ 1\le j\le L \\ j\ne i}} \sum_{n_i=1}^{m_i} p(\mathbf{n}) = p_i(n_i = 0)$

and

Normalization $\sum_{\mathbf{n}} p(\mathbf{n}) = 1$

Using (2.4.16b), we obtain $E(N_i)$ and $p_i(0)$,

$$E(N_i) = \frac{2\rho_i(1-\rho_i)(1-\alpha_i^{m_i} - m_i\alpha_i^{m_i} + m_i\alpha_i^{m_i+1})}{(1-\rho_i\alpha_i^{m_i})(1-\alpha_i)^2(1+c_{T_i}^2)^2} \tag{4.2.20}$$

$$p_i(0) = \frac{1-\rho_i}{1-\rho_i\,\alpha_i^{m_i}}$$

with

$$\rho_i = \frac{\lambda_i}{\mu_i}$$

$$\alpha_i = 1 - \frac{2(1-\rho_i)}{1+c_{T_i}^2}$$

It follows from (2.2.8)-(2.2.9) that the solution is given by

$$p(\mathbf{n}) = \prod_{i=1}^{L} p_i(n_i)$$

where

$$p_i(n_i) = \begin{cases} \dfrac{1-\rho_i}{1-\rho_i \alpha_i^{m_i}}, & \text{if } n_i = 0 \\ \dfrac{2(1-\rho_i)\rho_i \alpha_i^{n_i-1}}{(1-\rho_i \alpha_i^{m_i})(1+c_{T_i}^2)}, & \text{if } 1 \le n_i \le m_i \end{cases} \tag{4.2.21}$$

and the loss probability at node i is $p_i(m_i)$

To determine the flow process in the network, we shall first find the interdeparture time distribution of the GE/M/1/m queue. Adapted from the results obtained from Bocharov [67], the mean and the coefficient of variation of the interdeparture time are given by

$$E(D_i) = \frac{1 + c_{T_i}^2}{\lambda_i [1 + c_{T_i}^2 - 2p_i(m_i)]} \tag{4.2.22}$$

and

$$c_{D\,i} = \sqrt{1 - p_i(0) + p_i(0)\, c_{T_i}^2 - 2p_i(0)\, p_i(m_i)} \tag{4.2.23}$$

For decomposition and superposition processes, we shall use (4.2.7), (4.2.8), (4.2.11) and (4.2.12).

4.3 QUEUEING NETWORKS WITH WINDOW CONTROL [94,95]

A major advantage of the packet switching technology is its ability to share communication resources dynamically among the users. However, under a resource sharing environment, unless traffic demands are carefully controlled serious throughtput degradation may result. In extreme cases, protocol deadlocks may occur. In that case, a group of communications resources is occupied and none of which can be released before other resources in the same group are released, thus effectively locking up the entire group. Flow control strategies are employed in a communication network in order to achieve the following objectives :

- Prevention of throughput degradation and loss of efficiency because of overload;
- Deadlock avoidance;
- Fair allocation of resources among competing users; and
- Speed matching between the network and its attached users.

4.3.1 Networks of Single-Server Nodes

In this section we shall investigate the performance of a virtual route in a packet switching network with end-to-end flow control. A packet (or a customer) that originates at the source S traverses a series of L nodes, and then arrives at the destination D. The flow-controlled mechanism operates by limiting the number of packets in transit within this virtual route. The maximum number is called the window size.

Consider a single route consisting of L nodes in series from source to destination as shown in Fig. 4.1 satisfying the following assumptions:

(a) The message generation process from the source is described by an exponential interarrival time distribution with arrival rate λ_g messages/second and each message consists of a geometrically distributed number of packets with mean 1/p packets/message.

(b) Each node in the route is modeled as a GE/G/1 queue. The service times at node i are mutually independent, identically distributed random variables with mean $1/\mu_i$ and coefficient of variation c_{Y_i}, i=1,..., L.

(c) Whenever the number of packets in transit within the route is equal to the window size n_W, the source is halted. A quiescent source is later reactivated after an end-to-end acknowledgement returns from the destination indicating receipt of a packet. The acknowledgement delay is assumed to be negligible for simplicity.

Basically, this is an extension of the model described in [68-69] except for the bulk arrival assumption. Since there is no exact solution for such a model, we shall resort to approximate analysis. The main performance measures concerned include the mean waiting time, the throughput, and the mean number of packets stored in each node in the network.

In order to tackle the problem, we shall model the message arrival process by a packet arrival process with a GE interarrival time distribution specified by mean

p/λ_g sec/packet and coefficient of variation $\sqrt{2/p-1}$. For the window flow control mechanism , we let the packet input rate depend on the total number s_L of packets in transit over this virtual route as follows:

$$\lambda_p = \begin{cases} \dfrac{\lambda_g}{p} & , \text{if } 0 \le s_L < n_W \\ 0 & , \text{if } s_L = n_W \end{cases} \tag{4.3.1}$$

Then at any given instant $s_L = n_1 + n_2 + \ldots + n_L$ is at most n_W. Such an L-stage queueing system can be represented by a closed queueing network as shown in Fig. 4.2, where a fictitious server, the (L+1)*st* server, with service rate λ_g/p and coefficient of variation of the service time $\sqrt{2/p-1}$ is inserted in the reversed channel (or acknowledgement channel).

It has been shown in (4.2.5) that the steady-state joint state probability distribution of an open network with group arrivals, obtained by entropy maximization subject to the constraints on the means $E(N_i)$ and the utilizations $1-p_i(0)$, i=1,..., L+1, has a product form distribution

$$p(\mathbf{n}) = \prod_{i=1}^{L+1} p_i(n_i)$$

where $p_i(n_i)$ is given in (4.2.5).

In general, however, these constraints are unavailable for closed queueing networks. Rather, to obtain the state probability distribution $p^*(\mathbf{n})$ of our network Q^*, we interprete $p^*(\mathbf{n})$ as the state probability distribution of an open network Q

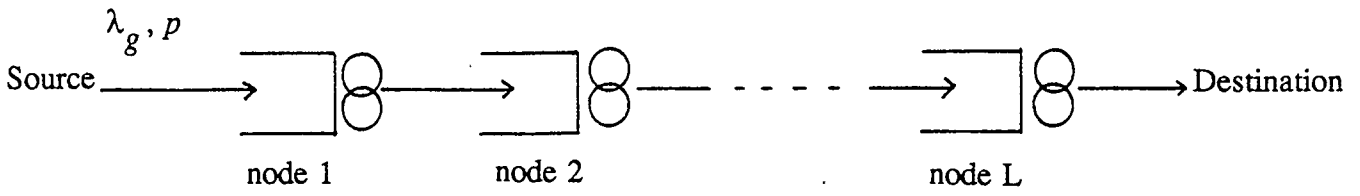


Fig. 4.1 A virtual route consisting of L nodes

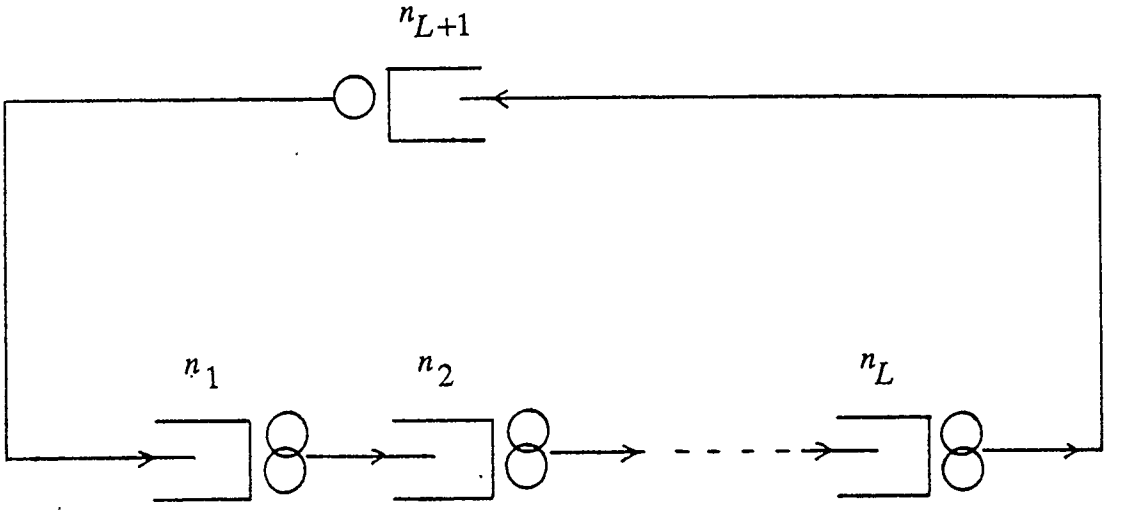


Fig. 4.2 A closed queueing network with population n_W

conditioned to a population of finite size n_W [70], that is,

$$p^*(\mathbf{n}) = \frac{p(\mathbf{n})}{\sum_{s_L=n_W} p(\mathbf{n})} \tag{4.3.2}$$

Furthermore, let the arrival rate λ_0 of Q be such that the total average population of Q is equal to n_W, i.e.,

$$\sum_{i=1}^{L+1} E(N_i) = n_W \tag{4.3.3}$$

where $E(N_i)$ are given in (4.2.4). These formulas express the mean number of packets stored in each node as a function of node utilization factors and coefficients of variation of the interarrival time . For each node the latter may be obtained from (4.2.6).

Since our result is an approximate solution, in order to be consistent with the fundamental Work Rate Theorem [71] which states that throughput rate of each node in a closed network should be the same, we propose the following solution.

$$p^*(\mathbf{n}) = \frac{\prod_{i=1}^{L+1} g_i(n_i)}{\sum_{\sum_{k=1}^{L+1} n_k = n_W} \prod_{k=1}^{L+1} g_k(n_k)} \tag{4.3.4}$$

where

$$g_i(n_i) = \begin{cases} 1, \text{ if } n_i = 0 \\ b_i \left[2\dfrac{\rho_i}{\rho_i + \rho_i c_{Y_i}^2 + c_{T_i}^2 - 1} \right] \left[\dfrac{\rho_i c_{Y_i}^2 + c_{T_i}^2 + \rho_i - 1}{\rho_i c_{Y_i}^2 + c_{T_i}^2 - \rho_i + 1} \right]^{n_i}, \text{ if } n_i \geq 1 \end{cases}$$

The parameter b_i is determined by the iteration equation

$$b_i^{(k+1)} = \frac{b_i^{(k)} n_W}{U_i \sum_{j=1}^{L+1} E(N_j) \frac{\mu_i}{\mu_j U_j}} \tag{4.3.5}$$

with initial values

$$b_i^{(1)} = 1$$

The iteration procedure terminates when two successive values of $b_i^{(k)}$ are sufficiently close, e.g.,

$$\left| b_i^{(k+1)} - b_i^{(k)} \right| \leq 10^{-4}$$

where the means $E(N_i)$ and the utilizations U_i, i=1,..., L+1, are obtained as follows

For convenience we let $G(k,l)$ be defined as

$$G(k,l) = \sum_{\substack{\sum_{j=1}^{l} n_j = k}} \prod_{i=1}^{l} g_i(n_j) \tag{4.3.6}$$

By virtue of (4.2.3), marginal state probabilities of nodes in the network can be obtained as

$$p_i(n_i) = \sum_{\substack{\sum_{j=1, j\neq i}^{L+1} n_j = 1}}^{n_W} \frac{\prod_{k=1}^{L+1} g_k(n_j)}{\sum_{\sum_{k=1}^{L+1} n_k = 1}^{n_W} \prod_{k=1}^{L+1} g_k(n_k)} \tag{4.3.7}$$

$$= g_i(n_i) \frac{G(L+1-i, n_W)}{G(L+1, n_W)}$$

The mean number of packets stored at node i is

$$E(N_i) = \sum_{k=0}^{n_W} k \; g_i(k) \frac{G(L+1-k, n_W)}{G(L+1, n_W)} \tag{4.3.8}$$

The server utilization of each node is defined as

$$U_i = \frac{\sum_{n_i=1}^{n_W} \min(C_i, n_i) \, p_i(n_i)}{C_i}$$

If C_i=1, then we have

$$U_i = 1 - p_i(0) \tag{4.3.9}$$

Finally, the throughput of node i is defined as

$$\theta_i = \mu_i \; U_i C_i \tag{4.3.10}$$

4.3.2 Networks of Nodes with Multiple Exponential Servers

Consider the network with assumptions (a), (c) and an additional assumption

(d) Each node in the route is modeled as a GE/M/C queue with C_i servers. The service time is exponentially distributed with mean $1/\mu_i$ sec/packet, i=1,..., L.

Recall from Section 4.2.2 that the steady-state probability distribution of an open network with group arrivals subject to constraints on the means and load dependent utilizations can be obtained from (4.2.14).

Similar to (4.3.4), we obtain the solution for the closed network

$$p^*(\mathbf{n}) = \frac{\prod_{i=1}^{L+1} g_i(n_i)}{\sum_{\sum_{k=1}^{L+1} n_k = n_W} \prod_{k=1}^{L+1} g_k(n_k)} \tag{4.3.11}$$

where

$$g_i(n_i) = \begin{cases} 1 \quad \text{if } n_i = 0 \\ \dfrac{b_i}{n_i\,!} \prod_{r=0}^{n_i-1} \dfrac{2\rho_i C_i + r(c_{T_i}{}^2 - 1)}{1 + c_{T_i}{}^2}, & \text{if } 1 \le n_i \le C_i \\ g_i(C_i) \left[\dfrac{c_{T_i}{}^2 - 1 + 2\rho_i}{1 + c_{T_i}{}^2} \right]^{n_i - C_i}, & \text{if } n_i \ge C_i \end{cases}$$

4.4 GENERAL QUEUEING NETWORKS WITH MULTISERVER NODES

One of the primary applications of modern queueing theory is the modeling of computer systems and data networks. However, the analysis of complex and general G/G/C queueing networks often runs into difficulties due to an extremely large number of system states. So far there is no exact method for the solution. Thus one must resort to develop approximate methods to solve complex queueing problems.

In this section we shall extend the notion of single queues to a complex network with G/G/C stations. The approach employed can be divided into two steps; first, we decompose the network into subsystems, each of which is modeled as a G/G/C queueing station. Then we compute the first two moments of the arrival process for each station. Secondly we determine the solution of each independent station through the methods of diffusion approximation and entropy maximization.

Then the performance measure for each station can be evaluated. Now we consider an open network with L G/G/C nodes under the following conditions :

1. Each node has C_i identical servers, i=1,..., L. The service time at node i has a general distribution with mean $1/\mu_i$ and coefficient of variation c_{Y_i}.

2. After being served the customers at station i are routed toward node j with probability r_{ij} which is independent of the system states.

3. The interarrival time has mean $1/\lambda_0$ and coefficient of variation c_{T_0}. An external customer joins the system at the j^{th} node with probability r_{0j} and leaves the system from the i^{th} node with probability $r_{i,L+1}$. We thus have

$$\sum_{j=1}^{L} r_{0j} = 1$$

Let $1/\lambda_i$ and $c_{T\,i}$ be respectively the mean and coefficient of variation for the interarrival time at node i. In an open network, the traffic entering a node i is the sum of the traffics from other nodes j and the external source with probability r_{ji}, i=1,..., L+1; j= 0,1,..., L. Thus we obtain

$$\lambda_i = \sum_{j=0}^{L} \lambda_j r_{ji} \quad , \text{for } 1 \le i \le L+1 \tag{4.4.1}$$

Let $\rho_i = \lambda_i/(\mu_i\, C_i)$ be the utilization of node i. Now the problem is to determine the value of the coefficient of variation $c_{D\,i}$ of the interdeparture time

D_i, i= 1,..., L. Suppose that the interarrival time distribution of each input traffic is known with mean λ_i and coefficient of variation c_{T_i}. Then we can define an equivalent single-server queue. This server is characterized by the first two moments, $\hat{\mu}_i$ and $\hat{c}_{Y_i}$, of its service time. We replace the C_i servers by an equivalent server whose service time distribution is chosen to satisfy $G_Y(t) = F_Y(C_i\ t)$, where $F_Y(t)$ is the original service time distribution. The first two moments of this equivalent service time are equal to

$$\hat{\mu}_i = C_i\ \mu_i$$

and

$$\hat{c}_{Y_i} = c_{Y_i}$$

Then the solution for c_{D_i} is obtained by replacing μ_i and c_{Y_i} by $\hat{\mu}_i$ and $\hat{c}_{Y_i}$ in (4.2.6), respectively.

Next, the means and coefficients of variation of the interarrival time for the decomposition and superposition processes are determined by (4.2.7) and (4.2.8), and (4.2.10), respectively.

Finally, after $1/\lambda_i$ and c_{Ti} have been determined, we can determine the state probabilities and mean number of customers in the G/G/C queues from (3.6.19)-(3.6.21) and (3.6.24) using the method of diffusion approximation and from (3.2.5) and (3.2.6) using entropy maximization. The procedure is as follows.

- Solving the set of L linear equations (4.4.1) for λ_i, i=1,..., L.
- Solving (4.2.6)-(4.2.8) and (4.2.11) and (4.2.12) for c_{T_i}.
- Using (3.6.24) or (3.2.6) to get the mean number of customers in the system.

4.5 SUMMARY

In Section 2, open networks with bulk Poisson inputs have been studied. The steady-state probability distribution of the network has been obtained by means of the entropy maximization method. Then using a decomposition technique, the traffic flow processes in the network are obtained. The solution has a product form and may be considered as an extension of the Jackson network.

In Section 3, a model of bursty traffic with window-flow control on a virtual route is developed. A closed queueing network is formulated under the scheme of flow control. The solution is consistent with the Work Rate Theorem for general closed queueing networks. Expressions for the throughput and the mean number of packets stored at each node are also obtained.

In Section 4, an open network consisting of G/G/C queueing stations is investigated. After analyzing the approximate output processes for the G/G/C queues and the flow processes of the network , the mean number of customers at each station is obtained via the methods of entropy maximization and diffusion approximation.

CHAPTER 5

SOLUTIONS FOR NETWORKS OF QUEUES

In this chapter we shall investigate networks of queues with bulk arrivals and FIFO scheduling using entropy maximization as proposed in Chapter 4. In the first section, open networks of queue in continuous time will be considered and examples shall be given to examine the accuracy of the expression for the mean number of packets. In section 2 we shall study the behavior of a buffered interconnection network. In section 3 we shall consider a window-based end-to-end flow control mechanism in a packet-switching network. We shall make a comparison study of the iterative solution method presented in Section 4.3 and the simulation method.

In section 4 we shall study the behavior of open networks with multiserver nodes, nonexponential service time distribution and FIFO scheduling using the method proposed in section 4.4. The accuracy of the expression for the approximate mean number of packets using the methods of diffusion approximation and entropy maximization will be investigated.

5.1 EXAMPLES OF OPEN NETWORKS OF QUEUES

Consider the open network of queues in which the interarrival times of messages are exponentially distributed with mean $1/\lambda_g$ and the message length is geometrically distributed with mean $1/p$. The sevice time distribution of node i is a

general distribution with mean $1/\mu_i$ and coefficient of variation c_{Y_i}. In the case of multiserver nodes, we consider only the case of exponential servers. The network state probability distribution is specified by (4.2.5) for single-server nodes and by (4.2.14) for multiple-server nodes. As discussed in Section 4.2.1 , the central issue in obtaining the mean number of packets is the characterization of the second moments of the internal interarrival time distributions. Assuming the interarrival time to be renewal, since bulk Poisson message arrivals can be characterized by a GE packet interarrival time distribution in the maximum-entropy sense, we may use equations (4.2.6), (4.2.8), (4.2.12), together with (4.2.19) to approximate the coefficient of variation c_{T_i} of the interarrival time distributions. Subsequently, the mean $E(N_i)$ may be obtained using (4.2.4) or (4.2.13a). Their accuracy will be compared with the results obtained by simulation. In the simulation, we use the independent-replication method [72]. The 95 percent confidence interval was computed for each simulation.

Example 1. Consider the network shown in Fig. 5.1 with a bulk Poisson input. Node 1 has two, three, or five exponential servers and node 2 has three servers with exponential or constant service times. Let the traffic intensity of node 1 , ρ_1, be 0.625. The performance measure is the mean number of packets stored at node 2 against its traffic intensity ρ_2. In the simulation, we chose ρ_2 to be 0.35, 0.6, and 0.85 which corresponds to light, moderate and heavy traffic conditions. The

analytic results presented in Fig. 5.2, 5.3 and Table 5.1 show fairly close agreement with the simulation results.

Example 2. Consider the network shown in Fig. 5.4 with bulk Poisson input and interfering Poisson traffic . The arrival rate of external traffic λ_2 is chosen to be 0, 0.85, and 2.15. Both nodes consist of a single exponential server. Analytic results are summarized in Fig. 5.5 and Table 5.2. It is seen that the accuracy is improved as λ_2 becomes large. Note that the accuracy of our approximate analysis of open networks will be improved as the number of nodes or traffic sources becomes large. The reason is that the renewal assumption of flow processes is more realistic when the number of component flow processes is large. In fact, if the component processes are mutually independent, in the limit as the number of component flow process approaches infinity, the superposition process becomes Poisson.

Example 3. Now consider a two-node network in Fig. 5.4 in which buffer capacities of nodes 1 and 2 are limited to m_1 and m_2 packets, respectively. The service mechanism is a partially rejected mechanism which means that part of the input message is lost if the buffer is full. The coefficient of variation of the interdeparture time can be obtained using the formula (4.2.23) and the mean number of packets is available from (4.2.20). Since the analytic result for the state probability distribution of node 1 is correct as shown in Section 2.4.3, we only consider the mean $E(N_2)$ against ρ_2. Numerical results are summarized in Table

5.3 for $m_2=\infty$ and displayed in Fig. 5.6 for $m_1=5$ and $m_2=10$. We see that the accuracy is better if queue capacity m_1 is large.

Example 4. Now consider the network shown in Fig. 5.7. External bulk Poisson arrivals enter node 1, 2 and 3. Let the number of exponential servers of node 1- 5 be 2, 2, 3, 1, and 3, respectively. The traffic intensities are assumed to be $\rho_1 = \rho_2 = 0.625$, 0.833, and $\rho_2 = 0.312$. We shall calculate the means $E(N_4)$ and $E(N_5)$ against ρ_4 and ρ_5 respectively. In the simulation, we use the independent-replication method [72]. From Fig. 5.8 and Table 5.4 we see that the analytic results show good agreement with the simulation results.

5.2 APPROXIMATE ANALYSIS FOR ASYNCHRONOUS INTERCONNECTION NETWORKS WITH BULK ARRIVAL [96]

Buffered interconnection networks have received increasing consideration for use in parallel computers. They are integral components of several machines currently under development, including the Cedar machine at the University of Illinois [73], the NYU Ultracomputer at New York University [74], and the RP3 machine [75] at IBM, where they are used to interconnect processors to share the memory. In order to study the multitude of options available in actually building a machine, it is extremely useful to have formulas that approximately measure the

performance of an interconnection network. Early work on interconnection networks was done by Goke and Lipovski [76], Lawrie [77], and Patel [78], among others. Also there have been a number of performance analyses of interconnection networks. Krushal et al [79] provided analysis of buffered banyan networks. Their analysis is mainly concerned with the distribution of waiting time at the first stage of the network. Of interest, nevertheless, are their simulation results which indicate the delays in successive stages are nearly independent. Jenq [80] made an analysis of the performance of a packet switch based on a single-buffered banyan network. He provided results on the throughput, delay and internal blocking. Bouras et al [81] studied discrete-time networks and provided upper bounds on the mean delays of the second stage and beyond for the case of infinite buffers. All the above network models dealt with external input of single-packet message arrivals. However, the performance analysis of networks with bulk arrivals has not yet been considered .

In this section we study the performance of a buffered, packet-switching, multistage interconnection network. The external input of message arrivals consists of a random number of packets. We propose an approximate method for analyzing the mean delay times and the number of packets of each node in the network.

5.2.1 Network Modeling and Assumptions

Consider a packet switching network built of switches connected by unidirectional lines. An r×q switch with r-input and q-output can receive packets at

each input, and send them through any one of its outputs. The nodes of a network are of the following three types: (a) source nodes, (b) sink nodes, and (c) switches. A banyan network is defined by Goke and Lipovsky [76] to be a network with a unique path from each source node to each sink node. A multistage network is a network in which the nodes can be arranged in stages, with all the source nodes at stage 0, and all the outputs at stage i connected to inputs at stage i+1. A uniform network is a multistage network in which all switches at the same stage have the same number of input ports and the same number of output ports. A square network is said to be of degree k if it is built of k × k switches. Fig. 5.9 shows a 4-stage square banyan network of degree 2, each input port of the switch has a buffer. Arriving packets that can not find a free output link are immediately stored in the buffer. The buffer capacity is assumed to be infinite.

To obtain tractable results for the performance of the n_b-stage square banyan network we make the following assumptions:

1. The interarrival times T_g of external messages are mutually independent , identically and exponentially distributed with mean $1/\lambda_g$ seconds.
2. The message size B is geometrically distributed with mean 1/p packets/message.
3. The service time (or packet transmission time) distribution at a switch of stage i is general with mean $1/\mu_i$ and coefficient of variation c_{Y_i}. The queue discipline is FIFO.

4. Packets arriving at each input port at stage 1 are destined uniformly (or randomly) for all output ports at stage n_b. The traffic loads are balanced in the whole network. Under these assumptions, the state of a switch at each stage is statistically the same or identical.

Note that operation of a switch in the first stage can be modeled as an $M^X/G/1$ queue. The generating function of the state probability distribution is given by (2.4.13).

Since accurate analysis of the performance does not seem tractable for later stages , we present here an approximate method. This method undertakes the approximation of the departure process from each queue as a stationary renewal process and the description of each stationary renewal process by the mean and the coefficient of variation of the underlying interarrival time distribution. First, we decompose our network into GE/G/1 queues. Specifically : (a) the departure process from a queue in stage i is approximated by a stationary renewal GE process with mean arrival rate λ_g/p and coefficient of variation c_{D_i} ; (b) this departure process is split into two stationary renewal processes (as shown in Fig. 5.9) each with mean arrival rate $\lambda_g/2$ and coefficient of variation $c_{i,i+1}$; and (c) if two GE processes, each with mean arrival rate $\lambda_g/2$ and coefficient of variation $c_{i-1,i}$, are superimposed, then the resultant process is described by (λ_g/p, c_{T_i}).

We now give three expressions corresponding to the above three cases. From (4.2.6), we have the formula

$$c_{D_i} = \left[\rho_i - \rho_i^2 + \rho_i^2 \, c_{Y_i}{}^2 + (1-\rho_i) \, c_{T_i}{}^2 \right]^{1/2} , \; i = 1, \ldots, \; n_b \tag{5.1}$$

where $\rho_i = \lambda_g / \mu_i \; p$.

Decomposing, or splitting, of renewal processes from (4.2.8) yields,

$$c_{i-1,i} = \sqrt{(c_{D_{i-1}}{}^2 + 1)/2} \; , \; i = 2, \ldots, \; n_b \tag{5.2}$$

For superposition we use (4.2.12) and obtain

$$c_{T_i} = c_{i-1,i} \; , \; i = 2, \ldots, \; n_b \tag{5.3}$$

Combining (5.1)-(5.3), we obtain

$$c_{T_i} = \sqrt{(1 + \rho_i - \rho_i{}^2 + \rho_i{}^2 \, c_{Y_i}{}^2 + (1-\rho_i) c_{T_{i-1}}{}^2)/2} \; , \; i=2, \ldots, \; n_b \tag{5.4}$$

with initial condition

$$c_{T_1} = \sqrt{2/p - 1}$$

Equations (5.4) provide a complete recursive formula for calculating $\{c_{T_i} \; ; \; i \geq 1\}$. Note that the number n_b of stages of the network does not appear

explicitly in the formula.

Also from (2.4.15a), the mean waiting time $E(W_i)$, the mean $E(N_i)$ and the variance $Var(N_i)$ for stage i can be calculated and are given by the following expressions for the GE/G/1 queue

$$E(W_i) = \frac{c_{T_i}^2 - 1 + \rho_i + \rho_i\, c_{Y_i}^2}{2(1-\rho_i)}$$

$$E(N_i) = \frac{\rho_i(1-\rho_i + c_{T_i}^2 + \rho_i\, c_{Y_i}^2)}{2(1-\rho_i)} \tag{5.5}$$

$$Var(N_i) = \frac{\rho_i(1-\rho_i+c_{T_i}^2+\rho_i c_{Y_i}^2)\left[2c_{T_i}^2-\rho_i(1-\rho_i+c_{T_i}^2+\rho_i c_{Y_i}^2)\right]}{4(1-\rho_i)^2}$$

where

$$\rho_i = \frac{\lambda_g}{\mu_i\, p}$$

Finally, the mean transfer delay from source to destination, T_d, for a packet, is given by

$$T_d = \sum_{i=1}^{n_b} \frac{1}{\mu_i} + \sum_{i=1}^{n_b} E(W_i) \tag{5.6}$$

The first term on the right-hand side is the transmission time and the second term is the cumulative queueing delay. Note that (5.6) is the only point in the procedure where n_b is used.

5.2.2 Numerical Results and Discussions

Tables 5.5-5.7 give numerical results for the mean and variance of number of packets and the average transfer delay of packets through the network. In addition to the analytic results, a simulation program is constructed to determine the performance measures, this being a means of validating the analytic results. Each simulation run is terminated after 60000 packets have been processed. Each run takes about 2800 cpu seconds for simulating a 3-stage network on the Cyber 860. It is observed that the means and variances of number of packets are small for low to medium traffic intensity but becomes larger for higher values of traffic intensity. It is further observed that the means and variances become larger as the message size increases. Analytic results compare favourably with simulation results. Further, the important observation is that the mean number of packets shrinks at the early stages rather quickly before reaching an asymptotic value. For any fixed $c_{Y_i} = c_Y$, $\mu_i = \mu$ and $\rho=\lambda_g/\mu p$, the asymptotic value of $c_{T_i}^2$, $c_{T_\infty}^2$, as $n_b \to \infty$ obtained simply from (5.4) is given by

$$c_{T_\infty}{}^2 = \frac{1 + \rho + \rho^2(c_Y^2 - 1)}{1 + \rho} \tag{5.7}$$

It may be substituted into (5.5) to obtain the asymptotic values $E(N_\infty)$ for mean and $Var(N_\infty)$ for variance as $n_b \to \infty$.

$$E(N_\infty) = \frac{\rho(1-\rho + c_{T_\infty}{}^2 + \rho\, c_Y^2)}{2(1-\rho)} \tag{5.8}$$

$$Var(N_\infty) = \frac{\rho(1-\rho+c_{T_\infty}{}^2+\rho c_Y{}^2)[2c_{T_\infty}{}^2-\rho(1-\rho+c_{T_\infty}{}^2+\rho c_Y^2)]}{4(1-\rho)^2}$$

5.3 WINDOW CONTROL ANALYSIS IN PACKET SWITCHING NETWORKS

Consider the network of two tandem queues shown in Fig. 5.10. We shall give two examples concerned with the window flow control mechanism proposed in Section 4.3.

Example 1. Two links in series are considered. Each link is modeled as a single-server node. Its transmission time is exponentially distributed with mean $1/\mu$. Message arrivals follow a Poisson process with mean λ_g message/sec. and the message length is geometrically distributed with mean $1/p$ packets/message. The window size is assumed to be n_W. Let λ_g=0.1, p=0.8, and μ=0.2. From (4.3.8)-

(4.3.10) , we can calculate the mean number of packets, the utilization and the throughput of each node. Numerical results are summarized in Table 5.8 and plotted in Figs. 5.11 and 5.12. They agree closely with the simulation results. We also see that the throughputs of both nodes are equal thus satisfy the Work Rate Theorem. Their utilizations are also equal because we chose identical service time distribution. The utilizations and the throughput increase proportional to the window size but tend to be constants. The reason is that as the window size becomes large, the network is close to an open network. Virtually the utilizations are equal to $\rho=\lambda_g/\mu p$ and the throughput will equal to λ_g/p. In our example, these values are 0.625 and 0.125, respectively. Also the mean number of packets at each node increases as the window size grows but tends to be constant. Based on our approximate analysis from Section 4.2 for open networks , we obtain

$$E(N_1) = \frac{\rho\, p}{(1-\rho)}$$

$$E(N_2) = \frac{\rho p - \rho + 1}{p(1-\rho)}$$

In our case these values are 2.083 and 1.823, respectively.

Example 2. Consider the same network shown in Fig. 5.10. Now node 2 is replaced by a 3-server node. Let λ_g=0.1, μ_1=0.2, μ_2=0.1, and p=0.8. Results for utilizations, throughput and mean number of packets stored at each node are summarized in Table 5.9. We display only the means in Fig. 5.13. Again we see

the throughput approaches 0.125 for large window size. The utilizations for node 1 and node 2 approach 0.625 and 0.4167, respectively. The means $E(N_1)$ and $E(N_2)$ approach 2.083 and 1.462, respectively. Asymptotic value for the mean $E(N_2)$ is calculated using (4.2.13a). Note that our analytic results agree closely with the simulation results.

5.4 OPEN NETWORKS WITH NONEXPONENTIAL MULTISERVER NODES

In this section we shall investigate open networks of queues with nonexponential multiple severs using the method of diffusion approximation and entropy maximization. Consider the network shown in Fig. 5.14 which represents an iterative two-stage system. Customers arriving at the system follow a Poisson stream with arrival rate λ. After passing through node 1 they either leave the system with probability $1-r_1$ or join queue 2 with probability r_1. Customers leaving node 2 join queue 1. The queue disciplines are FIFO for both nodes. Both nodes consist of 3 servers. We shall determine the mean number of customers of both nodes. Using the method proposed as in Section 4.4 and solving (4.2.6), (4.2.8), and (4.2.12) for coefficients of variation c_{T_i}, i=1,2 , we then apply these values to (3.2.6b) and (3.6.24) to obtain $E(N_i)$, i=1,2 using entropy maximizaton and diffusion approxiamtion, respectively. We have examined many different combinations of arrival rate λ, service rate μ_1 and μ_2, coefficients of variation c_{Y_1}

and c_{Y_2}. Some typical results are presented in Tables 5.10(a)-5.10(b). We see that the results obtained by the entropy maximization appear to be better than those obtained by the diffusion approximation. For large values of the coefficients of variation c_{Y_i}, the means $E(N_1)$ and $E(N_2)$ increase but analytic results deteriorate, hence a more accurate expression for the interdeparture time of a G/G/C system is needed in order to improve the accuracy.

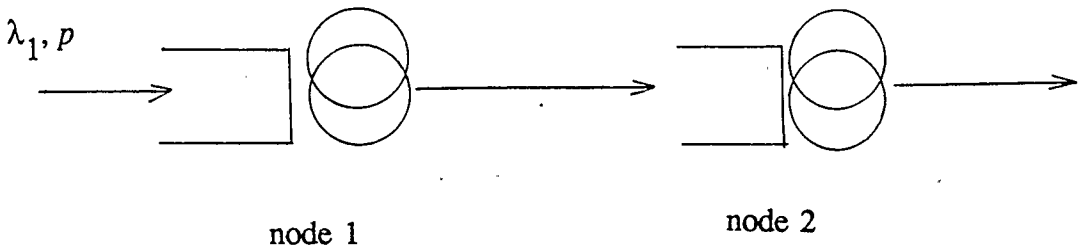


Fig. 5.1 A two-node network

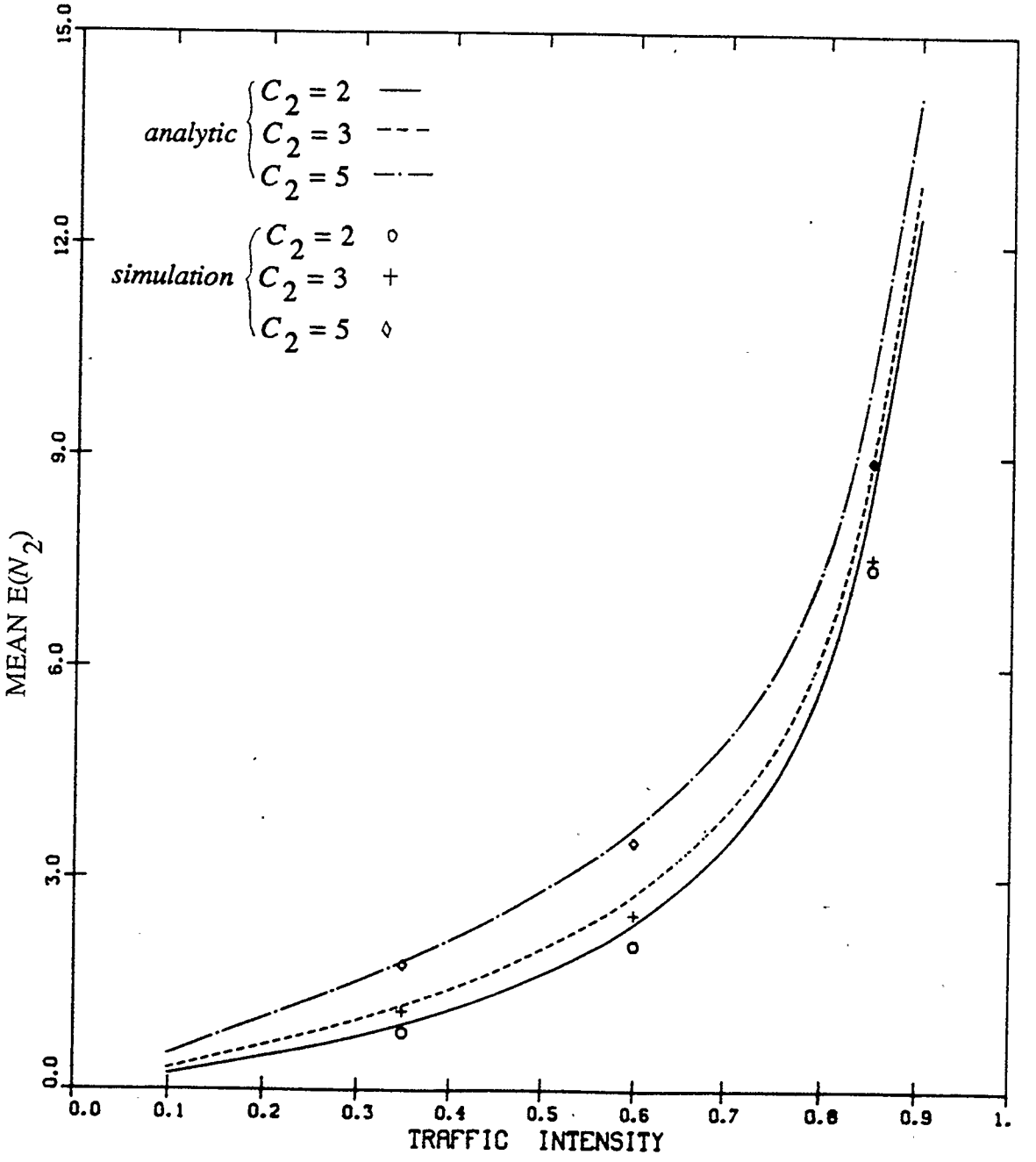


Fig. 5.2 Mean E(N_2) v.s ρ_2 (node 2: exponential service time)

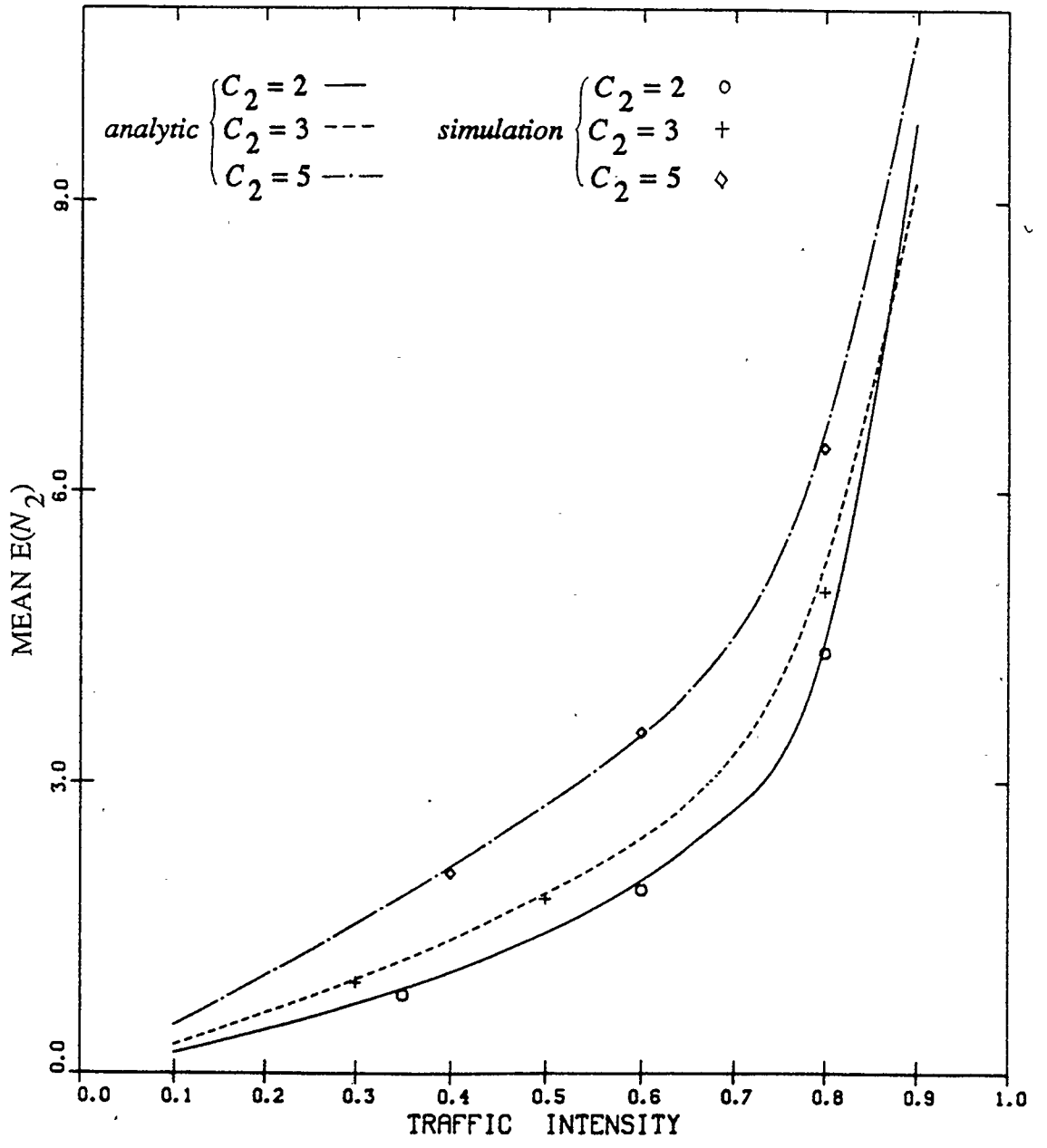


Fig. 5.3 Mean E(N_2) v.s ρ_2 (node 2: constant service time)

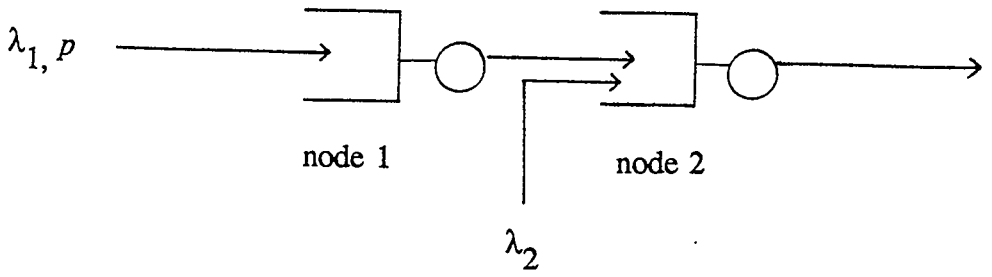


Fig. 5.4 A two-node network with interfering traffic

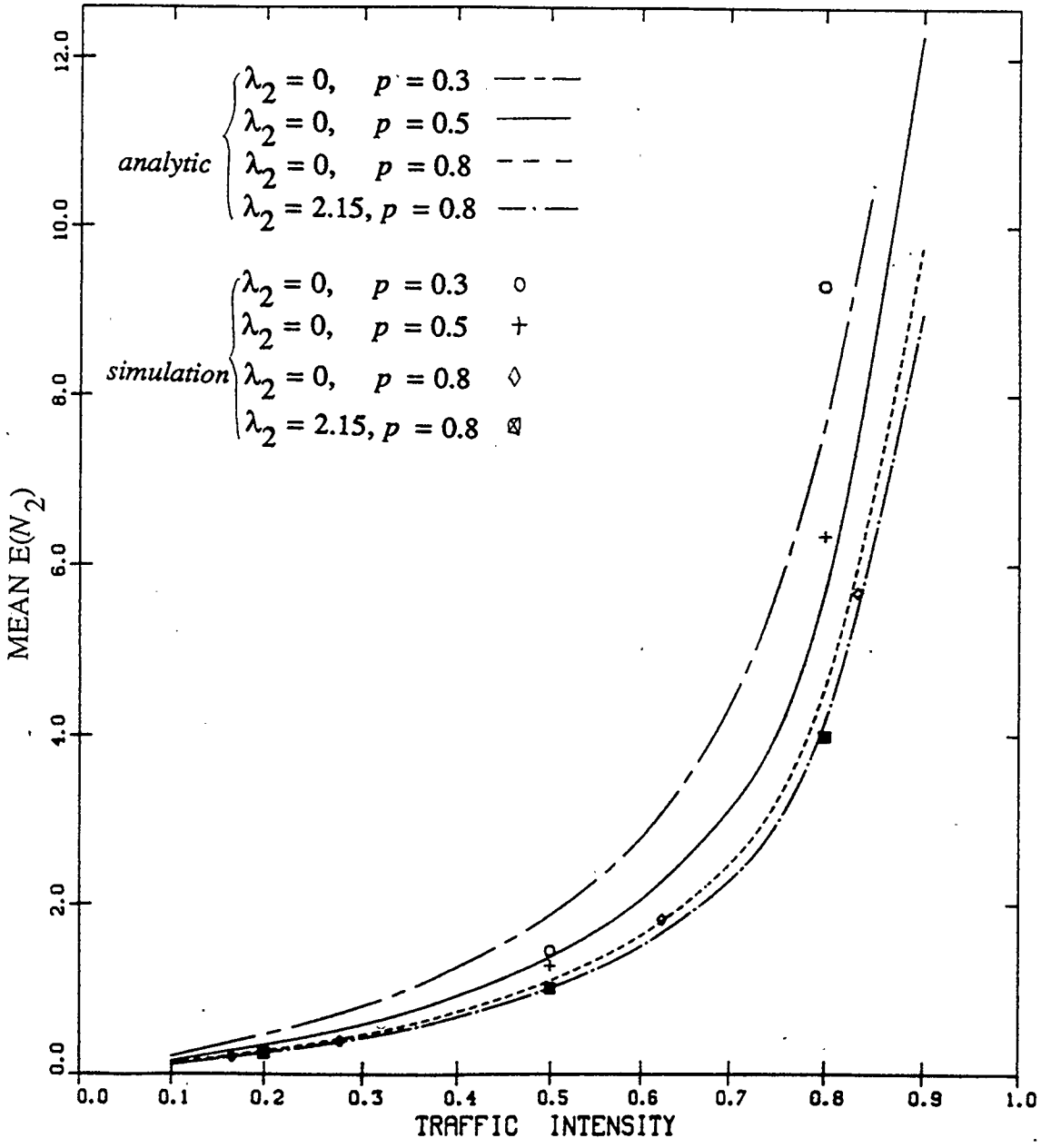


Fig. 5.5 Mean $E(N_2)$ v.s ρ_2 for the network of Fig. 5.4

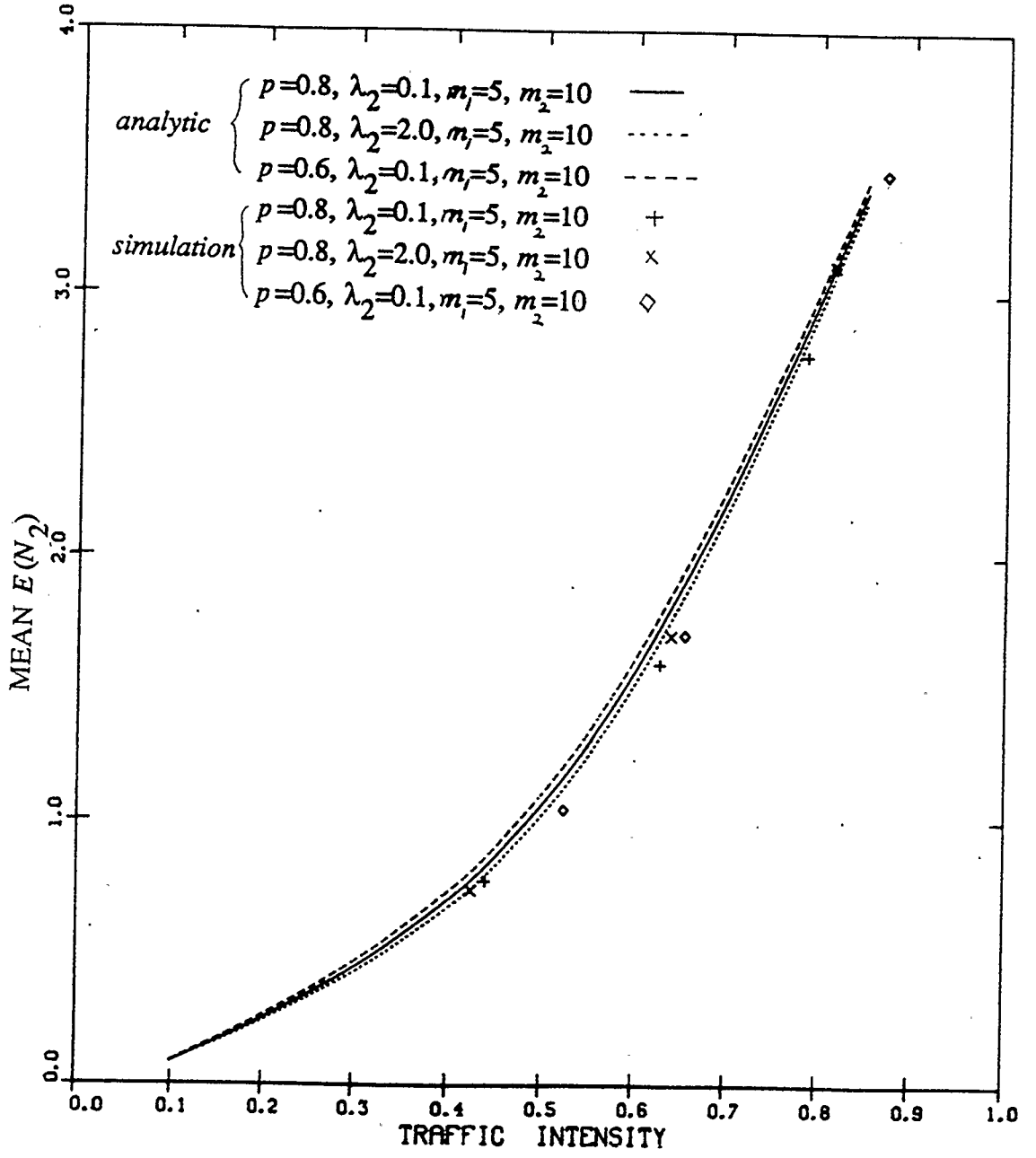


Fig. 5.6 Mean $E(N_2)$ v.s ρ_2

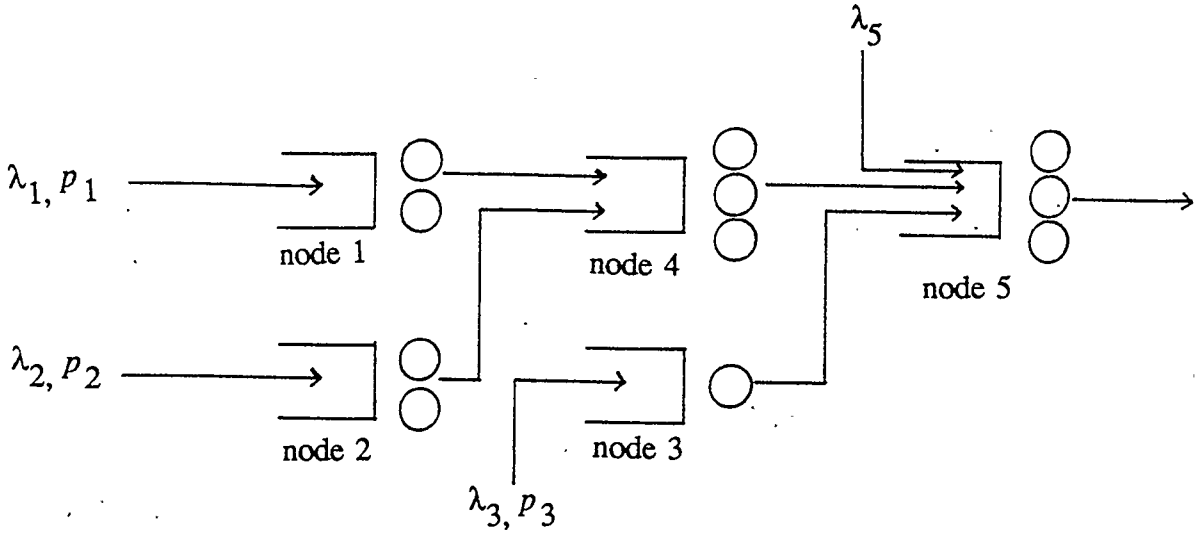


Fig. 5.7 A five-node network

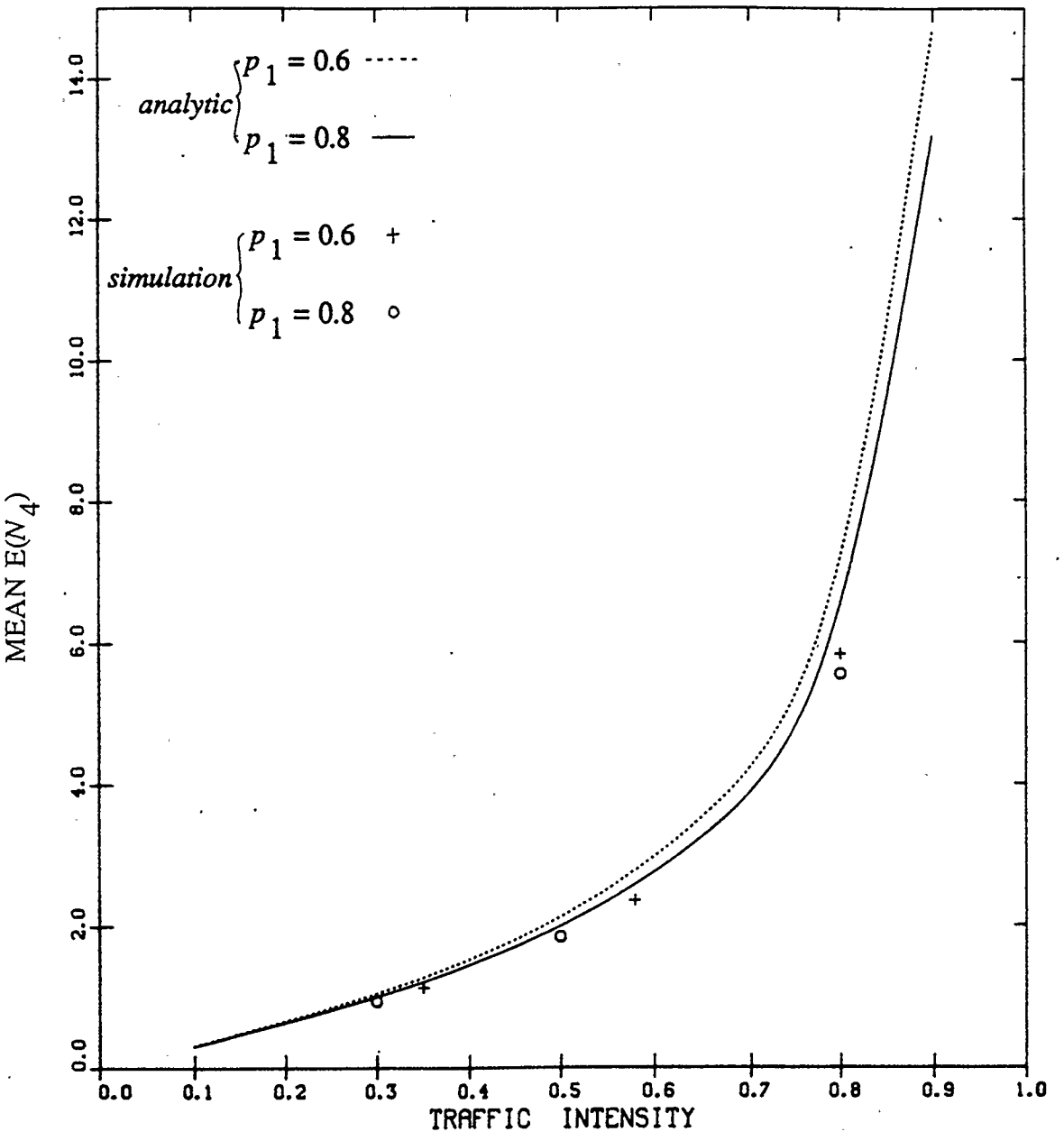


Fig. 5.8 Mean $E(N_4)$ v.s ρ_4 for network of Fig. 5.7

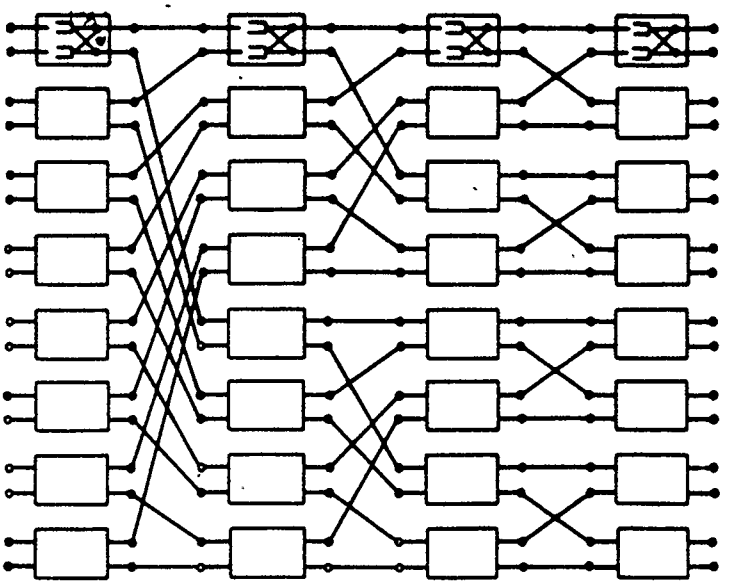


Fig. 5.9 A four-stage square banyan network with degree two [80]

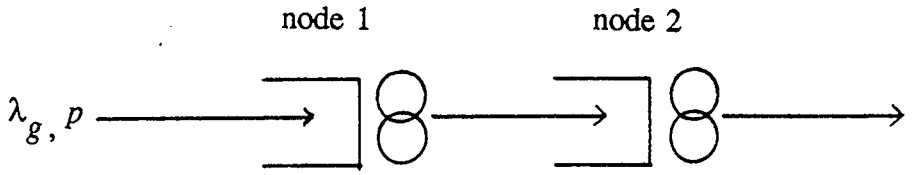


Fig. 5.10 A virtual route consisting of two nodes

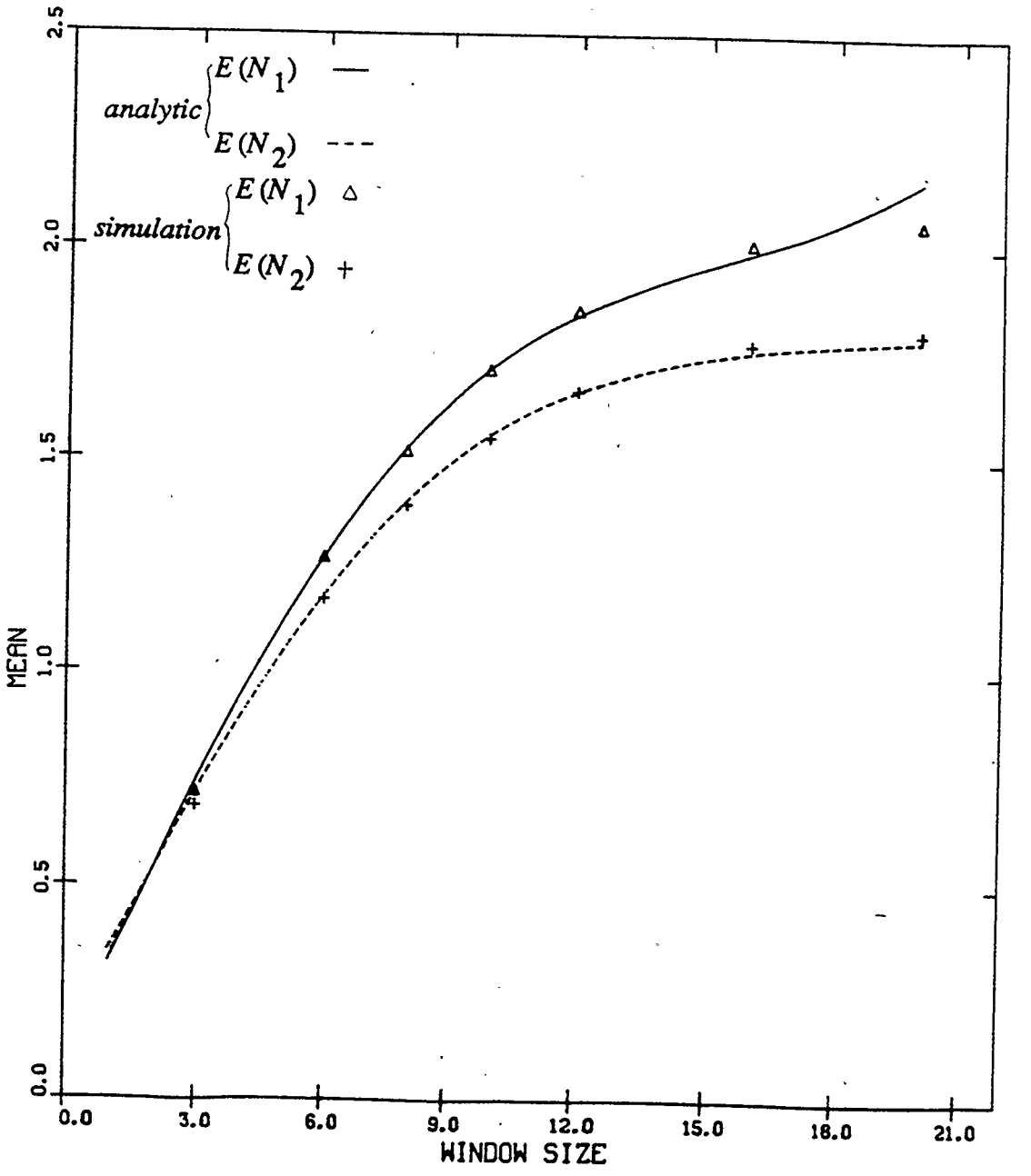


Fig. 5.11 Mean number of packets v.s window size n_W (C_1=C_2=1)

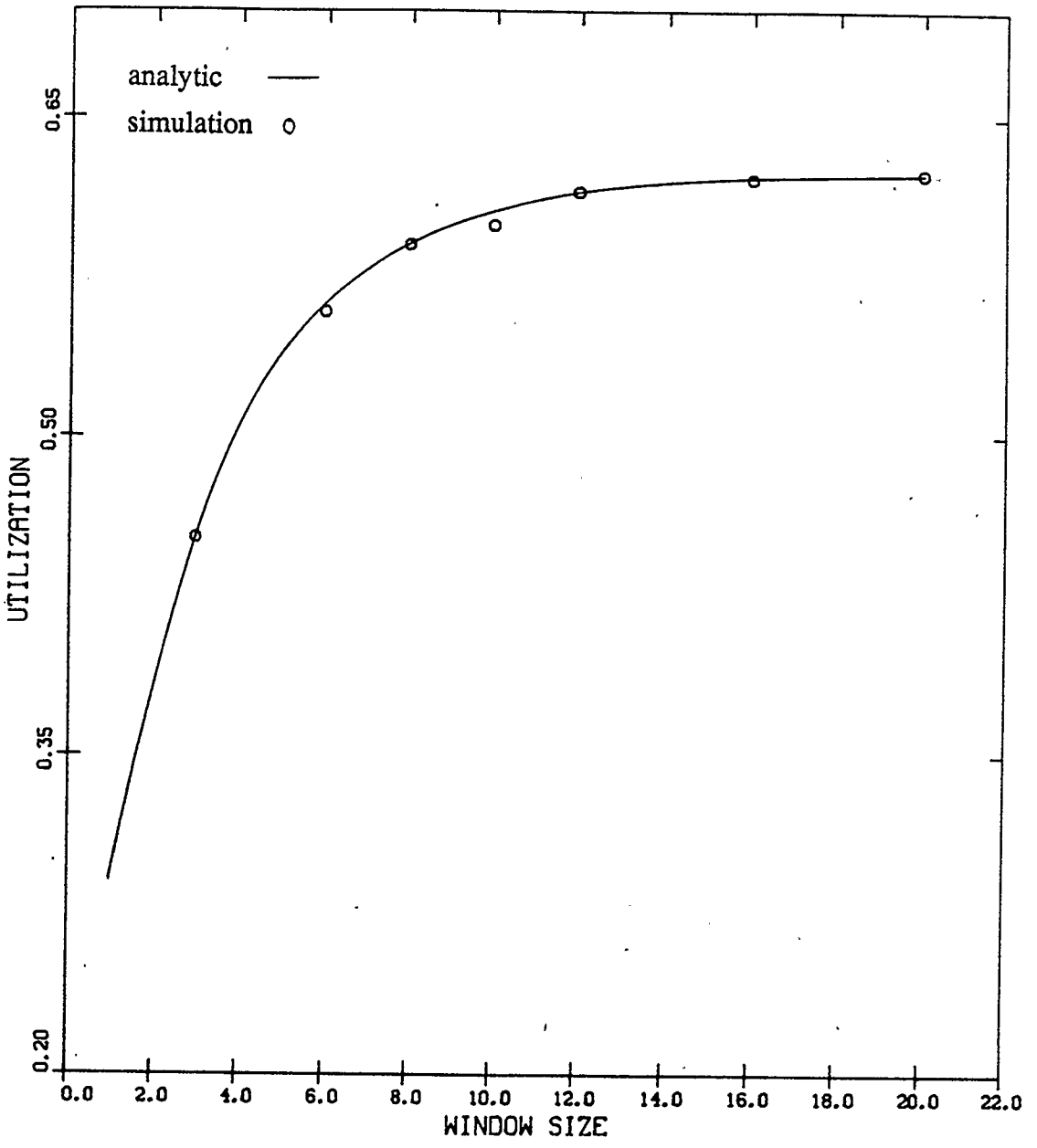


Fig. 5.12 Utilization U_1 v.s window size n_W

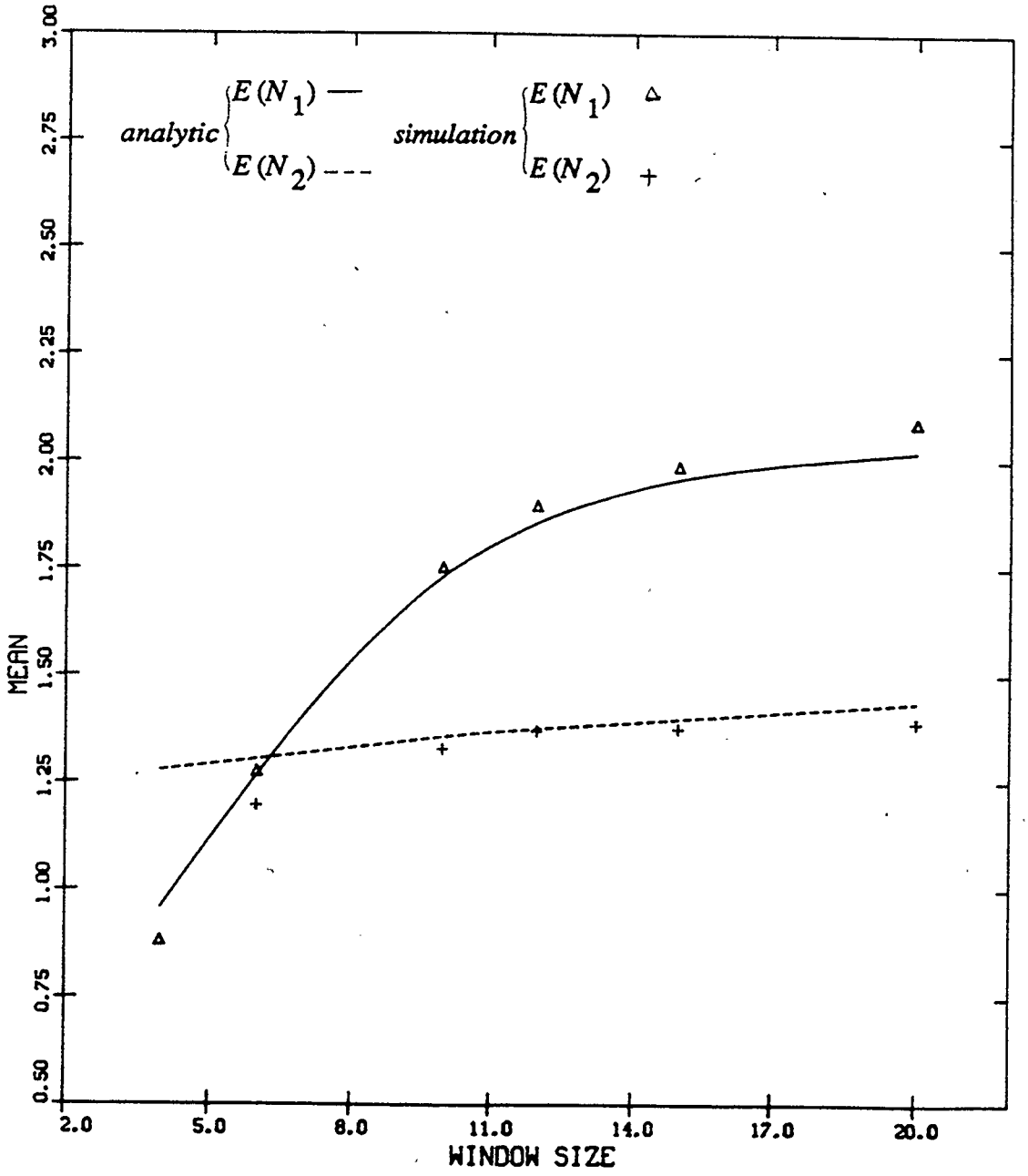


Fig. 5.13 Mean number of packets v.s window size n_W (C_1=1, C_2=3)

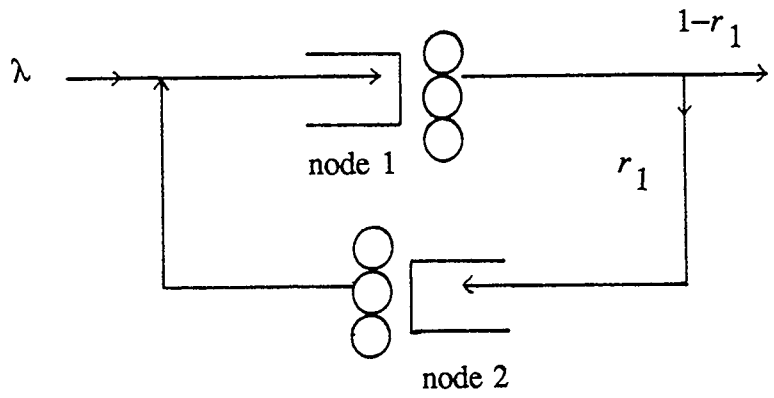


Fig. 5.14 The feedback network model

Table 5.1 Mean number of packets at node 2

C_2	ρ_2	$c_{Y_2}^2$	Analysis	Simulation
2	0.35	0	0.879	0.816 ± 0.068
2	0.60	0	1.986	1.873 ± 0.127
2	0.80	0	4.350	4.139 ± 0.151
3	0.30	0	0.988	0.945 ± 0.076
3	0.60	0	1.865	1.795 ± 0.115
3	0.80	0	4.985	4.983 ± 0.243
5	0.40	0	2.130	2.063 ± 0.042
5	0.60	0	3.546	3.527 ± 0.132
5	0.80	0	6.342	6.500 ± 0.391
2	0.35	1	0.942	0.852 ± 0.105
2	0.60	1	2.362	2.017 ± 0.195
2	0.85	1	8.157	7.451 ± 0.445
3	0.35	1	1.210	1.122 ± 0.088
3	0.60	1	2.760	2.467 ± 0.124
3	0.85	1	8.650	7.765 ± 0.579
5	0.35	1	1.827	1.775 ± 0.116
5	0.60	1	3.700	3.464 ± 0.142
5	0.85	1	9.825	8.916 ± 0.648

Table 5.2 Mean E(N_2)

p	λ_2	ρ_2	Analysis	Simulation
0.5	0	0.2	0.343	0.267 ± 0.052
0.5	0	0.5	1.375	1.273 ± 0.124
0.5	0	0.8	5.50	6.241 ± 0.51
0.8	0	0.167	0.217	0.204 ± 0.016
0.8	0	0.278	0.42	0.398 ± 0.024
0.8	0	0.625	1.823	1.824 ± 0.138
0.8	0	0.833	5.468	5.688 ± 0.358
0.8	0.85	0.2	0.252	0.251 ± 0.067
0.8	0.85	0.5	1.011	1.004 ± 0.106
0.8	0.85	0.8	4.044	4.045 ± 0.261
0.8	2.15	0.2	0.251	0.250 + 0.012
0.8	2.15	0.5	1.004	1.003 ± 0.087
0.8	2.15	0.8	4.018	3.998 ± 0.136

Table 5.3 Means $E(N_1)$ and $E(N_2)$ $(m_2 = \infty)$

p	ρ_2	m_1	$E(N_1)$	$E(N_2)$
0.8	0.365	3	0.923 (0.925 ± 0.066)	0.612 (0.495 ± 0.082)
0.8	0.566	3	0.923 (0.922 ± 0.019)	1.389 (1.047 ± 0.071)
0.8	0.731	3	0.923 (0.925 ± 0.053)	2.889 (1.996 ± 0.102)
0.8	0.387	5	1.349 (1.353 ± 0.061)	0.684 (0.614 ± 0.028)
0.8	0.600	5	1.349 (1.346 ± 0.054)	1.624 (1.395 ± 0.089)
0.8	0.774	5	1.349 (1.348 ± 0.077)	3.711 (3.062 ± 0.218)
0.8	0.400	10	1.881 (1.907 ± 0.115)	0.730 (0.698 ± 0.043)
0.8	0.621	10	1.881 (1.889 ± 0.114)	1.791 (1.777 ± 0.125)
0.8	0.801	10	1.881 (1.881 ± 0.227)	4.410 (4.459 ± 0.235)
0.8	0.403	20	2.072 (2.068 ± 0.155)	0.738 (0.718 ± 0.042)
0.8	0.625	20	2.072 (2.083 ± 0.121)	1.822 (1.867 ± 0.104)
0.8	0.806	20	2.072 (2.062 ± 0.106)	4.553 (4.765 ± 0.228)
0.55	0.390	10	4.065 (4.069 ± 0.165)	0.735 (0.578 ± 0.022)
0.55	0.548	10	4.065 (4.068 ± 0.136)	1.389 (1.046 ± 0.078)
0.55	0.702	10	4.065 (4.071 ± 0.178)	2.712 (1.893 ± 0.186)
0.55	0.358	20	7.630 (7.642 ± 0.271)	0.618 (0.543 ± 0.035)
0.55	0.560	20	7.633 (7.641 ± 0.201)	1.401 (1.239 ± 0.159)
0.55	0.779	20	7.631 (7.597 ± 0.177)	3.901 (3.323 ± 0.218)

Table 5.4 Means E(N_4) and E(N_5)

$\lambda_1=\lambda_2=0.1$, $\lambda_3=0.2$, $\mu_1=0.1$, $\mu_2=0.2$, $\mu_3=0.5$, $\mu_4=0.167$

(a)

$p_1=p_2$	p_3	ρ_4	λ_5	$E(N_4)$
0.8	0.8	0.30	0.5	1.005 (0.939 ± 0.056)
0.8	0.8	0.50	0.5	1.989 (1.794 ± 0.118)
0.8	0.8	0.80	0.5	6.375 (5.575 ± 0.296)
0.6	0.8	0.35	0.5	1.271 (1.122 ± 0.094)
0.6	0.8	0.58	0.5	2.701 (2.328 ± 0.157)
0.6	0.8	0.80	0.5	7.006 (5.684 ± 0.375)

(b)

p_1	$p_2=p_3$	ρ_5	λ_5	$E(N_5)$
0.6	0.8	0.416	0.5	1.441 (1.367 ± 0.107)
0.6	0.8	0.6	0.5	2.557 (2.373 ± 0.261)
0.6	0.8	0.8	0.5	5.727 (5.259 ± 0.466)
0.8	0.8	0.4	0.5	5.532 (5.264 ± 0.078)
0.8	0.8	0.6	0.5	1.348 (1.316 ± 0.186)
0.8	0.8	0.8	0.5	2.492 (2.372 ± 0.339)

(c)

$p_1=p_2=p_3$	ρ_5	λ_5	$E(N_5)$
0.8	0.4	5.0	1.302 (1.294 ± 0.008)
0.8	0.6	5.0	2.356 (2.331 ± 0.092)
0.8	0.8	5.0	5.079 (5.003 ± 0.138)

Table 5.5 Mean and variance

(a)

analytic (c_Y^2=0) λ_g=0.4, μ=1, p=0.8			analytic (c_Y^2=0) λ_g=0.4, μ=1, p=0.6		
stage i	$E(N_i)$	$Var(N_i)$	stage i	$E(N_i)$	$Var(N_i)$
1	1.000	2.000	1	2.667	11.56
2	0.750	0.938	2	1.333	2.222
3	0.687	0.730	3	1.111	1.358
4	0.672	0.682	4	1.074	1.233
5	0.668	0.671	5	1.068	1.213
6	0.667	0.668	6	1.067	1.209

(b)

analytic (c_Y^2=0.5) λ_g=0.4, μ=1, p=0.8			analytic (c_Y^2=0.5) λ_g=0.4, μ=1, p=0.6		
stage i	$E(N_i)$	$Var(N_i)$	stage i	$E(N_i)$	$Var(N_i)$
1	1.125	2.672	1	2.999	14.99
2	0.906	1.558	2	1.778	4.543
3	0.852	1.324	3	1.574	3.381
4	0.838	1.268	4	1.540	3.204
5	0.834	1.255	5	1.534	3.175
6	0.837	1.251	6	1.534	3.169

(c)

analytic (c_Y^2=1) λ_g=0.4, μ=1, p=0.8			analytic (c_Y^2=1) λ_g=0.4, μ=1, p=0.6		
stage i	E(N_i)	Var(N_i)	stage i	E(N_i)	Var(N_i)
1	1.250	3.438	1	3.333	18.89
2	1.063	2.324	2	2.222	7.654
3	1.016	2.079	3	2.037	6.262
4	1.004	2.019	4	2.006	6.043
5	1.001	2.005	5	2.001	6.007
6	1.000	2.001	6	2.000	6.001

(d)

analytic (c_Y^2=3) λ_g=0.4, μ=1, p=0.8			analytic (c_Y^2=3) λ_g=0.4, μ=1, p=0.6		
stage i	E(N_i)	Var(N_i)	stage i	E(N_i)	Var(N_i)
1	1.750	7.438	1	4.667	38.89
2	1.688	6.855	2	3.999	27.99
3	1.672	6.714	3	3.889	26.36
4	1.668	6.678	4	3.870	26.09
5	1.667	6.669	5	3.867	26.04
6	1.667	6.667	6	3.867	26.04

Table 5.6 Mean $E(N_i)$

Simulation				
λ_g =0.4, μ=1, p=0.8			λ_g =0.4, μ=1, p=0.6	
c_Y^2	stage i	$E(N_i)$	stage i	$E(N_i)$
0.0	1	1.001	1	2.657
0.0	2	0.677	2	1.222
0.0	3	0.649	3	1.148
0.5	1	1.124	1	3.000
0.5	2	0.892	2	1.798
0.5	3	0.866	3	1.715
1.0	1	1.252	1	3.336
1.0	2	1.068	2	2.250
1.0	3	1.034	3	2.166
3.0	1	1.748	1	4.667
3.0	2	1.693	2	3.993
3.0	3	1.662	3	3.912

Table 5.7 Average transfer delay

(a)

analytic								
	λ_g=0.4, μ=1, p=0.8				λ_g=0.4, μ=1, p=0.6			
n_b	c_Y^2=0	c_Y^2=0.5	c_Y^2=1	c_Y^2=3	c_Y^2=0	c_Y^2=0.5	c_Y^2=1	c_Y^2=3
2	3.500	4.062	4.625	6.875	5.999	7.166	8.333	13.00
3	4.875	5.765	6.656	10.22	7.666	9.527	11.39	18.83
4	6.218	7.441	8.664	13.55	9.277	11.84	14.39	24.64
5	7.554	9.110	10.67	16.89	10.88	14.14	17.39	30.44
6	8.888	10.78	12.67	20.22	12.48	16.44	20.39	36.24

(b)

simulation								
	λ_g=0.4, μ=1, p=0.8				λ_g=0.4, μ=1, p=0.6			
n_b	c_Y^2=0	c_Y^2=0.5	c_Y^2=1	c_Y^2=3	c_Y^2=0	c_Y^2=1	c_Y^2=3	
2	3.354	4.047	4.607	6.852	5.819	7.193	8.337	12.98
3	4.651	5.778	6.675	10.23	7.542	9.765	11.59	18.85

Table 5.8 Performance measures for various window size ($C_1 = C_2 = 1$)

([ ] - simulation results)

n_W	U_1	U_2	$E(N_1)$	$E(N_2)$	Throughput
3	0.4582 [0.4524 $\pm 7 \times 10^{-3}$]	0.4579 [0.4497 $\pm 4 \times 10^{-3}$]	0.755 [0.7145 $\pm 1.3 \times 10^{-3}$]	0.737 [0.6815 $\pm 1 \times 10^{-3}$]	0.0918
6	0.5602 [0.559 $\pm 9 \times 10^{-4}$]	0.5595 [0.564 $\pm 9 \times 10^{-4}$]	1.286 [1.266 $\pm 3 \times 10^{-3}$]	1.212 [1.1688 $\pm 3.5 \times 10^{-3}$]	0.112
8	0.5896 [0.5901 $\pm 8 \times 10^{-4}$]	0.5890 [0.5948 $\pm 10^{-3}$]	1.533 [1.5156 $\pm 4.4 \times 10^{-3}$]	1.417 [1.385 $\pm 5.8 \times 10^{-3}$]	0.118
10	0.6051 [0.599 $\pm 1.4 \times 10^{-3}$]	0.6043 [0.6044 $\pm 7 \times 10^{-3}$]	1.7091 [1.714 $\pm 9 \times 10^{-3}$]	1.554 [1.5435 $\pm 6.9 \times 10^{-3}$]	0.121
12	0.6146 [0.6121 $\pm 7 \times 10^{4}$]	0.6144 [0.6117 $\pm 10^{-2}$]	1.851 [1.851 $\pm 6.5 \times 10^{-3}$]	1.661 [1.661 $\pm 6 \times 10^{3}$]	0.122
14	0.6192	0.619	1.932	1.7177	0.1238
16	0.6217 [0.621 $\pm 10^{-3}$]	0.6215 [0.621 $\pm 1.6 \times 10^{-3}$]	1.983 [2.008 $\pm 1.3 \times 10^{-2}$]	1.753 [1.782 $\pm 1.3 \times 10^{-2}$]	0.124
20	0.6243 [0.627 $\pm 1.2 \times 10^{-3}$]	0.6243 [0.6287 $\pm 1.3 \times 10^{-2}$]	2.0196 [2.064 $\pm 1.7 \times 10^{-2}$]	1.778 [1.811 $\pm 1.1 \times 10^{-2}$]	0.1248

Table 5.9 Performance measures for various window size ($C_1 = 1$, $C_2 = 3$)

([ ] - simulation results)

n_W	U_1	U_2	$E(N_1)$	$E(N_2)$	Throughput
4	0.475	0.317	0.8968 [0.878 ±0.203]	1.2266 [1.01 ±0.0012]	0.095
6	0.561	0.375	1.2967 [1.273 ±0.004]	1.3044 [1.193 ±0.003]	0.112
8	0.594	0.397	1.5669	1.3577	0.119
10	0.609	0.407	1.7458 [1.756 ±0.0078]	1.385 [1.326 ±0.005]	0.122
12	0.615	0.412	1.857 [1.907 ±0.009]	1.397 [1.367 ±0.0031]	0.123
15	0.6204	0.415	1.954 [2.004 ±0.01]	1.405 [1.375 ±0.004]	0.124
18	0.6214	0.4158	1.898	1.4063	0.1247
20	0.6222	0.4162	2.0161 [2.111 ±0.019]	1.4064 [1.384 ±0.004]	0.1248

Table 5.10(a) Mean $E(N_1)$

$\lambda_0 = c_{T_0} = 1.0$, $\mu = 0.8$, $\mu_2 = 0.4$, $r_1 = 0.5$, $\rho_1 = \rho_2 = 0.833$

c_{Y1}^2	c_{Y2}^2	I	II	III	ME1	Simulation
0.0	0.0	3.09	3.07	3.24	3.31	3.83
0.0	0.5	3.47	3.45	3.72	3.84	4.22
0.0	1.0	3.79	3.80	4.15	4.21	4.84
0.0	1.5	4.07	4.09	4.51	4.574	4.28
0.0	3.0	4.69	4.78	5.36	5.386	5.13
0.5	0.0	4.39	4.20	4.21	4.272	4.55
0.5	0.5	4.8	4.6	4.72	4.767	5.67
0.5	1.0	5.14	4.97	5.15	5.188	4.98
0.5	1.5	5.42	5.28	5.53	5.482	5.81
0.5	3.0	6.04	5.98	6.38	6.36	6.78
1.5	0.0	7.17	6.64	6.27	6.25	5.16
1.5	0.5	7.57	7.07	6.79	6.75	6.45
1.5	1.0	7.91	7.44	7.23	7.17	5.87
1.5	1.5	8.18	7.75	7.60	7.52	6.18
1.5	3.0	8.76	8.45	7.93	8.33	6.67

I: Diffusion approximation (without traffic conservation)
II: Diffusion approximation (from section 3.6.1)
III: Diffusion approximation (from section 3.6.2)

Table 5.10(b) Mean E(N_2)

c_{Y1}^2	c_{Y2}^2	I	II	III	ME1	Simulation
0.0	0	3.18	3.17	3.36	3.43	3.91
0.0	0.5	4.49	4.30	4.336	4.39	4.58
0.0	1.0	5.86	5.5	5.35	5.374	6.43
0.0	1.5	7.25	6.73	6.38	6.35	4.98
0.0	3.0	11.48	10.47	9.51	9.34	6.54
0.5	0	3.48	3.47	3.74	3.82	4.26
0.5	0.5	4.81	4.63	4.74	4.79	5.72
0.5	1.0	6.19	5.85	5.76	5.67	5.17
0.5	1.5	7.58	7.08	6.8	6.71	5.61
0.5	3.0	11.79	10.83	9.64	9.75	10.38
1.5	0	4.09	4.11	4.54	4.56	4.57
1.5	0.5	5.45	5.31	5.57	5.59	5.85
1.5	1.0	6.83	6.55	6.61	6.56	5.12
1.5	1.5	8.22	7.79	7.65	7.52	6.19
1.5	3.0	12.39	11.54	10.80	10.07	6.73

CHAPTER 6

QUEUES AND NETWORKS IN DISCRETE TIME

6.1 INTRODUCTION

In the literature most of the works on queueing system analysis deals with the continuous time models. However, in many cases it is more natural to use discrete-time models for performance analysis. Examples include processor-based systems and time-slotted packet switching systems where instructions are executed in an integral number of fixed time slots.

There are a number of studies [82-88] which deal with the analysis of discrete-time single server queues. Konheim [88] obtained a result for the waiting time distribution of a discrete-time G/G/1 queue employing polynomial factorization. Effective algorithms for the calculation of the waiting time distribution of the discrete-time G/G/1 queue were presented by Ackroyd [82]. Tran-Gia [87] obtained the distribution of interdeparture time by using a numerical method based on the fast Fourier Transform. Pujolle et al [85] used Markov chains to analyze single queues and simple queueing networks with nodes connected in series. They obtained a product form solution for the state probability distribution of the network.

In this chapter, discrete-time queues and queueing networks will be investigated by using the method of entropy maximization. The state probability distributions of the G/G/1 and the $Geom^X/G/1$ queues are derived and the statistics

of packet arrivals in a compound-binomial input process will be obtained. A decomposition technique is then used to study queueing networks and the results obtained are applied to study a buffered multistage interconnection network.

6.2 EQUILIBRIUM STATE PROBABILITY DISTRIBUTION OF THE GEOM^X/G/1 QUEUE

In this section, we shall apply the method of entropy maximization to study the Geom^X/G/1 queue. Closed-form expression for the state probability distribution at equilibrium will be derived and Quasi-Geometric (QGeom) distribution will be introduced.

In order to formulate the problem we make the following assumptions :

1. Time axis is divided into equal units of length $\Delta t = 1$. Arrivals and services can take place only at those epochs of the discrete time axis.

2. The interarrival times T_g of batches are mutually independent, identically and geometrically distributed with mean $1/\lambda_g$, i.e.

$$P\{T_g=k\} = \lambda_g(1-\lambda_g)^{k-1} \quad , k=1,2,\cdots \tag{6.2.1}$$

where λ_g is the probability that one arrival occurs in $\Delta t = 1$.

3. The batch size B is also geometrically distributed with mean $1/p$, i.e.

$$P\{B=i\} = p(1-p)^{i-1} \quad , i=1,2,\cdots$$

4. The service time Y has independent and identical general distribution with mean $1/\mu$ and coefficient of variation c_Y, whose generating function is $G_Y(z)$. The service time Y, interarrival time T_g and batch size B are mutually independent random variables.

5. The queue discipline is FIFO and the traffic intensity is $\rho = \lambda_g / \mu p < 1$.

To determine the state probability distribution $\{p_k\}$, k= 0, 1, 2,..., of the $Geom^X/G/1$ queue, the information about the utilization $1-p_0$ and mean number of customers in the system E(N) are needed. According to Konheim [84], these two quantities can be expressed as

$$p_0 = 1 - \rho \tag{6.2.2}$$

and

$$E(N) = \frac{\rho(2-\lambda_g) + \rho^2 p (c_Y^2 - 1)}{2(1-\rho)} \tag{6.2.3}$$

This problem can be formulated as a constrained optimization problem :

$$\textit{Maximize} \quad H = -\sum_{k=0}^{\infty} p_k \ln p_k$$

subject to the constraints

$$\sum_{k=0}^{\infty} p_k = 1$$

$$p_0 = 1 - \rho$$

and

$$\sum_{k=0}^{\infty} k\, p_k = E(N)$$

By the use of the necessary conditions (2.2.4)-(2.2.5), we find the optimal distribution $\{p_k\}$ for maximum value of H as

$$p_k = \begin{cases} 1-\rho \quad , & \text{if } k = 0 \\ (1-\rho)a_2 a_2^k \ , & \text{if } k = 1, 2,... \end{cases} \tag{6.2.4}$$

where

$$a_1 = \frac{2\rho p}{\rho p - 2p + 2 - \lambda_g + \rho p c_Y^2}$$

and

$$a_2 = \frac{2 - 2p + p\rho - \lambda_g + p\rho c_Y^2}{2 - \rho p - \lambda_g + p\rho c_Y^2}$$

Konheim [84] also showed that the generating function of the state probability distribution for a $Geom^X/G/1$ queue is given by

$$G_N(z) = \frac{(1-\rho)(z-1)G_Y\left[1-\lambda_g + \lambda_g\, G_B(z)\right]}{z - G_Y\left[1-\lambda_g + G_B(z)\right]} \tag{6.2.5}$$

The generating function of the ME state probability distribution in (6.2.4) is

$$G_N(z) = \frac{(1-\rho)[1-a_2(1-a_1)z]}{1-a_2 z} \tag{6.2.6}$$

Suppose that the $G_N(z)$ in (6.2.5) and (6.2.6) are equal. Then equating the right-hand sides of (6.2.5) and (6.2.6) and solving for $G_Y(z)$, we obtain

$$G_Y(z) = \frac{c_Y^2+\mu-1 - (c_Y^2-\mu-1)z}{1+\mu+c_Y^2 - (c_Y^2-\mu+1)z} \tag{6.2.7}$$

Now taking the inverse z-transform results in the equivalent service time distribution

$$P\{\, Y = k \,\} = \begin{cases} \dfrac{c_Y^2+\mu-1}{1+\mu+c_Y^2}, & \text{if } k = 0 \\ \dfrac{4\mu}{(1+\mu+c_Y^2)^2}\left[\dfrac{1+c_Y^2-\mu}{1+\mu+c_Y^2}\right]^{k-1}, & \text{if } k = 1,2,\ldots \end{cases} \tag{6.2.8}$$

This probability distribution may be regarded as the ME probability distribution of the service time of a $\text{Geom}^X/G/1$ queue. In other words, if the

service time distribution is specified by (6.2.8), the associated ME state probability (6.2.4) of the single-server queue with Geom^X input becomes exact. Note that the distribution in (6.2.8) consists of two parameters, the mean $1/\mu$ and the coefficient of variation c_Y. In general, a random variable with distribution of the form of (6.2.8) will be said to have a quasi-geometric distribution.

In the case of constant service time where $\mu=1$, $c_Y^2=0$, (6.2.8) reduces to

$$p\{ Y = k \} = \begin{cases} 1, & \text{if } k=1 \\ 0, & \textit{otherwise} \end{cases}$$

However, if $c_Y^2 = 1-\mu$, then (6.2.8) reduces to a geometric distribution with mean $1/\mu$.

6.3 THE WAITING TIME DISTRIBUTION AND THE STATE PROBABILITY DISTRIBUTION OF THE QGeom/QGeom/1 QUEUE

In the following we shall first consider the waiting time distribution of the QGeom/QGeom/1 queue. Let W be the waiting time , V be the difference between the service time Y and the interarrival time T. Let $w(k) =$ P $\{W = k\}$ and let $v(k) =$ P$\{V = k\}$. Lindley's equation [82] of the discrete-time G/G/1 queue is given by

$$w(k) = \left[\sum_{l=0}^{\infty} w(l)\, v(k-l) \right] u(k) \tag{6.3.1}$$

where u(k) is the unit step function.

Denote the generating function of $P\{T = k\}$, $P\{Y = k\}$, and $P\{V = k\}$ by $G_T(z)$, $G_Y(z)$ and $G_V(z)$, respectively. Since V is the difference between Y and T, the generating function $G_V(z)$ of P{V=k} is given by

$$G_V(z) = G_T(1/z)\, G_Y(z) \tag{6.3.2}$$

For the QGeom/QGeom/1 queue, we have

$$P\{\ T=k\ \} = \begin{cases} \dfrac{(c_T^2+\lambda-1)}{(c_T^2+\lambda+1)}, & \text{if } k = 0 \\ \dfrac{4\lambda}{(1+c_T^2+\lambda)^2} \left[\dfrac{1+c_T^2-\lambda}{1+c_T^2+\lambda} \right]^{k-1}, & \text{if } k = 1, 2,... \end{cases}$$

and

$$P\{\ Y=k\ \} = \begin{cases} \dfrac{c_Y^2+\mu-1}{1+\mu+c_Y^2}, & \text{if } k = 0 \\ \dfrac{4\mu}{(1+\mu+c_Y^2)^2}\left[\dfrac{1+c_Y^2-\mu}{1+\mu+c_Y^2}\right]^{k-1}, & \text{if } k = 1, 2,... \end{cases}$$

where c_T and c_Y are the coefficients of variation of T and Y, respectively.

By introducing the notations

$$A = c_T^2-\lambda-1$$

$$B = c_T^2-\lambda+1$$

$$K = c_Y^2-\mu-1$$

$$R = c_Y^2-\mu+1$$

the generating functions of P{T=k} and P{Y=k} are given by

$$G_T(z) = \frac{A+2\lambda-Az}{B+2\lambda-Bz}$$

and

$$G_Y(z) = \frac{K+2\mu-Kz}{R+2\mu-Rz}$$

Since $G_T(z)$ and $G_Y(z)$ are rational functions of z , $G_V(z)-1$ can be decomposed into a product of two factors, i.e.,

$$G_V(z) - 1 = G_T(\frac{1}{z})G_Y(z) - 1$$

$$= \frac{(z-1)[RB - AK + 4\mu - (BR - AK + 4\lambda)z]}{[B - z(2\lambda + B)](R + 2\mu - Rz)}$$

The poles are

$$z_1 = \frac{B}{2\lambda + B}$$

$$z_2 = \frac{R + 2\mu}{R}$$

and the zero is

$$z_0 = \frac{BR - AK + 4\mu}{BR - AK + 4\lambda}$$

where z_0, z_2 are outside the unit circle, z_1 is inside the unit circle.

Let

$$H(z) = \frac{[BR - AK + 4\mu - z(BR - AK + 4\lambda)](z-1)}{R + 2\mu - Rz}$$

and

$$F(z) = \frac{1}{B - z(2\lambda + B)}$$

where F(z) contains all the poles and zeroes inside the unit circle and H(z) contains a zero at z=1 and all the other poles and zeroes outside the unit circle. Konheim [88] showed that the generating function of the waiting time distribution $G_W(z)$ is related to H(z) by

$$G_W(z) = \frac{\Pi_0 (z-1)}{H(z)} \tag{6.3.3}$$

where Π_0 is a normalization constant.

Thus we have

$$G_W(z) = \frac{2(1-\rho)(R+2\mu-Rz)}{BR-AK+4\mu-z(BR-AK+4\lambda)} \tag{6.3.4}$$

Using the fact that $G_W(1)=1$, we have

$$\Pi_0 = \frac{dH(z)}{dz}|_{z=1} \tag{6.3.5}$$

$$=2(1-\rho)$$

Using $G_W(z)$ we calculate the probability of zero waiting time to be

$$w_0 = G_W(0) \tag{6.3.6}$$

$$=\frac{(1-\rho)(1+\mu+c_Y^2)}{\mu-\lambda+c_T^2+c_Y^2}$$

From (6.3.3) the average waiting time is then given by

$$E(W) = \frac{\rho(1+c_Y^2)-1+c_T^2}{2\mu(1-\rho)} \tag{6.3.7}$$

Using Little's formula, we obtain

$$E(N) = \frac{\rho(1-\rho+c_T^2+\rho c_Y^2)}{2(1-\rho)} \tag{6.3.8}$$

Finally applying E(N) into (2.3.6a) and (2.3.6b), from (2.3.5) we obtain the state probability distribution $\{p_k\}$ of the QGeom/QGeom/1 queue

$$p_k = \begin{cases} 1-\rho, \text{ if } k = 0 \\ \dfrac{2\rho(1-\rho)}{\rho(1+c_Y^2)+c_T^2-1}\left[\dfrac{-1+\rho+c_T^2+\rho c_Y^2}{1-\rho+c_T^2+\rho c_Y^2}\right]^k, \text{ if } k = 1, 2,... \end{cases} \tag{6.3.9}$$

from which we can calculate the variance of number of customers in the system

$$\sigma_N{}^2 = \frac{\rho(1-\rho+c_T^2+\rho c_Y^2)(2c_T^2+2\rho c_Y^2-\rho+\rho^2-\rho c_T^2-\rho^2 c_Y^2)}{4(1-\rho)^2} \tag{6.3.10}$$

6.4 STATISTICS OF PACKET ARRIVALS IN A MESSAGE ARRIVAL PROCESS

This section will establish an equivalence relationship between a message arrival process and a QGeom distribution. This is a discrete time analog of the bulk Poisson process in continuous time. We shall apply the method of entropy maximization to obtain the interarrival time distribution for the packets.

Consider an input process, where the message arrivals follow a discrete-time random process. We shall make the following assumptions :

1. Message arrivals follow a random process with rate λ_g messages/second.

2. Each message consists of a random number B of packets with mean $E(B)$ packets/message.

3. Let T be a random variable denoting the interarrival time of packets with probability distribution P{T=k}.

First we note that the probability of zero interarrival time is nonzero. Therefore we may write

$$\phi_k = P\{ T = k \}, \quad k=0,1,2,...$$

Then

$$P\{ T = 0 \} = \phi_0 = 1-\frac{1}{E(B)}$$

Second, the arrival rate of packets is equal to λ_g/p which is the product of the message rate λ_g and the average message length $1/p$.

With these information, we shall derive the probability distribution for the interarrival time of packets. The problem is formulated to maximize the entropy function

$$H = -\sum_{k=0}^{\infty} \phi_k \ln \phi_k$$

subject to the constraints

$$\phi_0 = 1 - \frac{1}{E(B)}$$

$$\sum_{k=0}^{\infty} k\ \phi_k = \frac{1}{\lambda_g E(B)}$$

$$\sum_{k=0}^{\infty} \phi_k = 1$$

Then using the necessary conditions (2.3.4a)-(2.3.4c), we obtain

$$\phi_k = \begin{cases} 1-\dfrac{1}{E(B)}\ , & \text{if } k = 0 \\ \dfrac{\lambda_g\ (1-\lambda_g)^{k-1}}{E(B)}\ , & \text{if } k = 1,2,... \end{cases} \tag{6.4.1}$$

from which we calculate the coefficient of variation c_T as

$$c_T = [(2-\lambda_g)E(B) - 1]^{1/2}$$

We can express the probability ϕ_k in (6.4.1) as a QGeom distribution in terms of its mean $1/\lambda$ and coefficient of variation c_T as follows.

$$\phi_k = \begin{cases} \dfrac{(c_T^2+\lambda-1)}{(c_T^2+\lambda+1)}\ , & \text{if } k = 0 \\ \dfrac{4\lambda}{(1+c_T^2+\lambda)^2} \left[\dfrac{1+c_T^2-\lambda}{1+c_T^2+\lambda}\right]^{k-1}\ , & \text{if } k = 1, 2,... \end{cases} \tag{6.4.2}$$

We shall state an equivalent relationship between the message arrival process and the associated QGeom interarrival time distribution $\{\phi_k\}$ of packets as follows

Message	Packet
Mean message arrival rate λ_g Mean message size $E(B)$	*QGeom* packet interarrival-time distribution $\phi_k = \begin{cases} \dfrac{c_T^2 - 1 + \lambda}{c_T^2 + 1 + \lambda}, & \text{if } k = 0 \\ \dfrac{4\lambda}{(1+\lambda+c_T^2)^2}\left[\dfrac{1-\lambda+c_T^2}{1+\lambda+c_T^2}\right]^{k-1}, & \text{if } k = 1,2,\cdots \end{cases}$ with $\lambda = \lambda_g E(B),\ c_T^2 = (2-\lambda_g)E(B) - 1$ (6.4.3)

In particular, if message arrivals follow a binomial process with length or size following a geometric distribution, i.e., $P\{B=i\} = p(1-p)^{i-1}$, i=1,2,..., then

$$E(B) = \frac{1}{p}$$

and (6.4.1) becomes

$$\phi_k = \begin{cases} 1-p, & \text{if } k = 0 \\ \lambda_g\, p\,(1-\lambda_g)^{k-1} & \text{if } k=1,2,\ldots \end{cases} \qquad (6.4.4)$$

Note that for the case p=1, this equivalence reduces to a relation between the binomial process and its associated geometric interarrival time distribution.

We shall find an equivalence relation between the $Geom^X/G/1$ queue and the

QGeom/G/1 queue. Suppose that the message length has a geometric distribution with mean $1/p$, from (6.2.15) the generating function of the state probability distribution of the Geom^X/G/1 queue is given by

$$G_N(z) = \frac{(1-\rho)(z-1)G_Y\left[\dfrac{1-\lambda_g+z(\lambda_g-1+p)}{1-(1-p)z}\right]}{z - G_Y\left[\dfrac{1-\lambda_g+z(\lambda_g-1+p)}{1-(1-p)z}\right]}$$

Employing the equivalence equations (6.4.3), we obtain the generating function of the state probability distribution of the QGeom/G/1 queue.

$$G_N(z) = \frac{(1-\rho)(z-1)G_Y\left[\dfrac{1-\lambda+c_T^2+z(1+\lambda-c_T^2)}{1+\lambda+c_T^2+z(1-\lambda-c_T^2)}\right]}{z-G_Y\left[\dfrac{1-\lambda+c_T^2+z(1+\lambda-c_T^2)}{1+\lambda+c_T^2+z(1-\lambda-c_T^2)}\right]} \tag{6.4.5}$$

Suppose that the service time has the QGeom distribution as in (6.2.8). Applying (6.2.8) to (6.4.5) we obtain the generating function of the state probability distribution of the QGeom/QGeom/1 system.

$$G_N(z) = (1-\rho)\left[\frac{1-(1-a_1)a_2 z}{1-a_2 z}\right] \tag{6.4.6}$$

where

$$a_1 = \frac{2\rho}{\rho(1+c_Y^2) + c_T^2 - 1}$$

and

$$a_2 = \frac{c_T^2 + \rho c_Y^2 - 1 + \rho}{c_T^2 + \rho c_Y^2 + 1 - \rho}$$

Taking the inverse z-transform of (6.4.6), we obtain the state probabilities

$$p_k = \begin{cases} 1-\rho \quad , \text{ if } k = 0 \\ (1-\rho) a_1 \, a_2^{\,k} \, , \text{ if } k \geq 1 \end{cases}$$

Note that this result agrees with that obtained in (6.3.9).

6.5 SYNCHRONIZED OPEN QUEUEING NETWORKS

6.5.1 Description of the Open Network

Consider the synchronized queueing network consisting of L nodes and satisfying the following conditions (See Fig. 6.1).

1. Each node behaves like a G/G/1 queue. Service times of customers at node i are mutually independent and identically distributed with mean $1/\mu_i$ and coefficient of variation c_Y.

2. External arrivals constitute a renewal process with mean $1/\lambda_0$ and variation coefficient c_{T_0}. Customers first entering the system are directed to node i with probability r_{0i} , i=1,..., L.

3. Once served at node i, a customer requests service from node j, j=1,..L, with probability r_{ij} or leaves the system with probability

$$r_{i,L+1} = 1 - \sum_{j=1}^{L} r_{i\,j}$$

4. Queue discipline is FIFO at each node.

Let λ_i be the customer arrival rate at node i which must satisfy the law of conservation of traffic

$$\lambda_i = \lambda_0 r_{0i} + \sum_{j=1}^{L} \lambda_j r_{ji} \tag{6.5.1}$$

The utilization of server i is given by

$$\rho_i = \frac{\lambda_i}{\mu_i} < 1 \ , \ i = 1,2,..., L$$

The state of the network can be described by a vector **n**= $(n_1, \ldots, n_L)$, $0 \le n_i \le \infty$, $1 \le i \le L$. and $p(\mathbf{n})$ is defined as the steady state probability that the system is in state **n**. The marginal state probability that node i is in state n_i is defined as

$$p_i(n_i) = \sum_{\substack{all\ n_j \\ 1 \le j \le L,\ j \ne i}} p(\mathbf{n}) \tag{6.5.2}$$

The average number of customers at node i is given by (6.3.8) as

$$E(N_i) = \frac{\rho_i(1 - \rho_i + c_{T_i}{}^2 + \rho_i\, c_{Y_i}{}^2)}{2(1-\rho_i)} \tag{6.5.3}$$

In order to determine the joint state probability distribution of the network , we maximize the network entropy function

$$H = -\sum_{\mathbf{n}} p(\mathbf{n}) \ln p(\mathbf{n})$$

subject to the constraints

$$1 - p_i(0) = \rho_i \quad , 1 \le i \le L$$

$$\sum_{n_i=0}^{\infty} n_i\, p_i(n_i) = E(N_i) \quad , 1 \le i \le L$$

and

$$\sum_{n_i=0}^{\infty} p_i(n_i) = 1 \quad , 1 \le i \le L$$

Following the same procedure as Section 2.3, we obtain

$$p(\mathbf{n}) = \prod_{i=1}^{L} (1-\rho_i) \left[\frac{\rho_i^2}{(1-\rho)(E(N_i)-\rho_i)} \right]^{h_i(n_i)} \left[\frac{E(N_i)-\rho_i}{E(N_i)} \right]^{n_i} \tag{6.5.4}$$

where

$$h_i(n_i) = \begin{cases} 0 & \text{, if } n_i = 0 \\ 1 & \text{, if } n_i \geq 1 \end{cases}$$

Now we can write (6.5.4) in the product form

$$p(\mathbf{n}) = \prod_{i=1}^{L} p_i(n_i) \tag{6.5.5}$$

where

$$p_i(n_i) = (1-\rho_i)\left[\frac{\rho_i^2}{(1-\rho_i)(E(N_i)-\rho_i)}\right]^{h_i(n_i)} \left[\frac{E(N_i)-\rho_i}{E(N_i)}\right]^{n_i} \tag{6.5.6}$$

6.5.2 Flows in the Network

The significance of the solution (6.5.5) is that each node of the network can be treated as a separate G/G/1 queueing system. However, to implement (6.5.5), the flows in the network, mainly the interarrival and interdeparture processes, must be identified. In the following, we shall consider each queue of the network as a QGeom/QGeom/1 queue. First, we shall determine the interdeparture time distribution and then study the superposition and the decomposition of flow processes in the network.

(a) Interdeparture time of the QGeom/QGeom/1 queue

Marshall [89] obtained the idle time and the interdeparture time distributions of the G/G/1 queue. Starting from the discrete version of Lindley's equation (6.3.1)

and following similar arguments set out as Marshall, we obtain the generating function of the idle time , $G_{I_i}(z)$ and the generating function of the interdeparture time distributions, $G_{D_i}(z)$, in terms of the interarrival time and service time distribution as follows

$$G_{I_i}(z) = 1 - \frac{1}{w_{0_i}} G_{W_i}(\frac{1}{z}) \left[1 - G_{T_i}(z) G_{Y_i}(\frac{1}{z}) \right] \tag{6.5.7}$$

and

$$G_{D\,i}(z) = p_{e\,i} G_{I\,i}(z) G_{Y\,i}(z) + (1 - p_{e\,i}) G_{Y_i}(z) \tag{6.5.8}$$

where $G_{T_i}(z)$ and $G_{Y_i}(z)$ are respectively the generating function of the interdeparture time and service time distributions of node i, $G_{W_i}(z)$ and w_{0_i} are given in (6.3.4) and (6.3.6) respectively. The parameter p_{e_i} denotes the probability for a departure to leave an empty system which is given by [89]

$$p_{e_i} = \frac{(\frac{1}{\lambda_i} - \frac{1}{\mu_i})}{\frac{dG_{I_i}(z)}{dz}\Big|_{z=1}} \tag{6.5.9}$$

Employing (6.3.4), and (6.3.6) for $G_{W_i}(z)$, and w_{0_i}, we obtain the generating function of the idle time distribution

$$G_{I_i}(z) = 1 - \frac{2(1-z)(\mu_i - \lambda_i + c_{T_i}^2 + c_{Y_i}^2)}{(1+\mu_i + c_{Y_i}^2)[c_{T_i}^2 + \lambda_i + 1 - z(c_{T_i}^2 - \lambda_i + 1)]} \tag{6.5.10}$$

Applying (6.5.10) to (6.5.9) yields

$$p_{e_i} = \frac{(1-\rho_i)(1+\mu_i + c_{Y_i}^2)}{\mu_i - \lambda_i + c_{T_i}^2 + c_{Y_i}^2} \tag{6.5.11}$$

Applying p_{e_i}, $G_{I_i}(z)$ and $G_{Y_i}(z)$ to (6.5.8) gives the mean interdeparture time

$$E(D_i) = \frac{1}{\lambda_i} \tag{6.5.12}$$

and the coefficient of variation

$$c_{D_i} = \sqrt{\rho_i - \rho_i^2 + \rho_i^2 c_{Y_i}^2 + (1-\rho_i) c_{T_i}^2} \tag{6.5.13}$$

(b) The decomposition process

Suppose that a discrete-time arrival process with the generating function of the interarrival time distribution $G_{T_i}(z)$ is split into j components with probability r_{ij}. Then the generating function of the interarrival time distributiom of each component process is given by [90]

$$G_{T_{ij}}(z) = \frac{r_{ij}\, G_{T_i}(z)}{1 - [1 - r_{ij}\, G_{T_i}(z)]} \tag{6.5.14}$$

From which we can calculate the mean and the coefficient of variation as

$$E(T_{ij}) = \frac{1}{\lambda_i\, r_{ij}} \tag{6.5.15}$$

and

$$c_{T_{ij}} = \sqrt{1 + r_{ij}(c_{T_i}{}^2 - 1)} \tag{6.5.16}$$

(c) The superposition process

From the literature [83], it has been found that superposition of geometric distributions is not geometric in contrast to the exponential distribution case. As for the case of nongeometric discrete-time processes, there is no available information. Suppose that L component processes are specified by their means $1/\lambda_i$ and second moments $E(T_i{}^2)$ associated with the interarrival timess T_i, i=1,..,L. As an approximation, we let

$$E(T^2) = \sum_{i=1}^{L} \frac{\lambda_i\, E(T_i{}^2)}{\lambda} \tag{6.5.17}$$

to be the second moment of the interarrival time for this superposition process in which the mean arrival rate must satisfy the law of conservation of traffic

$$\lambda = \sum_{i=1}^{L} \lambda_i$$

Then the coefficient of variation is defined as

$$c_T = \sqrt{[E(T^2) - E^2(T)]/E^2(T)} \tag{6.5.18}$$

$$= \left[\frac{\lambda}{\sum_{i=1}^{L} \frac{\lambda_i}{1+c_{T_i}^2}} - 1 \right]^{1/2}$$

As an example, we consider the superposition of L geometric inputs with the same arrival rate $\lambda_i = p$ and coefficient of variation of the interarrival time $c_{T_i} = \sqrt{1-p}$. The service time is assumed to be a time slot. Substituting these λ_i and c_{Ti} into (6.5.16) for c_T, then using (6.3.8), we obtain the mean number of customers in the system

$$E(N) = \frac{(L-1)p}{2(1-Lp)}$$

Note that this result is correct and agrees with that obtained by Kruskal [79].

6.6 OPEN QUEUEING NETWORKS WITH BULK ARRIVALS

In this section, we shall study synchronized queueing networks with bulk arrivals. This model is an extension of the network decribed in Section 6.5.1. Consider the discrete-time open network at equilibrium with assumptions 1, 3, 4 in Section 6.5 and an additional assumption

5. The network is operating in discrete time, i.e., events such as arrivals and services occur only at the beginning of time slots. External arrivals constitute a compound-binomial process where the interarrival time and the number of customers of a group are geometric with means $1/\lambda_{g_i}$ and $1/p_i$, i=1,..., L, respectively.

To determine the state probability distribution of the network, the constraints on the means $E(N_i)$ must be given. Modeling each node as a QGeom/QGeom/1 queue, we use $E(N_i)$ in (6.5.3) and obtain the solution as in (6.5.6) and (6.5.5).

To determine the flow process in the network and implement the solution (6.5.5), we shall use (6.5.13) for the coefficient of variation of the interdeparture time. As for the merging and splitting processes , we use the formulas (6.5.16), and (6.5.18).

6.6.1 A Two - Node Network

Consider a network with two nodes in series. Suppose that the service time distributions of node 1 and node 2 are geometric with mean $1/\mu_1$ and $1/\mu_2$, respectively. Using (6.4.3), we specify the interarrival time distribution of packets entering node 1 by a QGeom distribution with mean p/λ_g and coefficient of variation $\sqrt{(2-\lambda_g-p)/p}$. Using (6.5.3), we obtain the mean $E(N_1)$. For node 2, we employ (6.5.13) for the coefficient of variation of the interdeparture time from node 1 and then use (6.5.3) for the mean $E(N_2)$. Analytic results are displayed in

Fig. 6.2. Compared with the simulation results, we see that the error increases for large values of $1/p$ (i.e., large variance of the interarrival time of packets) since the flow processes in the network become more dependent on each other so the decomposition of network will result in large errors.

6.6.2 Clocked Multistage Interconnection Networks with Bulk Arrival [97]

In the following, we shall consider the discrete time case of the network model as described in Section 5.2.

We also make the following assumptions:

1. The network is operated synchronously and the time axis is divided into equal units of length Δt=1. Arrivals and services can take place only at the discrete time, $t_k = k$, k=1, 2,....

2. The interarrival times T_g of external message are mutually independent , identically and geometrically distributed with mean $1/\lambda_g$ seconds, where λ_g is the average arrival rate. Thus the probability of having an arrival in unit time equals λ_g.

3. The message size B is also geometrically distributed with mean 1/p packets/message.

4. The service (or transmission) time of packet at a switch of stage i is constant which is equal to a unit time Δt=1. The queue discipline is FIFO (first-come-first-served).

5. Packets arriving at each input link of stage 1 are destined uniformly (or randomly) for all output links at stage n_b. Thus the traffic load for each input link at stage 1 is identical and the traffic loads in the whole network are balanced. Under these assumptions, the state probability distributions of switches at the same stage are statistically identical.

Observe that a queue of an arbitrary switch at the first stage can be modeled as a $\text{Geom}^X/D/1$ queue. From (6.2.6), the generating function of the state probability distribution is given by

$$G_N(z) = \frac{(1-\rho)[1-\lambda_g - z(1-\lambda_g - p)]}{1-\lambda_g - (1-p)z} \tag{6.6.1}$$

For later stages, we shall use an approximate method similar to the one described in Section 5.2. We first decompose the queueing network into a number of QGeom/D/1 queues. The reasons for using the QGeom input are as follows : the QGeom distribution is completely specified by its mean and coefficient of variation, and the QGeom distribution contains a zero interarrival time component which is a good description for batches entering into the system within a unit time for the queueing network under consideration. Since the input traffic is symmetric, it is possible to obtain simple and explicit expressions in the form of recursive equations for all traffic descriptors in the network. Specifically, (a) the departure process from a queue at stage i can be approximated by a stationary QGeom process with coefficient of variation c_{D_i}, which can be obtained from (6.5.13); (b)

if a stationary process is split into two stationary processes, the coefficient of variation $c_{i,i+1}$ of each component process is available from (6.5.16) ; and (c) if two QGeom processes with the same mean arrival rate $\lambda/2$ and coefficient of variation $c_{i-1,i}$ are superimposed, then the resultant process can be described by (λ, c_{T_i}) where the coefficient of variation c_{T_i} can be obtained from (6.5.18).

Combining (a)-(c), we obtain a formula similar to (5.4)

$$c_{T_i} = \left[[1+\rho_{i-1}-\rho_{i-1}{}^2+(1-\rho_{i-1})c_{T_{i-1}}{}^2] / 2 \right]^{1/2} \quad , i = 2,..., n_b \tag{6.6.2}$$

with the initial condition

$$c_{T_1} = \sqrt{(2-\lambda_g-p)/p} \tag{6.6.3}$$

where

$$\rho_i = \frac{\lambda_g}{p} \quad , \; i = 1,..., n_b$$

Equations (6.6.2) and (6.6.3) define a complete recursive formula for calculating $\{c_{T_i} ; i=1,..., n_b\}$.

Also from (6.3.8) and (6.3.10), the mean $E(N_i)$ and variance $Var(N_i)$ for a link queue at stage i are given by the following expressions for the QGeom/D/1 queue

$$E(N_i) = \frac{\rho_i(1-\rho_i+c_{T_i}^2)}{2(1-\rho_i)} \tag{6.6.4}$$

and

$$Var(N_i) = \frac{\rho_i(1-\rho_i+c_{T_i}^2)[2c_{T_i}^2 - \rho_i(1-\rho_i+c_{T_i}^2)]}{4(1-\rho_i)^2} \tag{6.6.5}$$

Using (6.6.4) and Little's formula, we obtain the mean waiting time

$$E(W_i) = \frac{c_{T_i}^2-1+\rho_i}{2(1-\rho_i)} \tag{6.6.6}$$

Finally, the mean transfer delay from source to destination for a packet, T_d , is given by

$$T_d = n_b + \sum_{i=1}^{n_b} E(W_i) \tag{6.6.7}$$

where n_b is the total number of stages. The first term is the packet transmission time and the second term is the cumulative queueing delay.

Figs. 6.3 and 6.4 give the mean and variance of the number of packets for stage 2 and 3, respectively against traffic intensity and average message size. It is observed that the mean number of packets at each node is small for low to medium traffic intensity but large for high traffic intensity. It is further observed that the means are larger for higher values of message size and variances are greater than

the means. Note the good agreement between the analytic and simulation results especially for low to medium traffic intensity; For $\rho>0.85$, the discrepancy is due to the invalidity of independence of network flow processes, so a more accurate approach is required for compensating this effect. We note that for each λ_g and p, the mean number of packets shrinks at the early stages rather quickly before reaching an asymptotic value. This indicates that $E(N_i)$ in successive stages are nearly independent. Asymptotic values of $c_{T_i}^2$, $c_{T_\infty}^2$, as $n_b \to \infty$ can be obtained from (6.6.2) and given by

$$c_{T_\infty}^2 = \frac{\lambda_g\, p - \lambda_g^2 + p^2}{\lambda_g + p}$$

It may be substituted in (6.6.4) to obtain the asymptotic values $E(N_\infty)$, as follows

$$E(N_\infty) = \frac{\lambda_g \left[1 - \frac{\lambda_g}{p} + \frac{\lambda_g\, p - \lambda_g^2 + p^2}{\lambda_g + p} \right]}{2p\left(1 - \frac{\lambda_g}{p}\right)}$$

In the case of single-packet message arrivals (p=1), Bouras et al [81] provided an upper bound b_{upper} for the mean $E(N_2)$ and beyond, i.e,

$$E(N_j) < b_{upper} = \frac{\lambda_g - \frac{3}{2}\lambda_g^2 + \lambda_g \left(1 - \frac{\lambda_g}{2}\right)^{-2}}{2(1-\lambda_g)} \quad ; i=2,\ldots, n_b$$

From (6.6.4), we find

$$E(N_j) < E(N_2) = \frac{\lambda_g\,(4-3\lambda_g)}{4(1-\lambda_g)} \quad , \; j = 3,\ldots, n_b$$

$$< b_{upper}$$

which agrees with [81].

6.7 DISCRETE-TIME QUEUEING NETWORKS WITH WINDOW CONTROL

In this section, a discrete time model with window control mechanism and bulk arrivals is presented. First, we replace assumptions 1 and 2 of Section 4.3.1 by the following assumptions :

1*. The interarrival times of messages and message lengths are geometrically distributed with mean $1/\lambda_g$ sec/message and 1/p packets/message, respectively;

2*. The service times are also geometrically distributed with mean $1/\mu_i$ sec/packet , i=1,2,..., L, where L is the number of nodes in the network. Each node has only one server.

We take a similar procedure as in Section 4.3.1 and model the message arrival process by an equivalent packet arrival process which has a QGeom interarrival time distribution with mean p/λ_g and coefficient of variation $\sqrt{(2-\lambda_g-p)/p}$.

Recall from Section 6.5 that the state probability distribution of an open network is given by (6.5.5). Now the network solution is similar to (4.3.4) with $g_i(n_i)$ replaced by

$$g_i(n_i) = \begin{cases} 1 \quad , \ if \ n_i=0 \\ \dfrac{2\rho_i b_i}{\rho_i - 1 + c_{Y_i}^2 + \rho_i \ c_{Y_i}^2} \left[\dfrac{\rho_i - 1 + c_{T_i}^2 + \rho_i \ c_{Y_i}^2}{1 - \rho_i + c_{T_i}^2 + \rho_i \ c_{Y_i}^2} \right]^{n_i} , \ if \ n_i \geq 1 \end{cases}$$

In the following example, a network of two links in series is considered. Each link is modeled as a single-server node. The transmission time is geometrically distributed with mean $1/\mu$. The message arrival process is a geometric process with parameter λ_g messages/sec and the message length is geometrically distributed with mean 1/p packets/message. The window size is assumed to be n_W. Let λ_g=0.15, p=0.5, and μ=0.6. From (4.3.8)-(4.3.10) , we can calculate the mean number of packets, the utiliuzation and the throughput of each node. Numerical results are summarized in Table 6.1. Note that there is a discrepancy in $E(N_2)$ due to the dependence of the second node on the first node. We see that the throughputs of both nodes are equal and thus satisfy the Work Rate Theorem [71]. The utilizations of the two nodes are also equal because of identical service time distribution. The utilizations and the throughputs increase proportionally to the window size but tend to be constant. The reason is that as the window size grows,

the network tends to become an open network. Virtually the utilizations is equal to $\rho=\lambda_g/\mu p$ and the throughput is equal to λ_g/p. In our example, these values are 0.5 and 0.3, respectively. Also the mean number of packets at each node increases as the window size grows but tends to be a constant. Based on our approximate analysis from Section 6.5 for open networks , we obtain from (6.5.3)

$$E(N_1) = \frac{\rho\,(1-\lambda_g)}{p\,(1-\rho)}$$

Then applying (6.5.13) for c_{D_1} to (6.5.3) , we obtain

$$E(N_2) = \frac{\rho\left[2\rho-\mu\rho-\mu\rho^2 + (1-\rho)\left(\frac{2-\lambda_g}{p}\right)\right]}{2(1-\rho)}$$

In our case these values are equal to 1.7 and 1.2, respectively.

6.8 SUMMARY

Discrete-time G/G/1 queues and networks have been investigated by means of entropy maximization. Distributions of the state probability and mean waiting time have been obtained. An equivalence relationship between the Geom^X and the QGeom processes is established. Maximum-entropy formalism has been used to derive a product-form solution for synchronous queueing networks at equilibrium. The results show that each node is analyzed as independent discrete-time G/G/1 system if the flow processes in the network have been determined. It reveals that an extension of the Jackson network has been established. Open networks with

bulk arrivals have also been discussed. The results have been applied to a buffered interconnection network and numerical results obtained showed good agreement with simulation results. Queueing networks with window control and bulk arrivals are also studied. Throughput, utilization and mean number of packets at each node are obtained and compared favourably with simulation results.

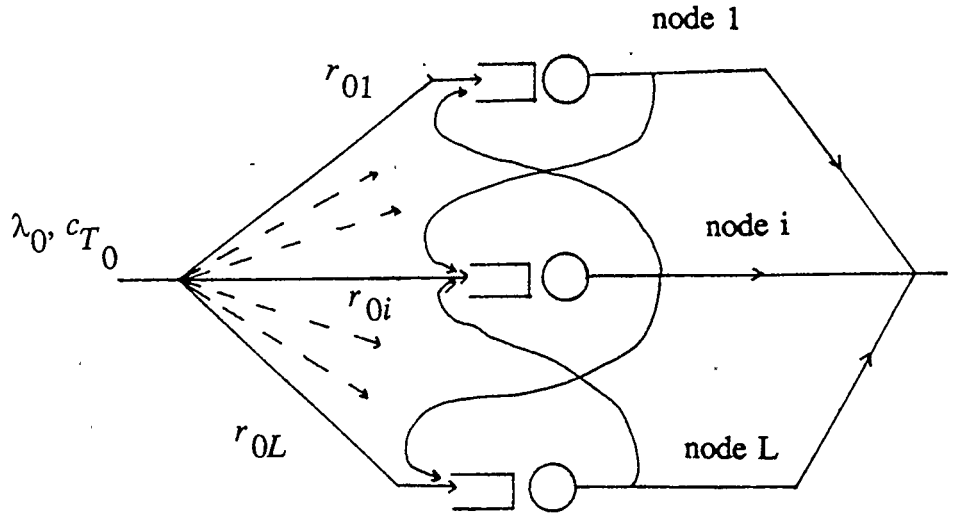


Fig. 6.1 A discrete-time open queueing network

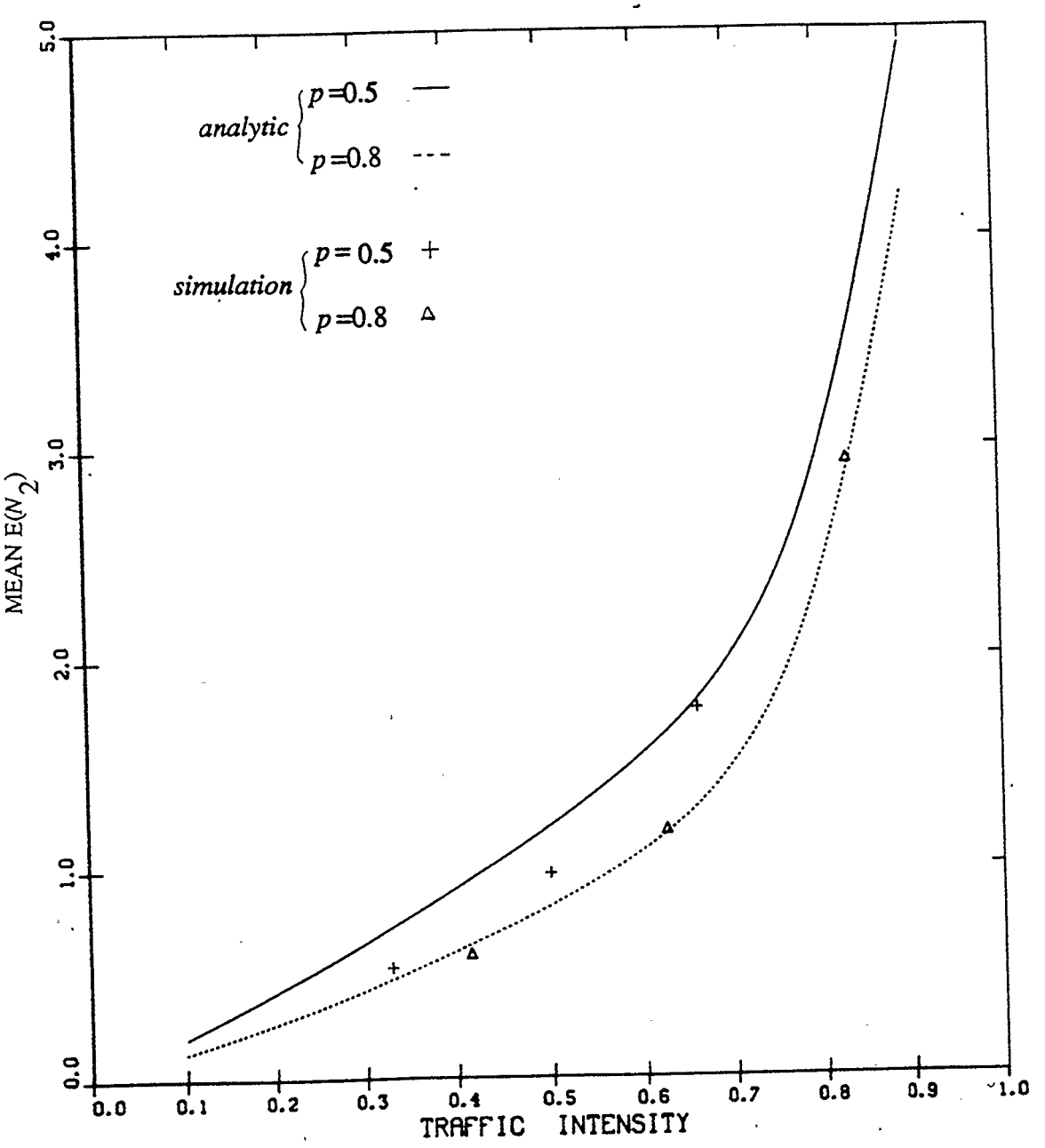


Fig. 6.2 Mean $E(N_2)$ v.s ρ_2

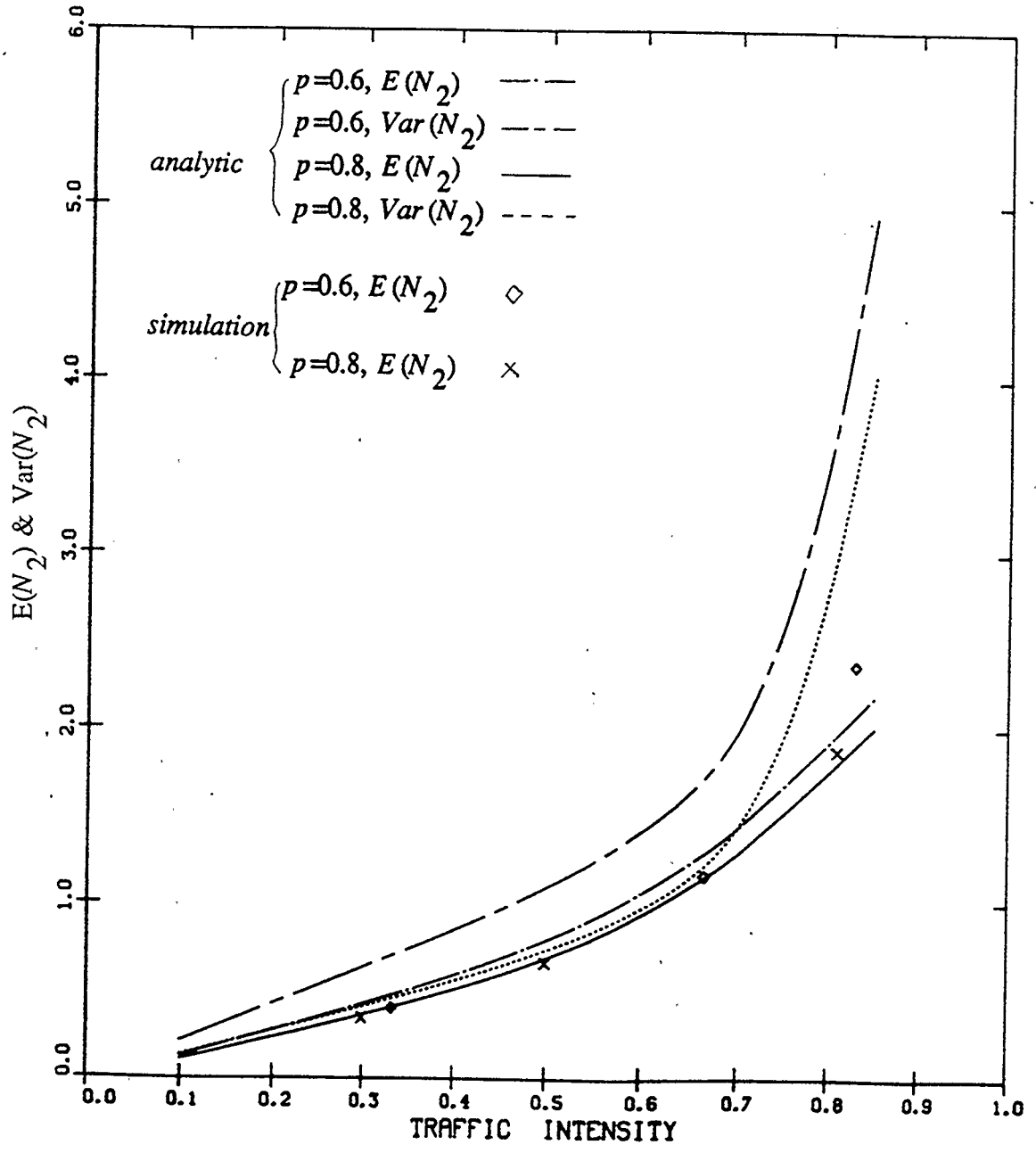


Fig. 6.3 Mean and variance of the number of packets in stage 2

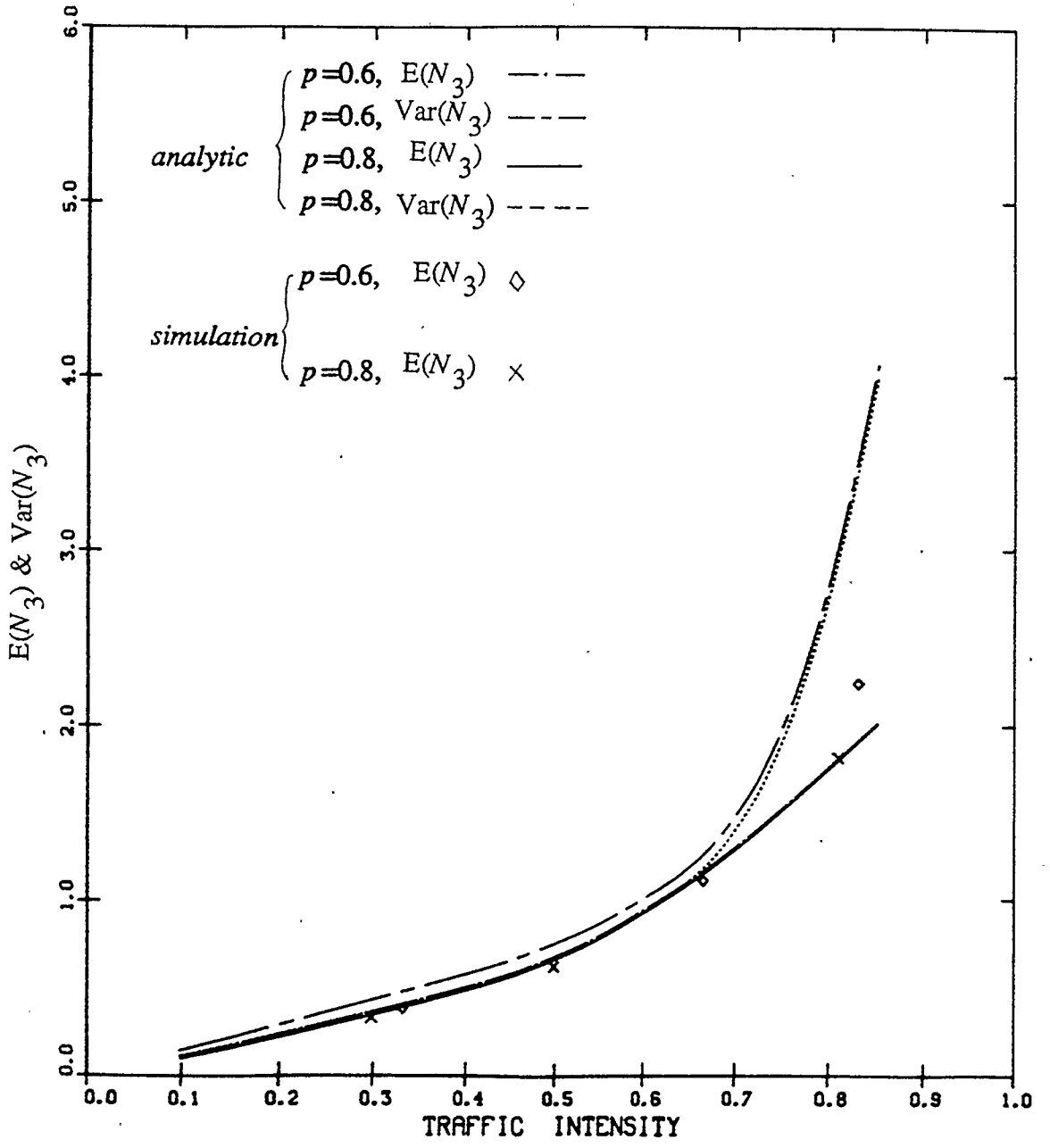


Fig. 6.4 Mean and variance of the number of packets in stage 3

Table 6.1 Performance measures for the network in discrete time

$C_1=C_2=1$ ([ ] - simulation results)

n_W	U_1	U_2	$E(N_1)$	$E(N_2)$	*Throughput*
4	0.484 [0.386 ±0.00066]	0.484 [0.384 ±0.0004]	0.761 [0.693 ±0.001]	0.727 [0.517 ±0.0008]	0.289
6	0.499 [0.439 ±0.0006]	0.499 [0.437 ±0.0006]	1.013 [0.994 ±0.0022]	0.908 [0.676 ±0.0015]	0.29
8	0.498 [0.467 ±0.0005]	0.498 [0.469 ±0.006]	1.209 [1.22 ±0.002]	1.023 [0.78 ±0.002]	0.291
10	0.488 [0.483 ±0.0004]	0.489 [0.481 ±0.0005]	1.421 [1.384 ±0.003]	1.114 [0.85 ±0.002]	0.293
12	0.497 [0.487 ±0.0001]	0.498 [0.489 ±0.0008]	1.505 [1.492 ±0.005]	1.156 [0.903 ±0.002]	0.298
15	0.5 [0.495 ±0.0008]	0.5 [0.497 ±0.011]	1.591 [1.61 ±0.005]	1.19 [0.947 ±0.005]	0.299
20	0.501 [0.5005 ±0.0008]	0.502 [0.4982 ±0.0008]	1.691 [1.667 ±0.007]	1.2 [0.965 ±0.003]	0.3

CHAPTER 7

CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

7.1 CONCLUSIONS

In this thesis a framework for the approximate analysis of single queueing systems with one or many servers based on the principle of maximum entropy has been presented. Then based on network decomposition, we developed a technique for the analysis of networks of queues with bulk arrivals and FIFO queue discipline in continuous time and discrete time.

For simple queueing systems, we have developed an equivalent arrival process for the bulk-Poisson input and applied it to study the $M^X/G/1$ queue, $M^X/M/1/m$ queue, $M^X/G/1$ queue with vacation, and the $M^X/M/C$ queue. It is shown that the equilibrium state probability distribution obtained by using the equivalent arrival process agrees with those of the GE/•/• queue.

Using the method of entropy maximization, two approximations for the G/G/C queueing systems have been studied. Expressions for the state probability distribution and the waiting time distribution are deduced. For the M/G/C queueing system, the accuracy of the second formulation (ME2) is better than that of the first one (ME1), especially as the number of servers and the coefficient of variation of service time or interarrival time are large. Moreover, the method of diffusion approximation has been modified by incorporating the law of conservation of traffic

and the asymptotic property from queueing theory. The accuracy of performance measures, e.g. mean number of customers, is also shown to be improved.

In extension to queueing networks, we have presented a method of solution for various types of network with bulk arrivals such as the open networks with single-server and exponential multiserver nodes, open networks with finite-capacity nodes, open networks in discrete time. This method is based on a network decomposition into single node subnetworks and the decomposition yields approximations of the performance measures for the network under consideration. Examples are given in Chapter 5 and the results are compared favorably with simulation. The proposed method thus provides a useful means for the study of packet switching networks with bulk-Poisson input. Furthermore, the problems of networks with multiple nonexponential servers have been tackled and the results indicated that the accuracy of the first entropy maximization formulation (ME1) is comparable to the method of diffusion approximation.

Using the method of entropy maximization, an approximation to networks of queues with bulk-Poisson input and window flow control mechanism has been proposed. In this approach an approximation for the state probability distribution of the network is obtained by entropy maximization subject to constraints on the mean numbers of packets and utilizations. These constraints are derived from the first few moments of the service time distributions. Maximum entropy approximations are obtained by iteration and produce results which are close with simulation results. The approximate method of entropy maximization therefore appears to be a

useful contribution to the analysis of queueing networks with window flow control and bulk-Poisson input.

The entropy maximization approach has also been used to study discrete-time queueing systems. An equivalent arrival process for the $Geom^X$ input has been developed and the framework for open networks similar to the Jackson network has been established. In the application to buffered interconnection networks, the mean and variance of the number of packets in each stage have been determined. Numerical results are presented in curve form for illustration. The analytic results agree with simulation results closely.

7.2 SUGGESTIONS FOR FURTHER RESEARCH

In this section we shall outline some suggestions for further studies. For the G/G/C queue, investigation of queueing systems with heterogeneous servers requires further study to find the state probability distribution and the waiting time distribution. The G/G/C queue in discrete time has not yet been explored. The approach of maximum entropy may be extended to queueing systems with priority classes of customers. In the area of queueing networks, further studies may include the extension of multiple routes with window flow-controlled mechanism and different acknowledgement strategies, networks with priority classes of customers, and networks with G/G/C queues and bulk arrivals in discrete time.

REFERENCES

[1] E. Fuchs and P.E. Jackson, "Estimates of Distributions of Random Variables for Certain Computer Communication Traffic Models," *Communications of the ACM,* Vol. 13, pp. 752-757, 1970.

[2] 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.

[3] D.R. Cheriton and C.L. Williamson, "Network Measurement of the VMTP Request-Response Protocol in the V Distributed System," *Proc. of ACM SIGMETRICS Conf. on Measurement and Modeling of Computer System,* pp. 216-225, 1987.

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

[5] P.J. Denning and F.P. Buzen, "The Operational Analysis of Queueing Network Models," *Comput. Surv.,* Vol. 10, pp. 225-261, 1978.

[6] J.E. Shore and R. Johnson, "The Principle of Maximum Entropy and the Principle of Minimum Cross Entropy," *IEEE Tran. Inf. Theory,* Vol. IT-26, pp. 26-37, 1980.

[7] A. E. Ferdinand, "A Statistics Mechanical Approach to System Analysis," *IBM. J. Res. Dev.,* Vol. 14, pp. 539-547, 1970.

[8] J. E. Shore, "Derivation of Equilibrium and Time-Dependent Solutions to M/M/∞/N and M/M/∞ Queueing Systems Using Entropy Maximization,"

AFIPS. Conf. Proc., Vol. 47, pp. 473-478, 1978.

[9] J. E. Shore, "Information Theoretic Approximations for M/G/1 and G/G/1 Queueing Systems," *Acta Inf.,* Vol. 17, pp. 43-61, 1982.

[10] D. D. Kouvatsos, " A Maximum Entropy Analysis of the M/G/1 and G/M/1 Queueing Systems at Equilibrium," *Acta Inf.,* Vol. 19, pp. 339-355, 1983.

[11] D. D. Kouvatsos, "Maximum Entropy and the G/G/1/N Queue," *Acta Inf.,* Vol. 23, pp. 545-656, 1986.

[12] Y. Bard, "Estimation of State Probabilities Using the Maximum Entropy Principle," *IBM J. Res. Dev.,* Vol.24, pp. 563-569, 1980.

[13] Y. Bard, "Modeling I/O Systems with Dynamic Path Selection and General Transmission Networks," *Perf. Eval. Rev.,* Vol. 11, no. 4, pp. 118-129, 1982.

[14] C. Bobrowski, *The Principle of Maximum Entropy in Some Computer System Modelling Problems,* M.Sc thesis, Univ. of Toronto, 1983.

[15] D. D. Kouvatsos, *Maximum Entropy Methods for General Queueing Networks,* Research Report RCC34, Univ. of Bradford, Bradford, U.K., 1983.

[16] C. E. Shannon, "A Mathematical Theory of Communication," *Bell Syst. Tech. J.,* Vol. 27, pp. 379-423, 623-656, 1948.

[17] V.E. Benes, *Mathematical Theory of Connecting Networks and Telephone Traffic,* Academic Press, New York, 1965.

[18] M. Chan, "System simulation and Maximum Entropy," *Operations Research,* Vol. 19, pp. 1751-1753, 1971.

[19] I.J. Good, "Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contigency Tables," *Annals Math. Stat.,* Vol. 34, pp. 911-934, 1963.

[20] R. L. Kashyap, "Prior Probability and Uncertainty," *IEEE Trans. Infor. Theory,* Vol. IT-17, pp. 641-650, 1971.

[21] M. Tribus, "The Use of Maximum Entropy Estimate in the Estimation of Reliability," *Recent Developement in Information and Decision Processes,* (R.E. Machol and D. Gray Eds.), Macmillan, New York, pp. 102-139, 1962.

[22] A. Katz, *Principles of Statistical Mechanics - The Information Theory Approach,* W.H. Freeman Company, New York, 1967.

[23] M. Tribus and G. Fitts, "The Widget Problem Revised," *IEEE Trans. on Systems Science and Cybernetics, Vol. SSC-4, pp. 241-248, 1968.*

[24] S. Kullback, *Information Theory and Statistics,* Dover, New York, 1968.

[25] J.G. Ables, "MAximum Entropy Spectral Analysis," *Astron. Astrophys. Suppl.,* Vol. 15, pp. 383-393, 1974.

[26] S.J. Wernecke and L.D. D'Addario, "Maximum Entropy Image Reconstruction," *IEEE Trans. on Computers,* Vol. C-26, pp. 351-364, 1977.

[27] M. Tribus, *Rational Descriptions: Decisions and Designs,* Pergaman Press, 1969.

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

[29] D. D. Yao, "Analyzing the Steady-State GI^X/G/1 Queue," *J. Opl. Res. Soc.*, Vol. 35, no. 11, pp. 1027-1030, 1984.

[30] J. D. C. Little, "A Proof for the Queueing Formula : L = λ W," Oper. Res., Vol. 9, no. 3, pp. 381-387, 1961.

[31] R. Weinstock, *Calculus of Variation with Applications to Physics and Engineering,* Dover, New York, 1974.

[32] L. Takacs, *Introduction to the Theory of Queues,* Oxford Univ. Press, New York, 1962.

[33] M. Schwartz, *Computer Communication Network Design and Analysis,* Prentice-Hall, 1977.

[34] D. Gross and C. M. Harris, *Fundamental of Queueing Theory,* John Wiley, 1974.

[35] Y. Baba, "On the M^X/G/1 Queue with Vacation Time," *Op. Res. Lett.,* Vol. 5, no. 2, pp. 93-98, 1986.

[36] J. Keilson and L. D. Servi, "Oscillating Random Walk Methods for G/G/1 Vacation Systems with Bernoulli Schedules," *J. Appl. Prob.,* Vol. 23, pp. 790-802, 1986.

[37] P. J. Kuehn, "Multiqueue Systems with Nonexhaustive Cyclic Service," *Bell Syst. Tech. J.,* Vol. 58, no. 4, pp. 671-698, 1979.

[38] M. L. Chaudhry and J. G. C. Templton, *A First Course in Bulk Queues,* Wiley, New York, 1983.

[39] Y. Takahashi and Y. Takami, "A Numerical Method for the Steady-State Probabilities of a GI/G/c Queueing in a General Class," *J. Opl. Res. Soc. of Japan,* Vol. 19, pp. 147-157, 1976.

[40] R.W. Johnson, "Determining Probability Distribution by Maximum Entropy and Minimum Cross-Entropy," *Proc. APL 79,* pp. 24-29, 1979.

[41] 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.

[42] K. Ogata, *Modern Control Engineering,* Prentice Hall, 1970.

[43] D.P. Gaver, nd G.S. Shedler,"Approximate Models for Processor Utilization in Multiprogrammed Computer Systems," *SIAM J. Comput.* Vol. 2, no.3, pp. 183-192, 1973.

[44] C. Palm, "Research on Telephone Traffic Carried by Full Avilability Groups," *Tele.(English ed.),* no. 1, pp. 1-107, 1957.

[45] H.C. Tijms et al, "Approximations for the Steady-State Probabilities in the M/G/c Queue," *Adv. Appl. Prob.,* Vol. 13, pp. 186-206, 1981.

[46] D.D. Yao, "Refining the Diffusion Approximation for the M/G/m Queue," *Oper. Res.,* Vol. 33, no. 6, pp. 1267-1277, 1985.

[47] J. Kollerstrom, "Heavy Traffic Theory for Queues with Several Servers," *J. Appl. Prob.,* Vol. 11, pp. 544-552, 1974.

[48] D.D. Kouvatsos, "On the Analysis of the G/G/1 Queue," *Research Report CCR18,* Univ. of Bradford, U.K., 1984.

[49] F.S. Hillier and O.S. Yu, *Queueing Tables and Graphs,* North Holland, New York.

[50] L.P. Seelen et al, *Tables for Multi-Server Queues,* North Holland, 1985.

[51] J.F.C. Kingman, "The Heavy Traffic Approximations in the Theory of Queues," In *Proc. of the Symposium on Congestion Theory,* pp. 137-159, 1965.

[52] D.L. Iglehart and W. Whitt, "Multiple Channel Queue in Heavy Traffic," *Adv. Appl. Prob.,* Vol. 2, pp. 150-177, 1970.

[53] B. Halachmi and W.R. Franta, "A Diffusion Approximation to the Multi-Server Queue," *Management Sci.,* Vol. 24, pp. 522-529, 1978.

[54] T. Sunaga et al, "An Approximation Method Using Continuous Models for Queueing Problems," *J. Opns. Res. Soc. Japan,* Vol. 21, pp. 29-44, 1978.

[55] H. Takahashi and H. Akimaru, "Analysis of Queueing System via Multi-Dimensional Elementary Return Process," *ITC-11,* pp. 432-438, 1985.

[56] G. Kimura and Y. Takahashi, "Diffusion Approximation for a Token Ring System with Nonexhaustive Service," *IEEE Sel. Area in Comm.,* Vol. SAC-4, no. 6, pp. 794-801, 1986.

[57] E. Gelenbe and I. Mitrani, *Analysis and Synthesis of Computer Systems,* Academic Press, 1980.

[58] T. Kimura, "Diffusion Approximation for an M/G/m Queue," *Oper. Res.*, Vol. 31, no. 2, pp. 304-321, 1983.

[59] W. Feller, "Diffusion Process in One Dimension ," *Trans. Am. Math. Soc.*, Vol. 77, pp. 1-31, 1954.

[60] S. Lam, "A New Measure for Characterizing Data Traffic," *IEEE Tran. Com.* , Vol. Com-26, 1978.

[61] J.L. Hammond and P.J.P. O'Reilly, *Performance Analysis of Local Computer Network,* Addision Wiley, 1986.

[62] P.J. Burke, "Delays in Single-Server Queues with Batch Input," *Bell Syst. Tech. J.*, Vol. 54, pp. 830-833, 1975.

[63] D.R. Manfield and P. Tran-Gia, "Analysis of a Finite Storage System with Batch Input Arising out of Message Packetization," *IEEE Tran. on Com.*, Vol. Com-30, pp.456-463, 1982.

[64] J.R. Jackson, "Jobshop-Like Queueing Systems," *Management Sci.*, Vol. 10, pp. 645-653, 1963.

[65] K.C. Sevick et al, "Improving Approximations of Aggregate Queueing Network Subsystems," In *Computer Performance* (K.M. Chandy ed.), North Holland, pp. 1-22, 1977.

[66] H. Heffes, "On the Output of a GI/M/N Queueing System with Interrupted Poisson Input," *Oper. Res.*, Vol. 24, no. 3, pp. 530-542, 1976.

[67] P.P. Bocharov and S.S. Spesivov, "Analysis of Coefficient of Variation of Intervals between Departures in PH/PH/1/r System," in *Methods of Teletraffic Theory in Decentralized Control Systems,* Moskow, pp. 23-32, 1986.

[68] M. Reiser, "A Queueing Network Analysis of Computer Communication Networks with Window Flow Control," *IEEE Tran. Comm. ,vol. Com-27,* pp. 1199-1209, 1979.

[69] M. Schwartz, ''Performance Analysis of the SNA Virtual Route Pacing Control,'' *IEEE Tran. Comm.,* Vol. Com-30, pp. 172-184, 1982.

[70] M. Reiser and H. Kobayashi, "Queueing Networks with Multiple Closed Chains: Theory and Computational Algorithms," *IBM J. Res. Dev.* , Vol. 19, pp. 283-293, 1975.

[71] A. Chang and S.S. Lavenberg, "Work Rates in Closed Queueing Networks with General Independent Servers," *Oper. Res.,* Vol. 22, pp. 838-847, 1971.

[72] H. Kobayashi, *Modeling and Analysis: An Introduction to System Performance Evaluation Methodology,* IBM Corp., 1978.

[73] D. Gajski et al, "Construction of a Large Scale Multiprocessor", *1983 ICPP,* pp. 524-529.

[74] A. Norton and G. Pfister, "A Methodology for Predicting Multiprocessor Performance," *1985 ICPP,* pp. 772-781.

[75] G. Pfister et al, "The IBM Research Parallel Processor Protype (RP3): Introduction and Architecture," *1985 ICPP,* pp. 764-771.

[76] G.R. Goke and G.J. Lipovski, "Banyan Networks for Partitioning Multiprcessor Systems," *1st Ann. Symp. on Computer Architecture,* pp. 21-28, 1973.

[77] D.H. Lawrie, "Access and Alignment of Data in an Array Processor," *IEEE Tran. on Computers,* Vol. C-24, pp. 1145-1155, 1975.

[78] J.A. Patel, "Performance of Processor-Memory Interconnections for Multiprocessors," *IEEE Trans. on Computers,* Vol. C-30, pp. 771-780, 1981.

[79] C.P. Kruskal, M. Snir and A. Weiss, "On the Distribution of Delays in Buffered Multistage Intercinnection Networks for Uniform and Nonuniform Traffic," *Proc. 1984 Intl. Conf. on Parallel Processing,* pp. 215-219.

[80] Y.C. Jenq, "Performance Analysis of a Packet Switch Based on Single-Buffered Banyan Network," *IEEE Tran. on Selected Areas in Comm., Vol. SAC-1, , pp. 1014-1021, 1983.*

[81] C. Bouras, J. Garofalakis, P. Spirakis, and V. Triantafillou, " Queueing Delays in Buffered Multistage Interconnection Networks," *Proc. of the 1987 ACM Sigmetrics Conf. on Measurement and Modeling of Computer Systems,* pp. 111-121.

[82] M. H. Ackroyd, "Computing the Waiting Time Distribution for the G/G/1 Queue by Signal Processing Methods," *IEEE Trans. on Commu.* Vol. Com-28, pp. 52-58, 1980.

[83] K. Bharath-Kumar, "Discrete-Time Queueing Systems and Their Networks," IEEE Trans. on Commu., Vol. Com-28, pp. 260-263, 1980..

[84] H. Kobayashi and A. G. Konheim, "Queueing Models for Computer Communications System Analysis," *IEEE Trans. on Commu.,* Vol. Com-25, pp. 2-29, 1977.

[85] G. Pujolle, J. P. Claude and D. Seret, "A Discrete Queueing System with a Product Form Solution," *Proc. Int. Seminar on Comp. Network and Perf. Evaluation,* pp. 139-147, 1985.

[86] J. A. Morrison, "Two Discrete Time Queues in Tandem,' *IEEE Trans. on Commu.,* Vol. Com-27, pp. 563-573, 1979.

[87] P. Tran-Gia, "Discrete-Time Analysis for the Interdeparture Distribution of G/G/1 Queue," in *Teletraffic Analysis and Computer Performance Evaluation,* O. J. Boxma and H. C. Tijms (eds), pp. 341- 357, 1986.

[88] A. G. Konheim, "An Elementary Solution of the Queueing System G/G/1," *SIAM J. Comput.,* Vol. 4, no. 4, pp. 540-545, 1975.

[89] K. T. Marshall, "Some Relationships between the Distributions of Waiting Time, Idle time, and Interoutput Time in the GI/G/1 Queue," *SIAM J. Appl. Math.,* Vol. 16, pp.324-327, 1968.

[90] D. R. Cox and W. L. Smith, *Queues*, Methuen & Co. Ltd., 1961.

[91] J.R. Spirn et al, "Bursty Traffic Local Network Modeling," *IEEE J. on Sel. Areas in Comm.,* Vol. SAC-2, pp. 250-257, 1984.

[92] M.J. Sobel, "Simple Inequalities for Multiserver Queues," *Management Science,* Vol. 26, pp. 951-954, 1980.

[93] J.S. Wu, W.C. Chan and D. Irvine-Halliday, "Information Theoretic Approximations for Multiple Server Queueing Systems," *Proc. of IASTED Applied Simulation and Modeling,* pp. 88-93, 1988.

[94] J.S. Wu, W.C. Chan and D. Irvine-Halliday, "Maximum ENtropy Analysis of Queueing Networks with Window-Based Flow Control," *ORSA/TIMS National Joint Meeting,* Denver, Oct. 24-26, 1988.

[95] J.S. Wu and W.C. Chan, "Performance Analysis of Queueing Networks with End-to-End Window Flow Control," *Accepted for Publication in IEE Proc., Part E.*

[96] J.S. Wu and W.C. Chan, "Performance Evaluation of Buffered Banyan Networks," *ISMM Int. Conf. on Mini and Microcomputers,* Miami, Dec. 14-16, 1988.

[97] J.S. Wu and W.C. Chan, "Approximate Analysis of Clocked Interconnection Networks with Bulk Arrivals," *ISMM Int. Conf. on Mini and Microcomputers,* Miami, Dec. 14-16, 1988.

[98] R.B. Cooper, *Introduction to Queueing Theory,* The Macmillan Company, 1981.

[99] Bruno De Finetti, *Theory of Probability,* Vol. 1, Wiley, 1974.