1

MADE IN JAPAN WITH
Micro-D
MICRO-DRUM
MICRO-DRUM

National Library
of Canada

Bibliothèque nationale
du Canada

Canadian Theses Service

Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, tests publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30.

TOPICS IN STATE DEPENDENT QUEUES AND QUEUEING NETWORKS

by

Rengarajan Srinivasan B.Sc., M.Sc.

A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Mathematics and Statistics

Carleton University

Ottawa, Canada

July, 1988

© Copyright, 1988
Rengarajan Srinivasan

Permission has been granted to the National Library of Canada to microfilm this thesis and to lend or sell copies of the film.

The author (copyright owner) has reserved other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without his/her written permission.

L'autorisation a été accordée à la Bibliothèque nationale du Canada de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

L'auteur (titulaire du droit d'auteur) se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation écrite.

ISBN 0-315-46249-3

The undersigned recommend to the Faculty of Graduate Studies and Research acceptance of this thesis

Topics in State Dependent Queues and Queueing Networks

submitted by Rengarajan Srinivasan, B.Sc., M.Sc., in partial fulfillment of the requirements for the degree of Doctor of Philosophy

D. Dawson

Dr. D.A. Dawson, *Thesis Supervisor*

R. Herz-Fischler

Dr. R. Herz-Fischler, *Thesis Supervisor*

P.W. Glynn

Dr. P.W. Glynn, *External Examiner*

Ward Whitt

Dr. W. Whitt, *External Examiner*

C. Garner

Dr. C. Garner, *Chairman*

Department of Mathematics and Statistics

Abstract

We analyse two state dependent single server queues, namely the $D^s/M_A/1$ queue and the $M/G_n/1$ queue. Stationary distributions for the number in the system are derived via the imbedded Markov chain and the regenerative and the up and down crossing techniques for the two queues. These stationary distributions are then used to calculate some key performance measures of semi-open networks by reducing the semi-open network to a state dependent single server queue. Approximate procedures for specifying the service time of the reduced queue are established.

Analogous to the "hydrodynamical study" of the non-equilibrium flow of fluids, density profile and hydrodynamic limit are developed for an infinite series of exponential queues with Poisson and non-Poisson input. This is done by studying an infinite particle system, namely, the *Road Hog* process. The results obtained enhance the understanding of how this queueing network reaches equilibrium.

ACKNOWLEDGEMENTS

I would like to thank my supervisors Dr. R. Herz-Fischler and Dr. D.A. Dawson for their constant encouragement, moral support and invaluable guidance all through my graduate years. Especially, I wish to express my sincere appreciation to them for volunteering to supervise this thesis under 'extraordinary' circumstances.

I am grateful to the authorities of this university, Dr. S.E. Mills and Dr. P. March for financial support. Special thanks to P. March for suggesting the name Road Hog process in Chapter 4 and to Sarah, Diane and Rick for helping me with the TEX.

Also, I thank my parents for their sacrifice and love and the faith they have always shown in me. Last but not the least, I would like to thank my girlfriend Karen for her love, patience and support.

TABLE OF CONTENTS

Abstract	(iii)
Acknowledgements	(iv)
Table of Contents	(v)
Introduction	(vii)
CHAPTER 1: The $D^*/M_n/1$ Queue	1
1.1 Preliminaries.....	2
1.2 Analysis.....	3
1.3 Delay Measures.....	7
1.4 Numerical Results.....	8
CHAPTER 2: The $M/G_n/1$ Queue	15
2.1 Preliminaries.....	16
2.2 Analysis.....	17
2.3 Discussion.....	23
CHAPTER 3: Approximation to Semi-Open Queueing Networks	26
3.1 Open and Closed Queueing Network Models.....	27
3.2 Semi-Open Queueing Network Model.....	30
3.3 Equivalent Single Server Queue Approximation.....	32
3.4 $M/M_n/1$ Approximation.....	35
3.5 Approximations to Semi-Open Queueing Network with Deterministic Input.....	37
3.5.1 $D/M_n/1$ Approximation.....	37
3.5.2 $D/M/1$ Approximation.....	38
3.5.3 $D/G/1$ Approximation.....	38
3.5.4 Numerical Results.....	40
3.6 Approximations to Semi-Open Queueing Networks with Poisson Input.....	46
3.6.1 Approximate Description of Exit times from Two Node Serial Semi-Open Networks.....	48
3.6.2 Approximate Description of Exit Times from Two Node Feedback Semi-Open Network.....	50

3.6.3 Approximate Description of Exit Times from Central Server Semi-Open Network	52
3.6.4 Simulation Experiments	53
CHAPTER 4: Hydrodynamic Limit for Infinite Series of Exponential Queues	62
4.0 Introduction	62
4.1 Asymmetric Simple Exclusion Process as a Model of Series of Exponential Queues	65
4.1.1 Asymmetric Simple Exclusion Process	65
4.1.2 The Model	67
4.2 Hydrodynamic Limit for the Infinite Series of Saturated Exponential Queues	69
4.2.1 Preliminaries	70
4.2.2 Density Profile and Hydrodynamic Limit	72
4.3 Hydrodynamic Limit for the Infinite Series of Unsaturated Exponential Queues with Poisson Input	74
4.3.1 Embedding Method	74
4.3.2 Density Profile and Hydrodynamic Limit	77
4.4 Asymmetric Zero Range Process as a Model of Infinite Series of Unsaturated Exponential Queues	81
4.4.1 Asymmetric Zero Range Process	81
4.4.2 The Model	82
4.4.3 Density Profile and Hydrodynamic Limit	83
4.4.4 Comparison of the Hydrodynamic Density Profiles	85
4.5 Density Profile for Infinite Series of Exponential Queues with Non-Poisson Input	85
4.5.1 Motivation Behind the Stochastic Comparison Method	87
4.5.2 Coupling for Multi-coloured Particle Systems	88
4.5.3 Definitions and Notations	89
4.5.4 Stochastic Comparisons	91
4.6 Density Profile and Hydrodynamic Limit for a Feedback Exponential Queues in Series	98
4.6.1 The Feedback Queues in Series	99
4.6.2 Density Profile and Hydrodynamic Limit	102
REFERENCES	105

INTRODUCTION

Among the important queueing systems in practical applications are those in which the service rate can be varied according to the number of jobs in the system. For example, when a queueing system is a machining center, the service rate of the machine can be increased by assigning more manpower, increasing the capacity of the tools etc., to speed the processing times of the jobs. Due to economies of scale, it may not be possible to operate the machine at the maximum capacity (the maximum service rate). But, it might be feasible to change the service rates according to the number of jobs in the system, i.e., by switching the service rates to a higher level when the number of jobs to be processed is high and to a lower level when the number of jobs is low.

In the first two chapters of this thesis, we consider such a queueing system, where the server works at different rates according to the number of jobs in the system. Chapter 1 deals with a single server state dependent queue with deterministic bulk arrivals and exponential service times, i.e., a $D^b/M_n/1$ queue. Here, the jobs are assumed to arrive at deterministic epochs in bulk quantities. Each arriving job is subjected to some inspection procedure before it is allowed into the queueing system. We assume that with probability p , a job is accepted and that with probability $(1 - p)$ it is rejected. Therefore, the number of jobs accepted from a batch is random. The service rates μ_n , when there are $n = 1, 2, \dots, N-1$ jobs in the system, are assumed to be non-decreasing, and we also assume that μ_N is the maximum service rate possible. We use the imbedded Markov chain technique to derive the stationary distribution for the number of jobs in the system and from this, performance measures such as average number in the system, average waiting time, etc. can be easily calculated.

In Chapter 2, we consider a state dependent single server queue with Poisson arrivals and general service times, i.e. an $M/G_n/1$ queue. Once again, we assume that the service rate varies according to the number of jobs in the system. Let μ_n , $n = 1, 2, \dots, N-1$ be the service rate when there are n jobs in the system, and μ_N be the maximum service rate. We use the powerful regenerative and up and down crossing techniques to derive the stationary distributions for the number of jobs in the system. The results of this chapter present the stationary distributions in the form of recursive relationships which are very suitable to numerical evaluation.

Queueing network models developed in recent years for computer, communication and manufacturing networks can be divided into two categories namely, open and closed models. In open queueing network models, jobs or customers enter the network from an outside source and receive service at one or more nodes of this network and then leave the network. Thus with this model, the external arrival rate (equivalently the throughput) is an independent variable and the number of jobs in the network is a dependent variable (Whitt(1984)). In closed queueing network models, on the other hand jobs neither arrive nor leave the network; there are a fixed number of jobs which circulate among the nodes indefinitely. Therefore, in closed models, the number of jobs in the system is an independent variable and the throughput is a dependent variable. In contrast to these two models, we consider another queueing network model, namely a semi-open queueing network model (the terminology is new here). In this network, jobs arrive from an external source and are admitted to the network according to the total number of jobs in the network. This would be the case for example, if the network can accommodate only a fixed number of jobs at any time. In this case, an arriving job which finds the network full in the above sense waits in an outside queue and enters the network whenever possible. After entering

the network, a job visits one or more nodes and leaves the network. With this arrival mechanism, this semi-open model fluctuates between being an open and a closed model and hence neither the throughput nor the number of jobs in the system (network + outside queue) is an independent variable.

In most computer, communications and manufacturing networks, the total number of jobs in the system is restricted, usually by resource constraints such as multi-programming level in computer systems, the number of tokens available in communication networks and the number of pallets and fixtures (which are used to mount the jobs in proper orientation) available in the manufacturing networks. Since open models assume that arriving jobs enter the network immediately, and also that there is no limit on the number of jobs which can reside in the network simultaneously, these models do not explicitly model the resource constraint mentioned above. In closed models, since a fixed number of jobs circulate in the network, we can by identifying an exit-entry node, think of an infinite supply of jobs waiting outside the network. As soon as one job enters this exit-entry node, another job from outside source replaces this job, thereby maintaining a fixed number of jobs in the network. Thus closed models definitely capture the resource constraint in terms of the number of jobs allowed in the network, but assume infinite supply and instantaneous replacement of jobs. In most applications, either jobs arrive at random (in computer and communications networks) or jobs are scheduled to arrive (in manufacturing networks). Even though closed models capture the resource constraints, they do not model the arrival pattern of jobs to the network. Semi-open models on the other hand, model the resource constraint as well as the arrival pattern and so are the most appropriate in practical situations.

The analysis of open and closed queueing models, depends basically on various as-

assumptions such as the nature of the arrival process, service processes at various nodes, queueing discipline, amount of waiting space available at each node and the routing of the jobs from one node to another. For example, the steady state solution of open models under Poisson arrivals, exponential service times, infinite waiting space and Markovian routing, are known to have product form solutions. Similarly, closed models with the above assumptions (with of course, no arrivals) are also known to have product form solutions; see Jackson (1963). Variations of these two models depending on the degree to which the above mentioned assumptions are relaxed, also result in product form solutions. However, semi-open queueing network models (even with simpler assumptions such as Poisson arrivals, exponential service time etc.) do not have product form solutions and therefore, in order to obtain some estimates on the performance measures, approximate analyses are necessary. We caution the reader that the semi-open model we study in this thesis is quite different from the ones discussed in Dubois (1983) and Kapur and Kumar (1984) where they assume that jobs arriving when the network is full are lost. Under this special assumption, the semi-open network with Poisson arrival, exponential service times and Markovian routing has a product form solution (See Jackson (1963)).

Chapter 3 investigates the effectiveness of an approximation procedure known as the 'equivalent single server queue approximation' for two semi-open queueing network models. The first model assumes deterministic input, exponential service times and Markovian routing and the second model assumes Poisson input. The first model is useful in manufacturing networks where jobs are usually scheduled to arrive at deterministic epochs. The second model is more useful in computer and communication networks where jobs do arrive at random epochs.

After a brief introduction to the open and closed models (Section 3.1); we present in

Sections 3.2 to 3.4, the semi-open model and the equivalent single server queue approximation. In Section 3.5, we introduce a $D/M_n/1$ queue approximation to the semi-open model with exponential service times and deterministic input. In that section, we also propose various other approximations and compare the performance of the various approximations against simulation results. We use the stationary distribution obtained in Chapter 1 for the $D/M_n/1$ queue to calculate some key performance measures.

In Section 3.6, we consider an $M/G_n/1$ approximation to the semi-open model with Poisson arrivals. Here we use the stationary distribution obtained in Chapter 2 for the $M/G_n/1$ queue to calculate the key performance measures. The performance of the $M/G_n/1$ approximation is compared with that of the frequently used $M/M_n/1$ approximation for this semi-open model. In order to use the $M/G_n/1$ approximation, one needs to obtain the general service time of the equivalent server. In that section, we also provide various approximate procedures to determine the service time of the equivalent server.

As we mentioned earlier, the steady state solution of an open model with Poisson arrivals and exponential service times has a product form solution. One of the simplest open models, is a serial network or queues in series. The product form solution of this simple network is due to the fact that the stationary output process from each queue is Poisson and hence in the steady state each queue in this network behaves like an $M/M/1$ queue. If we relax the assumption of Poisson arrivals by renewal arrivals to this simple network, we no longer have a product form solution. This is easily understood by the fact that the steady state output process of a $GI/M/1$ queue is not even a renewal process (Daley(1976)).

During our simulations experiments with deterministic input, we found that the expected number of jobs in downstream of this serial-queue seemed to behave like an $M/M/1$

queue. This observation led us to the conjecture, which had already been made by others, that the output process of the n^{th} queue in the steady state of this serial network with general renewal input approaches a Poisson process (mean identical to that of the input process) as $n \rightarrow \infty$. In the process of studying this conjecture, we found that the techniques which have been developed for interacting particle systems are useful for the understanding of the flow of customers in an infinite series of exponential queues. With the classical approach of queueing theory, when the input process is Poisson, we only know that this serial queue has a product form solution. In particular, we do not know how this network reaches equilibrium. Using simple models of interacting particle systems namely, the simple exclusion process and the zero range process, in Chapter 4, we study the flow properties of the customers in this serial network.

The main results of Chapter 4 are: the derivation of the density profile and the hydrodynamic limit for the unsaturated exponential queues in series, development of stochastic ordering methods for the unsaturated exponential queues in series with non-Poisson input and the derivation of the density profile for more complicated exponential queues in series with feedback mechanism and general input. In Section 4.2, we introduce a *Road Hog* process (in which the *leading particle* moves slowly when compared to other particles) to study the hydrodynamical limit of unsaturated exponential queues in series. In Section 4.3, we obtain this limit via an *embedding method* for the multi-coloured particle system. Section 4.5 deals with exponential queues in series with non-Poisson input. Here, we introduce a new coupling technique for the multi-coloured particle system which enables us to squeeze the density profile for the queues in series with non-Poisson input between known density profiles obtained in Section 4.3 for the queues in series with Poisson input. In Section 4.6, we consider feedback queues in series with non-Poisson input and derive

the hydrodynamic limit and density profile.

This hydrodynamical study of simple queueing networks provides some valuable insights into the understanding of these networks and the results obtained in Section 4.5 and 4.6 are a step toward proving the above mentioned conjecture.

CHAPTER 1

THE $D^s/M_n/1$ QUEUE

In this chapter, we consider a state dependent single server queueing system with deterministic bulk input and exponential service times which vary with the number of jobs in the system. Jobs arrive at the system at discrete time points which are equally spaced at unit time apart. At these points, s jobs are scheduled to arrive. Each of the jobs scheduled to arrive at these points arrives with probability p and does not arrive with probability $1 - p$, so that the total number of jobs arriving at each of these points is a random variable. In queueing literature, this would normally be designated as a $D^X/M_n/1$ queue. The server works at different speeds, depending on the number of jobs in the system. When there are n jobs in the system, service times are exponentially distributed with rate μ_n . We also assume that the service rate changes instantaneously. There is also a limit on the maximum rate at which the server can operate. As regards related work, the analysis of $D/M/1$ and $D/M/r$ queues can be found in standard text books like Prabhu(1965), Cohen(1969), Page(1972) and Conolly(1975). The $D^s/M/r$ queue was studied by Cox and Hinkley(1970) by means of the imbedded Markov chain associated with this queue.

Following Cox and Hinkley (1970), we study the imbedded Markov Chain arising from this $D^s/M_n/1$ queue by considering the number of jobs in the system immediately after an arrival of jobs. Recall that the conventional imbedded Markov chain studied in the literature considers the number of jobs in the system immediately before an arrival epoch. The main reason for studying the imbedded Markov chain immediately after an arrival

epoch is to use some of the interesting waiting measures introduced by Cox and Hinkley (1970). We also develop formulae for general equilibrium probabilities in addition to the conditional ones. The general and conditional equilibrium probabilities are complicated functions of the interarrival and service rates and it is not possible to gain any insights from them. These formulae are however, very suitable for computational purposes and numerical results are provided at the end of the chapter. This chapter is organized as follows. After introducing some notations in Section 1.1, we provide the analysis of this queue via the embedded Markov chain in Section 1.2. In Section 1.3, we discuss the delay measures and in Section 1.4, we provide the numerical comparisons for various parameters of this queue. We also compare the results for simpler systems like $D/M/1$ and $D/M/r$ queues which are deduced from our results as special cases. In Chapter 3 (Section 3.5), the $D/M_n/1$ queue ($s = 1, p = 1$) will be used as an approximation to calculate the total average number in a semi-open queueing network with deterministic input and exponential service times.

1.1 Preliminaries

Jobs are scheduled to arrive at time points $t_n, n = 0, 1, \dots$ in batches of size s . There is an independent probability p that any one of the jobs scheduled to arrive at t_n in fact does arrive: while $1 - p$ is the probability it does not arrive at t_n . Without loss of generality, let $t_{n+1} - t_n = 1$ for all n and $t_0 = 0$. Let A_i be the (random) number of jobs that arrive in a given batch; then

$$P\{A_i = h\} = a_h = \binom{s}{h} p^h (1-p)^{s-h} \quad h = 0, 1, \dots, s. \quad (1.1)$$

with $E\{A_i\} = sp$ and $Var\{A_i\} = spq$.

Jobs are served on a FCFS (First Come First Served) basis. The service rates depend

on the number of jobs in the system. There is limit on the maximum service rate at which the server can complete the jobs. Let μ_n be the service rate when there are n jobs in the system. When there are n jobs in the system, the service times are exponential with parameter μ_n . We also assume that $\mu_1 < \mu_2 < \mu_3 < \dots < \mu_N$ and that μ_N is the maximum service rate, i.e.

$$\mu_n = \begin{cases} \mu_n, & \text{if } n < N; \\ \mu_N, & \text{otherwise.} \end{cases} \quad (1.2)$$

Note that these are instantaneous state dependent service rates.

1.2 Analysis

Let $N(t)$ be the number of jobs in the system at time t and $N_n = N(t_n + 0)$ be the number of jobs in the system immediately after an arrival. We could analyse the system by considering the process $N(t)$ in the interval $0 \leq t \leq 1$, but we prefer to study the embedded process N_n . Let Y_n be the number of service completions in the unit interval. Recalling that A_n is the number of jobs that arrive at the n^{th} epoch we have

$$N_{n+1} = N_n - Y_n + A_{n+1}. \quad (1.3)$$

Since the service times are exponential with instantaneous state dependent rates, the distribution of Y_n given $N_n = l$ is independent of N_{n-1} , N_{n-2} etc. and hence N_n is a Markov chain. Suppose that $\{N_n\}$ converge in distribution as $n \rightarrow \infty$. To find the conditional equilibrium distribution which exists when $sp/\mu_N < 1$ (Chaudhry and Templeton(1983)), we proceed as follows. The condition $sp/\mu_N < 1$ can be explained as follows: as long as the arrival rate of the jobs(sp) is less than the maximum service rate μ_N , the queue will be stable.

Now to find the transition probability matrix P of the imbedded chain N_n , consider the distribution of Y_n given $N_n = l$. Let $q_{lm} = P(N_n - Y_n = m | N_n = l)$. To calculate

q_{lm} , we first consider the quantities $q_{lm}(t)$ defined for $l \geq m$ as

$$\begin{aligned} q_{lm}(t) &= P\{\text{exactly } l-m \text{ service completions in time } t / l \text{ jobs in the system initially}\} \\ &= P\{X_l + X_{l-1} + \dots + X_{m+1} \leq t, X_l + X_{l-1} + \dots + X_m > t\} \end{aligned} \quad (1.4)$$

where X_i is the time to complete the next service when there are i jobs in the system. Let

$$K_n(s, t) = P\{X_s + X_{s-1} + \dots + X_{s-n+1} > t\}$$

then

$$q_{lm}(t) = K_{l-m+1}(l, t) - K_{l-m}(l, t) \quad (1.5)$$

or equivalently let $Z_1 = \sum_{i=m+1}^l X_i$, then

$$q_{lm}(t) = f_{Z_1} * \bar{F}_{X_m}(t) \quad (1.6)$$

where $f_{Z_1}(\cdot)$ is the density of Z_1 , $\bar{F}_{X_m}(\cdot)$ is the complimentary cumulative distribution of X_m and $*$ denotes convolution. Note that X_i is distributed exponentially with parameter μ_i , and $q_{lm} = q_{lm}(1)$. In order to calculate q_{lm} explicitly, we consider the following cases:

Case i: We first note that when $l < m$, from the definition of $q_{lm}(t)$ (equation (1.4)), $q_{lm} = 0$.

Case ii: When $l, m \geq N$, Z_1 in (1.6) is distributed as the sum of $(l - m)$ independent exponential random variables with rate μ_N . Therefore for $l, m \geq N$, (from (1.6))

$$q_{lm} = \frac{\mu_N^{l-m} e^{-\mu_N}}{(l-m)!} \quad (1.7)$$

Case iii: When $l < N$, Z_1 in (1.6) is distributed as the sum of $(l - m)$ independent exponential random variables with rates $\mu_l, \mu_{l-1}, \dots, \mu_{m+1}$. Therefore,

$$q_{lm} = \begin{cases} 1 - \sum_{i=1}^l A_i(1, l) e^{-\mu_i} & \text{if } 1 \leq l \leq N, m = 0; \\ \frac{1}{\mu_m} \sum_{i=m}^l A_i(m, l) \mu_i e^{-\mu_i} & 1 \leq l, m \leq N. \end{cases} \quad (1.8)$$

where

$$A_i(m, n) = \prod_{j=m}^n \frac{\mu_j}{\mu_j - \mu_i} \quad (1.9)$$

Case iv: When $l > N$ and $m \leq N$, Z_1 in (1.6) is distributed as the sum of the two independent random variables. The first is a Erlang random variable with $l - N$ phases and rate μ_N and the second is the sum of $N - m$ random variables with rates $\mu_N, \mu_{N-1}, \dots, \mu_{m+1}$. For $l > N$ and $m = 0$,

$$\begin{aligned} q_{l0} &= P(\mu_N, l - N - 1) - \frac{\mu_N^{l-N}}{(l - N)!} e^{-\mu_N} A_N(1, N) \\ &\quad - \sum_{i=1}^{N-1} A_i(1, N) e^{-\mu_i} \left(\frac{\mu_N}{\mu_N - \mu_i} \right)^{l-N} P(\mu_N - \mu_i, l - N - 1). \end{aligned} \quad (1.10)$$

For $l > N, 0 < m < N$,

$$\begin{aligned} q_{lm} &= \frac{e^{-\mu_N} \mu_N^{l-N+1}}{\mu_m (l - N)!} A_N(m, N) \\ &\quad + \sum_{i=m}^{N-1} A_i(m, N) \mu_i e^{-\mu_i} \left(\frac{\mu_N}{\mu_N - \mu_i} \right)^{l-N} P(\mu_N - \mu_i, l - N - 1). \end{aligned} \quad (1.11)$$

where

$$P(\mu, l - N - 1) = 1 - \sum_{j=0}^{l-N-1} \frac{e^{-\mu} \mu^j}{j!}. \quad (1.12)$$

Now the elements of the transition probability matrix P are

$$\begin{aligned} p_{ij} &= P\{N_{n+1} = j | N_n = i\} \\ &= \sum_{h=0}^j q_{ih} a_{j-h}. \end{aligned} \quad (1.13)$$

by the independence of A_{i+1} and $N_i - Y_i$. With the matrix of the imbedded chain P determined by these equations, the conditional equilibrium distribution π satisfying $\pi = \pi P$ can be found in principle. Note that $\pi_0 = \pi_1 = \pi_2 = \dots = \pi_{p-1} = 0$ if $p = 1$.

The above equilibrium distribution π applies only at the epoch immediately after the arrival. The general equilibrium distribution $\hat{\pi}$ at an arbitrary epoch can be defined as follows:

$$\hat{\pi}_j = \sum_{k=j}^{\infty} \pi_k \int_0^1 q_{kj}(s) ds. \quad (1.14)$$

Now substituting for $q_{kj}(s)$ and simplifying, we obtain

$$\begin{aligned} \hat{\pi}_0 = & \sum_{k=0}^N \pi_k \left\{ 1 - \sum_{i=1}^k A_i(1, k) d_0(\mu_i) \right\} \\ & + \sum_{k=N+1}^{\infty} \pi_k \left[1 - \sum_{i=0}^{k-N-1} \frac{\mu_N d_i(\mu_N)}{i!} - \frac{\mu_N^{k-N} A_N(1, N) d_{k-N}(\mu_N)}{(k-N)!} \right. \\ & \left. - \sum_{i=1}^{N-1} A_i(1, N) \left(\frac{\mu_N}{\mu_N - \mu_i} \right)^{k-N} \left(d_0(\mu_i) - \sum_{j=0}^{k-N-1} \frac{(\mu_N - \mu_i)^j d_j(\mu_N)}{j!} \right) \right] \end{aligned} \quad (1.15)$$

For $j = 1, 2, \dots, N-1$,

$$\begin{aligned} \hat{\pi}_j = & \pi_j d_0(\mu_j) + \sum_{k=j+1}^N \frac{\pi_k}{\mu_j} \sum_{l=j}^k A_l(j, k) \mu_l d_0(\mu_l) \\ & + \frac{1}{\mu_j} \sum_{k=N+1}^{\infty} \pi_k \left[\frac{\mu_N^{k-N+1} A_N(j, N) d_{k-N}(\mu_N)}{(k-N)!} \right. \\ & \left. - \sum_{i=j}^{N-1} A_i(j, N) \left(\frac{\mu_N}{\mu_N - \mu_i} \right)^{k-N} \mu_i \left(d_0(\mu_i) - \sum_{l=0}^{k-N-1} \frac{(\mu_N - \mu_i)^l d_l(\mu_N)}{l!} \right) \right] \end{aligned} \quad (1.16)$$

For $j = N, N+1, \dots, \infty$,

$$\hat{\pi}_j = \pi_j d_0(\mu_N) + \sum_{k=j+1}^{\infty} \frac{\pi_k \mu_N^{k-j} d_{k-j}(\mu_N)}{(k-j)!} \quad (1.17)$$

where $d_j(\mu_i)$ can be calculated recursively by

$$d_j(\mu_i) = \frac{-e^{-\mu_i}}{\mu_i} + \frac{j}{\mu_i} d_{j-1}. \quad (1.18)$$

1.3 Delay Measures

In the last section, we obtained the equilibrium distributions of the number of jobs in the system immediately after an arrival epoch as well as at an arbitrary epoch. The frequently used delay measure, namely, the average waiting time of a job could be easily calculated using Little's formula from $\hat{\pi}$. We could also calculate delay measures directly from the equilibrium distribution obtained from the imbedded Markov chain, π .

In general, for queueing systems with deterministic arrivals, the following delay measures introduced by Cox and Hinkley(1970) seem more appropriate: 1. the probability that an arriving job is subjected to any delay and 2. the delay is more than a half time unit. The motivation for considering these two delay measures is that a job arriving at the lattice point usually has a good chance of completing its service before the next lattice point at which further jobs arrive. In the case of jobs scheduled to arrive at the system, these measures would serve best as a criteria to minimize the delay of the jobs.

The first delay measure is defined more precisely by

$$\phi_0 = \frac{\sum_{j=1}^{\infty} j \pi_{N+j}}{sp} \quad (1.19)$$

This is the ratio of the expected number of delays to expected number of arrivals. In many practical applications, a job delayed at the start has a good chance of starting its service before the next time point at which jobs are scheduled to arrive, whereas a delay beyond the half time point would be regarded as serious. Therefore in a way similar to ϕ_0 , we define $\phi_{\frac{1}{2}}$ as

$$\phi_{\frac{1}{2}} = \sum_{h=0}^{\infty} \sum_{j=1}^{\infty} \frac{j \pi_{N+h+j} b_h}{sp} \quad (1.20)$$

where b_h is $q_{lm}(\frac{1}{2})$ when $l, m \geq N$. These delay measures can be used to calculate the optimum batch size to be delivered to the system such that the delay measures given in

(1.19) or (1.20) achieves its minimum.

1.4 Numerical Results

In order to evaluate the equilibrium distributions just after the arrival instants and at arbitrary time points, we have set an artificial bound on the queue size in order to keep the ranges of probabilities and summation finite. A bound was selected such that there is a small probability of it being exceeded in the unrestricted system. We then calculated the transition probability matrix, assuming a finite number of states. Due to the truncation of the state space, the row sum of this matrix P would not be 1. We distributed the difference in the row sum and 1 uniformly over the state space so that the elements in a row add up to 1. With this adjusted transition matrix, we solved for π and $\hat{\pi}$.

To test the accuracy of the truncation procedure and the theoretical results, we calculate the conditional and general equilibrium probabilities for $D/M/1$, $D/M/r$ queues. Also, numerical examples involving more complicated service rates will be presented in Section 3.5, where we use the $D/M_n/1$ queue as an approximation to the semi-open queueing network with deterministic input. Note that $D/M/1$, $D/M/r$ queues are special cases of the $D^*/M_n/1$ queue, i.e. when $s = 1$ and $p = 1$, the $D^*/M_n/1$ becomes a $D/M_n/1$ queue. For the cases $N = 1$, $\mu_n = \mu$ and $N = r$, $\mu_n = n\mu$, the $D/M_n/1$ queue reduces to a $D/M/1$ and $D/M/r$ queue respectively. Since no explicit results are available for the conditional equilibrium probabilities π , we compared the general equilibrium probabilities $\hat{\pi}$ for these queues with the tables in Hillier and Yu (1981). Tables 1.1 to 1.3 compare the general equilibrium probabilities for the $D/M/1$ queue when $\rho = 0.2, 0.5, 0.8$. Tables 1.4 to 1.6 provide the comparisons for $D/M/2$ queue when $\rho = 0.2, 0.5, 0.8$. Tables 1.7 to 1.9 give the results for $D/M/3$ queue when $\rho = 0.2, 0.5, 0.8$. As the tables indicate, our method is very accurate when compared to Hillier and Yu's (1981) figures, and the

truncation of the state space does not have much influence on the results.

Equilibrium Distribution of the Number in the system For The $D^s/M_n/1$ Queue

$\rho = 0.2, s = 1, p = 1, N = 1$		
N	D/M/1	Hillier & Yu
0	.800000E+00	.800000E+00
1	.198605E+00	.198605E+00
2	.138569E+00	.138570E+00
3	.966820E-02	.966834E-02
4	.674565E-07	.674578E-07
5	.470655E-09	.470668E-09
6	.328382E-11	.328394E-11
7	.229075E-13	.229126E-13
8	.157048E-15	.159866E-15

TABLE 1.1
D/M/1 QUEUE with $\rho = 0.2$

$\rho = 0.5, s = 1, p = 1, N = 1$		
N	D/M/1	Hillier & Yu
0	.500000E+00	.500000E+00
1	.398406E+00	.398405E+00
2	.809513E-01	.809520E-01
3	.164483E-01	.164486E-01
4	.334210E-02	.334220E-02
5	.679074E-03	.679100E-03
6	.137980E-03	.137987E-03
7	.280358E-04	.280375E-04
8	.569653E-05	.569695E-05
9	.115747E-05	.115756E-05
10	.235183E-06	.235205E-06
11	.477863E-07	.477912E-07
12	.970960E-08	.971071E-08
13	.197287E-08	.197311E-08
14	.400864E-09	.400917E-09

TABLE 1.2
D/M/1 QUEUE with $\rho = 0.5$

$\rho = 0.8, s = 1, p = 1, N = 1$		
N	D/M/1	Hillier & Yu
0	.200000E+00	.200000E+00
1	.297097E+00	.297096E+00
2	.186764E+00	.186763E+00
3	.117405E+00	.117405E+00
4	.738045E-01	.738044E-01
5	.463957E-01	.463957E-01
6	.291657E-01	.291657E-01
7	.183345E-01	.183344E-01
8	.115256E-01	.115256E-01
9	.724532E-02	.724532E-02
10	.455462E-02	.455462E-02
11	.286316E-02	.286318E-02
12	.179986E-02	.179988E-02
13	.113143E-02	.113146E-02

TABLE 1.3
D/M/1 QUEUE with $\rho = 0.8$

$\rho = 0.2, s = 1, p = 1, N = 2$		
N	D/M/2	Hillier & Yu
0	.617771E+00	.617771E+00
1	.364459E+00	.364459E+00
2	.176466E-01	.176466E-01
3	.121230E-03	.123123E-03
4	.859049E-06	.859048E-06
5	.599372E-08	.599370E-08
6	.418191E-10	.418190E-10
7	.291773E-12	.291778E-12
8	.203131E-14	.203578E-14
9	.142145E-16	.142039E-16

TABLE 1.4
D/M/2 QUEUE with $\rho = 0.2$

$\rho = 0.5, s = 1, p = 1, N = 2$		
N	D/M/2	Hillier & Yu
0	.254565E+00	.254564E+00
1	.490871E+00	.490871E+00
2	.202840E+00	.202840E+00
3	.412147E-01	.412147E-01
4	.837432E-02	.837430E-02
5	.170156E-02	.170156E-02
6	.345736E-03	.345737E-03
7	.702494E-04	.702495E-04
8	.142738E-04	.142739E-04
9	.290027E-05	.290027E-05
10	.589300E-06	.589300E-06
11	.119739E-06	.119738E-06
12	.243294E-07	.243294E-07
13	.494346E-08	.494344E-08

TABLE 1.5
D/M/2 QUEUE with $\rho = 0.5$

$\rho = 0.8, s = 1, p = 1, N = 2$		
N	D/M/2	Hillier & Yu
0	.721589E-01	.721584E-01
1	.255683E+00	.255683E+00
2	.249620E+00	.249620E+00
3	.156919E+00	.156918E+00
4	.986439E-01	.986432E-01
5	.620105E-01	.620103E-01
6	.389816E-01	.389815E-01
7	.245050E-01	.245050E-01
8	.154046E-01	.154046E-01
9	.968377E-02	.968376E-02
10	.608750E-02	.608750E-02
11	.382677E-02	.382678E-02
12	.240560E-02	.240563E-02
13	.151221E-02	.151225E-02

TABLE 1.6
D/M/2 QUEUE with $\rho = 0.8$

$\rho = 0.2, s = 1, p = 1, N = 3$		
N	D/M/3	Hillier & Yu
0	.468127E+00	.468141E+00
1	.465427E+00	.465418E+00
2	.647461E-01	.647409E-01
3	.168833E-02	.168815E-02
4	.117797E-04	.117785E-04
5	.821890E-07	.821804E-07
6	.573445E-09	.573386E-09
7	.400101E-11	.400060E-11
8	.279102E-13	.279128E-13
9	.190735E-15	.194752E-15

TABLE 1.7
D/M/3 QUEUE with $\rho = 0.2$

$\rho = 0.5, s = 1, p = 1, N = 3$		
N	D/M/3	Hillier & Yu
0	.121888E+00	.129890E+00
1	.396613E+00	.396613E+00
2	.341104E+00	.341102E+00
3	.111869E+00	.111868E+00
4	.227303E-01	.227302E-01
5	.461853E-02	.461849E-02
6	.938429E-03	.938420E-03
7	.190677E-03	.190676E-03
8	.387433E-04	.387430E-04
9	.787217E-05	.787210E-05
10	.159953E-05	.159952E-05
11	.325005E-06	.325003E-06
12	.660371E-07	.660365E-07
13	.134180E-07	.134179E-07

TABLE 1.8
D/M/3 QUEUE with $\rho = 0.5$

$\rho = 0.8, s = 1, p = 1, N = 3$		
N	D/M/3	Hillier & Yu
0	.230712E-01	.230713E-01
1	.137191E+00	.137190E+00
2	.256407E+00	.256405E+00
3	.216633E+00	.216633E+00
4	.136182E+00	.136182E+00
5	.856082E-01	.856080E-01
6	.538159E-01	.538156E-01
7	.338302E-01	.338301E-01
8	.212667E-01	.212666E-01
9	.133688E-01	.133688E-01
10	.840403E-02	.840408E-02
11	.528300E-02	.528303E-02
12	.332102E-02	.332107E-02
13	.208766E-02	.208773E-02

TABLE 1.9
D/M/3 QUEUE with $\rho = 0.8$

CHAPTER 2

THE $M/G_n/1$ QUEUE

One of the important queueing models in manufacturing systems is the one in which the service rate can be varied depending on the number of jobs in the system. Especially, if a queueing system represents a machining centre, when the number of jobs to be processed is very large, by assigning more man power and increasing the capacity of the machine tools, the processing times of the jobs can be shortened and thereby decreasing the in-process inventory. However, the resources required to increase the service rates are very expensive and it may not be possible to operate the machine at the highest capacity. In order to be cost effective, we can switch the service rates depending on the number of jobs in the system. In this chapter, we consider such a single server queue in which N types of service are available to the server. The server operates as follows. If at the initiation of service, there are $k = 1, 2, \dots, N - 1$ jobs in the system, type k service will be used; and when there are N or more jobs in the system, type N service will be used. We also assume that type $k + 1$ service is faster than the type k service. Jobs arrive at the system according to a Poisson process. The purpose of this chapter is to find an efficient recursive relation for the stationary distributions of the number of jobs in this system.

Previously, Federgruen and Tijms (1980) considered a queueing system in which there were only two types of service available to the server. In their work, the service type changes from type 1 to type 2 only when the number of jobs in the system at the initiation of service epoch is larger than, say R_1 , and the service type switches from type 2 to type 1 when the number of jobs in the system is smaller than, say R_2 ($R_1 \geq R_2$). See Gross

and Harris (1985) and the references cited there for related papers. After introducing some preliminary notations in Section 2.1, in Section 2.2, we use the powerful regenerative approach and up and down crossing analysis to derive a recursion relationship for the stationary distribution of the number of jobs in the system. This recursion relationship is very suitable for computational purposes. In Section 2.3, for particular choices of service types considered in Section 2.2, we show that our recursion relationship reduces to that of Federgruen and Tijms (1980) and Tijms (1986). In that section, we also provide a numerical example to highlight the efficiency of the recursion relationship.

2.1 Preliminaries

Consider a state dependent single server queue in which jobs arrive according to a Poisson process with parameter λ . An arriving job is serviced immediately if no other jobs are present in the system and it waits if the server is busy. Each job is serviced by one of finite number of service types available from the server. We assume that there are N types of service available. We also assume that type $k + 1$ service is faster than the type k service for $k = 1, 2, \dots, N - 1$, and that type N service is the fastest one.

Let

S_k = a positive random variable representing the service time of a job under the service type k , $k = 1, 2, \dots, N$.

$F_k(\cdot)$ = the distribution of S_k , $k = 1, 2, \dots, N$.

$E(S_k)$ = mean of S_k , $E(S_k) < \infty$ for $k = 1, 2, \dots, N$; $E(S_1) > E(S_2) > \dots > E(S_N)$

The service rule is characterized as follows. If at the service initiation epoch, there are $k = 1, 2, \dots, N - 1$ jobs in the system, the new service starts with service type k (the service time random variable is S_k and the distribution is F_k), and if there are N or more jobs

in the system, the service type is N . The service type changes only at the initiation of service epoch. We assume that $\lambda E(S_1) < 1$. The queueing discipline is FCFS.

2.2 Analysis

We assume that a job who arrives at epoch 0 finds the server idle. In order to derive recursion relationship for the stationary distribution of the number of jobs in the system, first we introduce the following random variables.

T = the next epoch at which an arriving job finds the system empty, i.e. T is the length of the busy cycle of the queue.

L = number of jobs served in a busy cycle $(0, T]$.

$T_k(i)$ = the amount of time during which i jobs are in the system and type k service is used in the busy cycle $(0, T]$, $k = 1, 2, \dots, N$ and $i \geq k$. Note that type k service implies that the service starts with k jobs in the system.

$L_k(i)$ = the number of service completion epochs at which the departing job leaves i jobs behind and was served by a type k service in the busy cycle $(0, T]$, $k = 1, 2, \dots, N$ and $i \geq k - 1$.

Let

$p_k(i, t)$ = probability that there are i jobs in the system and type k service is used at time t ($i \geq 1, k = 1, 2, \dots, N$).

$p(0, t)$ = probability that the server is idle at time t .

Now if we define the stochastic process $X(t)$ as the number of jobs in the system at time t , then $X(t)$ is a regenerative process and this process regenerates at the epochs when an arriving job finds the system empty or server idle. Since the expected cycle time

$E(T) < \infty$, we have

$$p_k(i) = \lim_{t \rightarrow \infty} p_k(i, t) \text{ and } p(0) = \lim_{t \rightarrow \infty} p(0, t) \text{ exists.}$$

Now by the memoryless property of the Poisson arrival process $p(0)$ reduces to

$$p(0) = \frac{1/\lambda}{E(T)} = \frac{1}{\lambda E(T)} \quad (2.1)$$

From the theory of regenerative process (Heyman and Sobel(1982)), $p_k(i)$ is given by

$$p_k(i) = \frac{ET_k(i)}{E(T)} \quad \text{for } k = 1, 2, \dots, N \text{ and } i < k. \quad (2.2)$$

We can interpret the $p_k(i)$ as the long run expected fraction of the time during which i jobs are present and service type k is used. Let $p(i) = \sum_{k=1}^N p_k(i)$ and note that $\sum_{i=0}^{\infty} p(i) = 1$.

Since Poisson arrivals see time averages, i.e. the limiting distribution of the number of jobs in the system at an arbitrary epoch is the same as the limiting distribution of the number of jobs in the system seen by an arriving job (see Cooper(1981) and Wolf (1982)). In other words,

$$\text{Time average probability} = \text{Job average probability.}$$

Note that this relationship holds only for the queueing system with Poisson arrival. Following Stidham(1972), we can interpret $p(i)$ as the long run expected fraction of the jobs who find i jobs in the system upon arrival. The quantity $\sum_{k=1}^N EL_k(i)/E(L)$ is the long run fraction of the jobs which leave i jobs in the system upon departure. Since in any time interval, the number of times the queue size increases from i to $i+1$ cannot differ by more than one from the number of times the queue size decreases from $i+1$ to i (this is basically an upcrossing and downcrossing argument given by Stidham(1972)), we have

from (2.2)

$$p(i) = \frac{\sum_{k=1}^N EL_k(i)}{E(L)}, \quad i = 0, 1, \dots \quad (2.3)$$

Also from the theory of regenerative process, $E(L)/E(T)$ is the long run average number of jobs served per unit time and hence

$$E(L) = \lambda E(T). \quad (2.4)$$

Because $p_k(i)$ is the probability that i jobs are in the system and type k service is used and since type k service starts only when the k jobs are in the system we have $p_k(i) = 0$ when $i < k$. Therefore $p(i)$, which was defined as $\sum_{k=1}^N p_k(i)$ reduces to

$$p(i) = \sum_{k=1}^i p_k(i), \quad \text{for } i = 1, 2, \dots \quad (2.5)$$

Note also that in equation (2.3) the quantities $L_k(i)$'s are non-zero only when $i \geq k - 1$.

Therefore equation (2.3) reduces to two expressions.

For $i = 1, 2, \dots, N - 1$

$$p(i) = \frac{\sum_{k=1}^{i+1} EL_k(i)}{E(L)} \quad (2.6)$$

and for $i \geq N$,

$$p(i) = \frac{\sum_{k=1}^N EL_k(i)}{E(L)}. \quad (2.7)$$

Using (2.6) and (2.7), equation (2.4) reduces to

$$\sum_{k=1}^{i+1} EL_k(i) = p(i)\lambda E(T) \quad i = 1, 2, \dots, N - 1. \quad (2.8)$$

and

$$\sum_{k=1}^N EL_k(i) = p(i)\lambda E(T), \quad i \geq N. \quad (2.9)$$

Now to obtain a relationship between $ET_k(i)$ and $EL_k(i)$, we first divide the cycle into random number of disjoint intervals by means of service completion epochs and then calculate ET_k as the sum of the contributions to ET_k from various intervals. For this purpose, we define for $k = 1, 2, \dots, N$ and $j \geq 1$,

$A_k(i, j)$ = expected amount of time during which i jobs are present in one service time of type k and this service started with j jobs.

Note that $A_k(i, j)$ is valid only for $i \geq j$. Also, when the service type is j (which starts with j jobs in the system), we have $A_k(i, j) = A_j(i, j)$. Given that at epoch 0, service type k starts when j jobs are present, the probability that at epoch t this service is still in progress and $i \geq j$ jobs are present in the system is

$$(1 - F_k(t)) \frac{e^{-\lambda t} (\lambda t)^{i-j}}{(i-j)!}, \quad t > 0.$$

and hence

$$A_k(i, j) = \int_0^\infty (1 - F_k(t)) \frac{e^{-\lambda t} (\lambda t)^{i-j}}{(i-j)!} dt. \quad (2.10)$$

Now using Wald's identity for the expectation of random sum of independent random variables we obtain the following system of equations relating $ET_k(i)$ and $EL_k(i)$. See Fig 2.1 for a schematic representation of a typical sample path of a busy cycle.

When $k = 1$, the expected amount of time during which i jobs are in the system and type 1 service is used in one busy cycle, i.e. $ET_1(i)$ is obtained by splitting the busy cycle as follows. Since a job always arrives to an empty system at epoch 0, type one service starts at epoch 0 and the expected amount of time during which i jobs are present is $A_1(i, 1)$. Also $L_1(1) + L_2(1)$ is the total number of service completion epochs at which service type one starts with one job. Therefore $A_1(i, 1)\{EL_1(1) + EL_2(1)\}$ is the contribution to $ET_1(i)$. Combining these two quantities we have for $k = 1$:

$$ET_1(i) = A_1(i, 1) + A_1(i, 1)\{EL_1(1) + EL_2(1)\}, \quad i \geq 1. \quad (2.11)$$

Figure 2.1

Sample Path of a Busy Cycle

Similarly, when $k = 2, 3, \dots, N - 1$, the contribution to $ET_k(i)$, $i \geq k$ comes from the number of epochs at which type k service starts with k jobs and $i \geq k$ jobs are in the system before this service ends. Since type 1 to type $k + 1$ services could leave k jobs at the end of service, the total number of such epochs is $\sum_{l=1}^{k+1} EL_l(k)$. Therefore for $k = 1, 2, \dots, N - 1$,

$$ET_k(i) = A_k(i, k) \sum_{l=1}^{k+1} EL_l(k), \quad i \geq k. \quad (2.12)$$

When $k = N$, the calculation of $ET_N(i)$, $i \geq N$ as the contribution from various intervals of the busy cycle becomes much more complicated because type N service starts whenever there are more than N jobs in the system. We explain the calculation of $ET_N(i)$ only when $i = N + 1$. This argument is easily extended for other i 's.

When $i = N + 1$, service types 1 to N could leave N jobs at the service completion epochs and type N service starts with N jobs. The total number of such service completion epochs is $\sum_{l=1}^N EL_l(N)$. Starting at these epochs, the expected amount of time during which $N + 1$ jobs are in the system is $A_N(N + 1, N)$. Similarly, service types 1 to N could

leave $N+1$ jobs at the service completion epochs and this quantity is $\sum_{l=1}^N EL_l(N+1)$ and the expected amount of time during which $N+1$ jobs are in the system is $A_N(N+1, N+1)$.

Now generalizing the above argument for any $i \geq N$ we have for $k = N$:

$$\begin{aligned} ET_N(i) &= \sum_{j=N}^i \left\{ \sum_{l=1}^{N-1} EL_l(j) + EL_N(j) \right\} A_N(i, j) \\ &= \sum_{j=N}^i \left\{ \sum_{l=1}^N EL_l(j) \right\} A_N(i, j), \quad i \geq N. \end{aligned} \quad (2.13)$$

Equations (2.11), (2.12) and (2.13) can be simplified using equations (2.2) and (2.8). For example, when $i = 1$ the LHS of (2.11) reduces, using (2.2), to $p_1(i)E(T)$ and the RHS reduces, using (2.8), to $A_1(i, 1)\lambda E(T)p(1)$. Therefore (2.11) becomes

$$p_1(i)E(T) = A_1(i, 1) + A_1(i, 1)\lambda E(T)p(1), \quad i \geq 1. \quad (2.14)$$

Similarly, for $k = 2$ equation (2.12) reduces to

$$p_2(i)E(T) = A_2(i, 2)\lambda E(T)p(2), \quad i \geq 2. \quad (2.15)$$

and for $k = 3$, equation (2.12) becomes

$$p_3(i)E(T) = A_3(i, 3)\lambda E(T)p(3), \quad i \geq 3. \quad (2.16)$$

In general, for $k = 2, 3, \dots, N-1$ we have

$$p_k(i)E(T) = A_k(i, k)\lambda E(T)p(k), \quad i \geq k. \quad (2.17)$$

For $k = N$

$$p_N(i)E(T) = \sum_{j=N}^i p(j)\lambda E(T)A_N(i, j), \quad i \geq N. \quad (2.18)$$

Since

$$p(i) = \sum_{k=1}^i p_k(i), \quad 2 \leq i \leq N-1 \quad (2.19)$$

and

$$p(i) = \sum_{k=1}^N p_k(i), \quad i \geq N. \quad (2.20)$$

Adding the corresponding equations in (2.14), (2.17) and (2.18) we obtain

$$p(i)E(T) = A_1(i, 1) + \sum_{j=1}^i \lambda E(T) A_j(i, j) p(j), \quad 1 \leq i \leq N-1, \quad (2.21)$$

and

$$p(i)E(T) = A_1(i, 1) + \sum_{j=1}^{N-1} \lambda E(T) A_j(i, j) p(j) + \sum_{j=N}^i \lambda E(T) A_N(i, j) p(j), \quad i \geq N \quad (2.22)$$

These recursion equations for $p(i)$, together with

$$p(0) = \frac{1}{\lambda E(T)},$$

are very natural for computing $p(i)$'s. We do not have to evaluate the term $E(T)$ explicitly in order to evaluate the $p(i)$'s. In equations (2.21) and (2.22) let $q(i) = p(i)E(T)$ for $i = 0, 1, \dots$. Given an M such that $\sum_{i=0}^M p(i) \geq 1 - \epsilon$, we can calculate $q(i)$ recursively by

$$p(i) = \frac{q(i)}{E(T)} = \frac{q(i)}{\sum_{i=0}^M q(i)}. \quad (2.23)$$

since

$$\sum_{i=0}^M q(i) = \sum_{i=0}^M p(i)E(T).$$

Once the bound M is known, starting with $q(0) = 1/\lambda$, we can calculate $q(i)$, $i > 0$ recursively from the equations (2.21) and (2.22).

2.3 Discussion

In order to verify our analysis, we evaluate the recursive equations for the special case when there are two types of service treated by Federgruen and Tijms (1980). They

specified two levels of service for the $M/G_n/1$ queue namely, the service type changes from first level to second level only when the number of jobs at the service initiation epoch is greater than R_1 and the service type changes from level two to level one only when the number of jobs at the service initiation epoch is less than or equal to R_2 . In particular, when $R_1 = R_2 = 1$, type one service is offered whenever there is one job at the service initiation epoch; otherwise type two service is used. This special case corresponds to $N = 2$ in our case, since there are only two types of service, one type when exactly one job is present and the other when more than one job is present at the service initiation epoch. Equations (9) and (12) of Fredergruen and Tijms(1980) are

$$p(1)E(T) = A_1(1, 1) + \lambda A_1(1, 1)p(1)E(T) \quad (2.24)$$

and

$$p(i)E(T) = A_1(i, 1) + \lambda A_1(i, 1)p(1)E(T) + \sum_{j=2}^i \lambda A_2(i, j)p(j)E(T), \quad i \geq 2. \quad (2.25)$$

It is easy to verify that when $N = 2$ our equations (2.21) and (2.22) are same as the ones given above.

As a further check, we consider the case when only one service type is available to the server, i.e, $N = 1$. In this case (when $N = 1$), the $M/G_n/1$ queue reduces to an $M/G/1$ queue. When $N = 1$, equations (2.21) and (2.22) reduces to

$$p(i) E(T) = A(i, 1) + \sum_{j=1}^i \lambda E T A(i, j) p(j), \quad i \geq 1. \quad (2.26)$$

where we have used the fact that $A_1(i, 1) = A(i, 1)$. Dividing the above equations by $E(T)$

and noting that $p(0) = 1/(\lambda E T)$, this equation reduces to

$$p(i) = \lambda p_0 A(i, 1) + \sum_{j=1}^i \lambda A(i, j) p(j), \quad i \geq 1 \quad (2.27)$$

which is exactly the same as the equation (4.28) in Tijms(1986).

The recursion equations developed for the $M/G_n/1$ queue in this chapter (i.e. (2.21) and (2.22)) will be used in Chapter 3 (Section 3.6) as an approximation to a semi-open queueing network with Poisson arrivals to calculate the total average number in the system. In Table 2.1, we demonstrate the accuracy of our recursion scheme, by comparing the expected number in the system of a single server $M/G/1$ queue calculated via (2.21) and (2.22) and to that of exact value obtained by Pollaczek-Khinchine formula when the service time is a mixture of two exponential random variables with parameters μ_1, μ_2 whose density is given by

$$\frac{\mu_1 \mu_2}{\mu_2 - \mu_1} e^{-\mu_1 t} + \frac{\mu_1 \mu_2}{\mu_1 - \mu_2} e^{-\mu_2 t}$$

In Table 2.1, $\mu_1 = 5, \mu_2 = 2$, so that the mean service time is $(1/\mu_1 + 1/\mu_2) = (1/5 + 1/2) = 0.7$. The expected number(EN) in the system calculated using the truncated state probabilities match very well with the exact value. The state probabilities were calculated using a preassigned truncation number which depended on the utilization factor ρ . As a rule of thumb, we set this value to 30 when $\rho \leq 0.5$ and to 50 when $\rho > 0.5$.

λ	ρ	$M/G/1$ (2.21)&(2.22) (EN)	$M/G/1$ P-K (EN)
0.25	0.175	0.2045	0.2045
0.50	0.350	0.4981	0.4981
0.75	0.525	0.9864	0.9864
1.00	0.700	1.9984	1.9984
1.25	0.875	5.7269	5.7377

Comparison of Average Number in the System

Table 2.1

CHAPTER 3

APPROXIMATIONS TO SEMI-OPEN QUEUEING NETWORKS

One of the important queueing networks which has received very little attention among researchers is the semi-open queueing network. Semi-open queueing networks, as we shall explain in detail in Section 3.2, combine features of both the open and closed queueing networks in the sense that they model the arrival process to the network as well as resource constraints such as the maximum number of jobs which can reside in the network simultaneously.

In contrast to open and closed queueing networks, even the simplest semi-open queueing networks (simple in terms of exponential service times, Poisson arrivals etc) do not have the elegant product form solution. In order to obtain any decent estimates of some of the performance measures of this semi-open queueing network, simple and reasonably accurate approximate procedures are necessary. In this chapter, we propose two approximate procedures namely, a $D/M_n/1$ approximation to the semi-open queueing network with deterministic input and an $M/G_n/1$ approximation to a semi-open queueing network with Poisson input. We use the stationary distributions obtained in Chapters 1 and 2 for the $D^*/M_n/1$ and the $M/G_n/1$ queues to calculate some key performance measures.

This chapter is organized as follows. After a brief introduction (Section 3.1) to the open and closed queueing network models, we explain the semi-open queueing network model in detail in Section 3.2. Section 3.3 deals with the well known equivalent single server queue approximation to the semi-open network. The frequently used $M/M_n/1$ single server queue approximation is discussed in the Section 3.4. In this section, we also

provide some of the key performance measures of the $M/M_n/1$ queue which will be used in the later sections for comparison purposes. In Section 3.5, we investigate the effectiveness of the $D/M_n/1$ approximation for the semi-open queueing network with deterministic input.

In Section 3.6, the $M/G_n/1$ approximation to the semi-open network with Poisson input requires the explicit knowledge of the service time distribution. In that section, we first provide several methods of calculating the service time distributions for the $M/G_n/1$ queue, and then investigate the effectiveness of this approximation against the $M/M_n/1$ approximation and the simulation results.

3.1 Open and Closed Queueing Network Models

In this section, we briefly indicate the steady state analysis of the open and closed models due to Jackson (1963) and Gordon and Newell (1967), which lead to product form solutions.

Firstly, we consider an open queueing network with M nodes each having a single server. Job flows through the network in a probabilistic manner. When a job completes its service at the i^{th} node, it proceeds to the j^{th} node with probability p_{ij} . Jobs arrive from an external node to the network at a rate λ and upon entering the network, go to the i^{th} node with probability p_{0i} . On leaving the i^{th} node, a job will leave the network with probability $1 - \sum_{j=1}^M p_{ij} = p_{i0}$. We assume that the arrival process is Poisson with parameter λ . The average total arrival rate λ_i to the i^{th} node in the network satisfy the following traffic rate equations:

$$\lambda_i = \lambda p_{0i} + \sum_{j=1}^M \lambda_j p_{ji}, \quad i = 1, 2, \dots, M. \quad (3.1)$$

For this model, with exponentially distributed service times, FCFS service discipline and

unlimited waiting space at each node, Jackson(1963) showed that the nodes of the network can be decomposed and analysed separately as $M/M/1$ queues, i.e. if we denote $P(n_1, n_2, \dots, n_M)$ as the joint probability that there are n_i jobs at node i in the steady state, then this probability is given by

$$P\{n_1, n_2, \dots, n_M\} = \prod_{i=1}^M P_i(n_i) \quad (3.2)$$

where the marginal probabilities $P_i(n_i)$, $i = 1, 2, \dots, M$ are exactly the same as that of the queue length probability of an $M/M/1$ queue with traffic intensity ρ_i , i.e.

$$P_i(n_i) = \rho_i^{n_i} (1 - \rho_i) \quad (3.3)$$

where $\rho_i = \lambda_i / \mu_i$ (μ_i is the service rate at node i and $\lambda_i \leq \mu_i$, $i = 1, 2, \dots, M$). This celebrated product form solution holds because in the steady state, each queue in this network acts like an $M/M/1$ queue even though the effective arrival process to any node is not at all Poisson. Moreover, this product form solution holds when the single server queues are replaced by multi-server queues and also when the external arrival and the service process at each queue are state dependent. See Kelly(1979) for more general queueing networks which have this product form solution.

In a closed model, there are no arrivals and no departures. Let N be the total number of jobs which circulate in the network and let p'_{ij} , $i, j = 1, 2, \dots, M$ be the transition probabilities for this closed network. Instead of specifying these transition probabilities, we modify the transition probabilities p_{ij} 's for the open model mentioned above to obtain the p'_{ij} 's for the closed model by eliminating the external arrivals and departures to and from the network as follows:

$$p'_{ij} = p_{ij} + p_{0j} \left(1 - \sum_{k=1}^M p_{ik} \right), \quad i, j = 1, 2, \dots, M. \quad (3.4)$$

Note that p_{0j} is the probability that an arriving job on its first visit to the network enters the node j and $(1 - \sum_{k=1}^M p_{ik}) = p_{i0}$ is the probability that a job leaves the network after completing its service at node i . For this closed network, with the transition probabilities p'_{ij} given above, the traffic rate equations are given by the following M linear equations:

$$\eta_i = \sum_{j=1}^M \eta_j p'_{ji}, \quad i = 1, 2, \dots, M. \quad (3.5)$$

Unlike an open model, the above set of traffic rate equations have unique solution only up to a multiplicative constant. Let $X_i = \eta_i / \mu_i$, then the joint probability that there are n_i jobs in the i^{th} queue is given by

$$P\{n_1, n_2, \dots, n_M\} = \frac{1}{G(N)} \prod_{i=1}^M X_i^{n_i} \quad (3.6)$$

where $G(N)$ is the normalisation constant so that the probabilities sum to one over the set of M -tuples such that $n_1 + n_2 + \dots + n_M = N$. Once the constant $G(N)$ is known, various quantities such as utilisation, throughput etc. can be easily calculated. For example, the utilisation of the i^{th} node is given by

$$u_i(N) = X_i \frac{G(N-1)}{G(N)} \quad (3.7)$$

and the throughput $Th_i(N)$ is

$$\begin{aligned} Th_i(N) &= \mu_i p_{i0} X_i \frac{G(N-1)}{G(N)} \\ &= \eta_i p_{i0} \frac{G(N-1)}{G(N)} \end{aligned} \quad (3.8)$$

We note that in the closed models, unlike open models, the nodes can not be decomposed and analysed separately; this is due to the fact that the total number of jobs in the network is kept at constant level N all the time. Once again the product form solution can be shown to hold for more general networks (see Kelly(1979)).

3.2 Semi-Open Queueing Network Model

In this section, after a brief introduction to the semi-open queueing network model, we explain the difficulties involved in obtaining an analytic solution and stress the need for approximation schemes in order to obtain estimates of the performance measures for this network. Once again, we consider the open model described in Section 3.1 with the following modifications. For the sake of clarity, we repeat some of the descriptions mentioned in that section. The queueing network consists of M single server nodes; service times at these nodes are exponential with rates μ_i , $i = 1, 2, \dots, M$. We note that this assumption is not crucial for the equivalent single server queue approximation to be introduced later in Section 3.3. Let the routing matrix be given by the transition probabilities p_{ij} , $i, j = 1, 2, \dots, M$ such that $\sum_{j=1}^M p_{ij} \leq 1$ and let $p_{i0} = 1 - \sum_{j=1}^M p_{ij}$ be the probability that a job leaves the network after completing its service at node i . Also let p_{0i} , $i = 1, 2, \dots, M$ be the probability that an arriving job joins the node i on its first visit to the network. We will assume that the arrival process to be arbitrary but renewal. In this network, unlike the open model, an arriving job can enter the network only when the total number of jobs in the network is less than a prespecified quantity, say N . Jobs arriving at the network when the network is full (i.e. when there are N jobs in the network) are blocked and wait outside the network and enter the network as soon as possible (this can happen only when a job leaves the network after its service completion at one of the nodes). We call the waiting area outside the network either as outside queue or a dispatch queue (see Fig.3.1).

By limiting the number of jobs which can reside simultaneously in the network to N , and forcing the blocked jobs to wait in outside queue, this network can model resource constraints such as multi-programming, maximum number of pallets, number of tokens etc

Figure 3.1

Semi-Open Queueing Network

as well as the arrival pattern of the jobs to the network. We could also think of this outside queue as a dispatcher who monitors the status of the network in terms of the number jobs in the network and then releases the job to the network. Since the total number of jobs in the network can not exceed N at any time, this network fluctuates between open and closed model. Note that as long as the total number of jobs in the network is less than N , the outside queue will be empty and arriving jobs enter freely the network and hence the system behaves like an open model. When the total number of jobs in the system (network + outside queue) is greater than or equal to N , the network has always N jobs in it and hence it acts like a closed model. It is because of this feature that we call this network the semi-open model.

We point out that Dubois(1983) and Kapur and Kumar(1984) studied a model which they also called semi-open model. They assumed the following input pattern:

$$\lambda_n = \begin{cases} \lambda & \text{if } n < N \\ 0 & \text{otherwise,} \end{cases}$$

i.e., an arriving job which finds the network full (N jobs in the network) is assumed to be lost. This network with the arrival pattern mentioned above can be reduced directly as a

closed queueing network with $(M + 1)$ nodes and N jobs where the $(M + 1)^{th}$ is a state dependent single server queue with service rates λ_n given above. Note that despite the names being the same, the semi-open model considered in this chapter is quite different from that discussed by the above mentioned authors.

Our aim is to analyse this semi-open model under the additional assumption that each node has a waiting space of at least N so that there is no blocking. It is easy to show (Konheim and Reiser(1978)) that with all the simple assumptions such as Poisson arrivals, exponential service times, unlimited waiting space etc, this semi-open queueing network does not have a product form solution. Though it is possible to consider the multi-dimensional queue length process which is Markov when the arrival process is Poisson and then derive the steady state performance measures, the analysis become very tedious and cumbersome even for small networks. There is only one such analysis available in the literature and it is due to Konheim and Reiser(1978) who consider a three single server node central sever semi-open network with exponential service times and Poisson arrivals. In order to obtain simple and reasonable estimates of the performance measures for the general semi-open network, we seek approximation procedures. This approximation is discussed in the next section.

3.3 Equivalent Single Server Queue Approximation

The gist of this approximation is to view the semi-open queueing network model as two subsystems namely, one containing the arrival process and the outside queue and the other containing the M nodes of this network. Once we have identified the two subsystems, the approximation calls for the latter subsystem to be replaced by a single equivalent server. The schematic representation of this reduction is given in Fig 3.2.

The intuitive reasoning behind this approximation is that between successive arrivals of

Figure 3.2

Equivalent Single Server Queue

jobs to the semi-open network, the M nodes of the network acts like a closed model.

Having replaced the second subsystem by an equivalent server, one must determine the service time characteristics of this equivalent server. This is done by determining throughput $Th(n)$ of the associated closed queueing network with M nodes when there are n , $n = 1, 2, \dots, N$ jobs in the network. In other words, the service rates of the equivalent server are set equal to the throughput of the corresponding closed network, i.e., if we denote the service rates of the equivalent server by μ_n , then these service rates are given by

$$\mu_n = \begin{cases} Th(n) & n = 1, 2, \dots, N - 1 \\ Th(N) & n \geq N \end{cases} \quad (3.9)$$

Note that $Th(N)$ is the maximum throughput from the closed queueing network and hence the maximum service rate possible for this equivalent server. Since the throughput is a non-decreasing function of the number of jobs in the closed queueing network, the service rates of the equivalent server given in (3.9) are also non-decreasing.

This equivalent single server queue approximation for the semi-open model can be carried out via the following three steps given below.

- i Identify the two subsystems namely, the arrival process and the outside queue and

the M nodes of the network.

- ii Form the associated closed queueing network, by modifying the original transition probabilities p_{ij} by the following formula:

$$p'_{ij} = p_{ij} + p_{0j} \left(1 - \sum_{k=1}^M p_{ik} \right), \quad i, j = 1, 2, \dots, M.$$

Calculate the throughput $Th(n)$'s of this associated closed queueing network with M nodes when there are $n = 1, 2, \dots, N$ jobs in the network.

- iii Form the equivalent single server queue with service rates set equal to the throughput and analyse this state dependent server queue.

The performance measures such as the total number of jobs in the system (network + outside queue) and the average waiting time in the system obtained from this equivalent server would then correspond to the same measures of the semi-open model. Unfortunately, we can not obtain performance measures for the individual nodes in the semi-open network from the above approximation.

We note that from the above approximation, we will only have obtained the first moment of the state dependent service times. The analysis of the state dependent server queue mentioned in step (iii), depends on the complete characterisation of the service times. From the definition of the throughput, $Th(n)$ represents the rate at which jobs leave the closed queueing network when there are n jobs in the network and hence represents the average inter-departure times or exit times from this network. Therefore, the characterisation of the service times of the equivalent server depends on the characterisation of the departure process from the closed queueing network.

In Sections 3.4, 3.5.1 and 3.5.2, we will assume that the exit times are exponentially distributed and hence the service times of the equivalent server are exponentially distributed with rate $Th(n)$ i.e., a $./M_n/1$ approximation to the semi-open model. With

this assumption, as we shall see in these sections, the throughput of the associated closed queueing network alone is sufficient to analyse this state dependent queue given in step (iii). In Sections 3.5.3 and 3.6, we will explore the possibilities of using a non-exponential equivalent server for the semi-open model which in turn requires the general distribution of the exit times from the associated closed queueing network. The knowledge of the throughput alone is not sufficient. In these sections, we will provide approximate procedures to determine the general service times of the equivalent server. In particular, in Section 3.5.3, we will propose a $D/G/1$ approximation to the semi-open queueing network with deterministic input and in Section 3.6, we will propose several $M/G_n/1$ approximations to semi-open queueing networks with Poisson input.

3.4 $M/M_n/1$ Approximation

In this section, we consider a semi-open model with exponential service times and Poisson input. The equivalent single server approximation discussed in the last section, reduces to a single server queue with Poisson input and state dependent service times with rates given in (3.9). As we mentioned in Section 3.3, in order to analyse this state dependent single server queue, further information about the service times is necessary. Now, if we assume that the service times of the state dependent server are exponentially distributed with the service rates given in (3.9), then the equivalent single server queue is simply an $M/M_n/1$ queue.

This $M/M_n/1$ approximation was first reported by Avi-Itzhak and Heyman(1973) and later by Shanthikumar and Sargent(1981) and Buzacott and Shanthikumar(1985) (who calls this approach the Controlled Arrival Single server Queue (CASQ)) for the semi-open model considered in this section.

By assuming that the service times are exponentially distributed, we are indirectly

assuming that the interdeparture times from the closed queueing network are exponentially distributed, but this is only an approximation. We note however when $N = \infty$, the semi-open model considered here reduces to the open model discussed in Section 3.1. In this case, the $M/M_n/1$ equivalent single server queue with service rates $\mu_n = Th(n)$ for $n = 1, 2, \dots$ (note that there is no maximum service rate) is exactly equivalent to the open model, i.e., the stationary distribution of the number of jobs in this queue is exactly the same as the stationary distribution of the total number of jobs in the open network. See Shanthikumar and Sargent(1981) for a proof of this exact equivalence. This exact equivalence of the $M/M_n/1$ single server queue to the semi-open model with Poisson input when $N = \infty$, is the motivation for considering a $M/M_n/1$ equivalent single server queue with service rates specified in (3.9) when $N < \infty$.

This single server queue with the specified arrival (λ) and service rates (3.9) are very easy to solve, and we now provide results for some of the performance characteristics. This $M/M_n/1$ approximation will be used as a benchmark against an $M/G_n/1$ approximation in Section 3.6.4. The flow balance equations for this single server queue are:

$$\lambda P_{n-1} = \begin{cases} \mu_n P_n, & \text{if } n = 1, 2, \dots, N-1; \\ \mu_N P_n & \text{otherwise;} \end{cases} \quad (3.10)$$

where P_n is the probability that there are n jobs in the single server queue. Now using the normalisation condition $\sum_{n=0}^{\infty} P_n = 1$, we obtain

$$P_n = \begin{cases} \frac{\lambda^n}{\mu(n)} P_0, & \text{if } n = 1, 2, \dots, N-1; \\ \frac{\lambda^N}{\mu(N)} \rho_N^{n-N} P_0, & \text{otherwise.} \end{cases} \quad (3.11)$$

where

$$P_0 = \left[1 + \sum_{n=1}^{N-1} \frac{\lambda^n}{\mu(n)} + \frac{\lambda^N}{\mu(N)} \left(\frac{1}{1 - \rho_N} \right) \right]^{-1} \quad (3.12)$$

and $\mu(n) = \prod_{i=1}^n \mu_i$, $n = 1, 2, \dots, N$ and $\rho_N = \lambda/\mu_N$. From these quantities, all the characteristics of interest can be derived easily. Let $E(N_0)$, $E(N_i)$ and $E(N_s)$ be the

expected number of jobs in the outside queue, inside the network and in the whole system respectively. Then

$$E(N_o) = \frac{\lambda^N}{\mu(N)} \frac{\rho_N}{(1 - \rho_N)^2} P_0, \quad (3.13)$$

$$E(N_i) = \left\{ \sum_{n=1}^{N-1} n \frac{\lambda^n}{\mu(n)} + \frac{\lambda^N}{\mu(N)} \frac{N}{(1 - \rho_N)} \right\} P_0, \quad (3.14)$$

and

$$E(N_s) = E(N_i) + E(N_o). \quad (3.15)$$

Once the μ_n , $n = 1, 2, \dots, N$ are calculated using (1.8), the quantities given in (3.13) to (3.15) are easy to compute.

3.5 Approximations to Semi-Open Queueing Network with Deterministic Input

In this section, we consider a semi-open model (described in Section 3.2) with exponential service times and deterministic input. Jobs are scheduled to arrive at the network at discrete time points. In order to analyse this model, we use the equivalent single server approximation described in Section 3.3. We also present two other approximations, which are variants of the procedure given in that section.

3.5.1 D/M_n/1 Approximation

Following the steps (i) to (iii) given in Section 3.3, the equivalent single server queue for this semi-open model is a state dependent single server queue with deterministic input. Similar to the M/M_n/1 approximation discussed in Section 3.4 for Poisson arrivals, we propose a D/M_n/1 approximation for the semi-open model with deterministic input. The equation (3.9) describes the service rates for this queue. We use the results obtained in Chapter 1 (when $s = 1$ and $p = 1$, the D^{*}/M_n/1 queue reduces to a D/M_n/1 queue) to calculate the average number of jobs in the system.

3.5.2 D/M/1 Approximation

Instead of considering the state dependent equivalent single server $D/M_n/1$ queue as an approximation to the semi-open model, we simply consider a $D/M/1$ queue approximation. Here the service rate is set equal to the maximum throughput $Th(N)$ of the associated closed queueing network, i.e., $\mu_n = Th(N)$ for all $n = 1, 2, \dots$

3.5.3 D/G/1 Approximation

The approximation to be discussed here is similar to the one introduced in Section 3.5.2, i.e., not a state dependent server. The motivation for considering a general distribution for the equivalent server is as follows: When $N = 1$, even though the equivalent single server queue approximation suggests a $D/M/1$ queue, the semi-open model is actually a single server queue with a mixture of exponential service times. In this sub section, we will only consider a two node serial semi-open network.

$N = 1$ Case: The two node serial semi-open network with exponential service times is exactly a $D/E_2/1$ queue (assuming that the service times at the two nodes are identical) where E_2 represents the Erlang random variable with two stages. When the service times at the nodes are different, this network reduces to a $D/E_2^*/1$ queue where E_2^* represents the generalised Erlang random variable. Therefore, the $D/M/1$ approximation is not correct. We will use the $D/E_2/1$ queue (which is exact) for the two node serial semi-open network when $N = 1$ to compare the performance of the $D/M/1$ approximation suggested by the equivalent single server approach.

The expected number in the system for this queue is calculated using the Kramer and Langenbach-Belz(1976) approximation for the $D/G/1$ queue which is given by

$$E(L) \approx \rho + \frac{\rho^2 C_s^2}{2(1-\rho)} \exp\left(\frac{-2(1-\rho)}{3\rho C_s^2}\right) \quad (2.21)$$

where C_s^2 is the squared service coefficient of variation.

When $N = \infty$, this network becomes open network with deterministic input and exponential service times which does not have a product form solution. For the two node serial semi-open network when $1 < N < \infty$, we will propose a $D/G/1$ approximation which is based on Buzacott and Shanthikumar(1985).

$1 < N < \infty$ Case: We observe that as long as the second server is busy, then the interdeparture times from the network will be exponentially distributed with rate 1 (assuming that the service rates at the two nodes are identical at rate 1), otherwise, the interdeparture times will be distributed according to an Erlang random variable with two stages. Let p be the probability that the second server is busy in the network, then the interdeparture times from the network are a mixture of exponential and Erlang random variables with expectation $(2 - p)$ and second moment $(6 - 4p)$.

Now we use this mixture as the correct random variable representing the service times of the equivalent single server queue instead of just an exponential random variable. The values of p can be calculated by equating the mean of this mixture to the inverse of the maximum throughput $Th(N)$ obtained from the closed queueing network which in this case is a cyclic network, i.e.

$$(2 - p) = \frac{1}{Th(N)}$$

Note that from the definition of the throughput, the $Th(N)$ is the rate at which jobs leave the network when there are N jobs in the network. Once again, the expected number in the system for this $D/G/1$ queue is calculated using equation (2.21). We point out that the argument presented here can be easily extended to the case when the service times at the two nodes are different. See Sections 3.6.1 to 3.6.3 for a variation of the argument presented in this section.

3.5.4 Numerical Results

In this section, we provide numerical results which compare the performance of the various approximations suggested in the previous sections against the results obtained by simulation. The simulation experiments were carried out using the QNAP2 simulation programme. In all the tables provided here, we only consider the average total number of jobs in the semi-open model. The simulation results are reported with 95% confidence interval estimates. We consider two types of semi-open networks namely, two node serial semi-open network with equal and unequal exponential service times, and four node symmetric semi-open network with equal exponential service times.

Tables 3.1 through 3.5 refer to the two node semi-open serial network with identical exponential service times. In Table 3.1, we compare, for various arrival rates, the $D/M/1$ and $D/E_2/1$ approximations against the simulation results when $N = 1$. As we can see from this table, the $D/E_2/1$ approximation outperforms the $D/M/1$ approximation. Also the $D/M/1$ approximation always over estimate the average total number in the semi-open model.

When $1 < N < \infty$, we use $D/M_n/1$, $D/M/1$ and $D/G/1$ approximations for this network. Tables 3.2 to 3.5 refer to this case. In these tables, we compare the effectiveness of these approximations by varying the input rate from 0.5 to 0.8 by increments of 0.1. Each tables provide the average total number in the network for various values of N . As we observe from Tables 3.2 through 3.5, the $D/M_n/1$ approximation always overestimates; the $D/M/1$ and the $D/G/1$ approximations perform poorly.

The poor performance of the $D/G/1$ approximation is worth a few comments. Buza-cott and Shanthikumar(1985) reported that for the two node semi-open serial network with Poisson input and exponential service times, the $M/G/1$ approximation outperformed the

$M/M/1$ approximation. As we saw in Table 3.1, the $D/E_2/1$ approximation indeed performs better than the $D/M/1$ approximation when $N = 1$. But when $1 < N < \infty$, the $D/M/1$ approximation always performs poorly and hence the $D/G/1$ approximation also performs poorly. This is expected since the $D/G/1$ approximation uses a service time random variable whose coefficient of variation is less than one when compared to the exponential service time which has unit coefficient of variation. Based on this observation, we might suggest a $D/G/1$ approximation with high coefficient variation but, it is yet to be determined how to find a service time random variable with this property.

In Table 3.6, we report the results for this network by keeping N fixed and varying the service rates at the two nodes. Once again, the $D/M_n/1$ approximation over estimates and the $D/M/1$ and $D/G/1$ approximation perform poorly.

Tables 3.7 to 3.9 refer to the four node symmetric semi-open network. In this case, we have not been able to carry out the arguments as in the two node serial case to approximately characterise the exit times from the associated closed queueing network; we only know the throughput. Therefore, for various arrival rates, we only compare the $D/M/1$ and $D/M_n/1$ approximations against simulation results. For this symmetric network, all the approximations perform poorly (under estimation).

In general, based on the limited numerical experiments, we can conclude that for the two node semi-open serial networks, the $D/M_n/1$ approximation performs relatively better than the other approximations except for the case when $N = 1$. For general semi-open networks (such as symmetric networks), the approximations proposed here are not good. In our opinion, there are two sources of error namely, 1. the approach of reducing the semi-open model to a single server queue, 2. characterisation of the service times (such as exponential service times) of the equivalent server. In the next section, we explore the

possibility of improving the characterisation of the equivalent server service times (general instead of exponential) for a simple semi-open model with Poisson input and exponential service times.

Average Total Number in the System			
ρ	SIMUL	$D/M/1$	$D/E_2/1, \rho = 0.4$
0.666	$0.8531 \pm .0392$	1.1685	0.8370
0.5	$0.5320 \pm .0146$	0.628	0.5329
0.4	$0.4062 \pm .0097$	0.448	0.4090

TABLE 3.1
Two Node Serial Semi-open Network with $N=1$
 $\mu_1 = 1, \mu_2 = 1$

Average Total Number in the System				
N	SIMUL	$D/M_n/1$	$D/M/1$	$D/G/1$
2	$1.539 \pm .0840$	1.9381	1.6500	1.4074
3	$1.471 \pm .0641$	1.6144	1.1439	1.0652
4	$1.414 \pm .0626$	1.5560	0.9735	0.9352
5	$1.423 \pm .0553$	1.5473	0.8888	0.8655
6	$1.422 \pm .0466$	1.5445	0.8364	0.8218

TABLE 3.2
Two Node Serial Semi-open Network with $\lambda = 0.5$ (N varies)
 $\mu_1 = 1, \mu_2 = 1$

Average Total Number in the System				
N	SIMUL	$D/M_n/1$	$D/M/1$	$D/G/1$
2	$4.001 \pm .6697$	4.8177	4.5733	3.7622
3	$2.423 \pm .2725$	2.7335	2.1539	1.9572
4	$2.177 \pm .1327$	2.4122	1.6504	1.5629
5	$2.168 \pm .1087$	2.3524	1.4340	1.3846
6	$2.025 \pm .0799$	2.3000	1.3134	1.2817

TABLE 3.3
Two Node Serial Semi-open Network with $\lambda = 0.6$ (N varies)
 $\mu_1 = 1, \mu_2 = 1$

Average Total Number in the System				
N	SIMUL	$D/M_n/1$	$D/M/1$	$D/G/1$
3	6.465 ± 1.2380	5.4266	6.4655	6.3472
4	$4.305 \pm .8312$	4.5082	3.6430	3.4150
5	$3.665 \pm .3542$	3.8686	2.7810	2.6610
6	$3.361 \pm .3239$	3.4895	2.3828	2.3099
7	$3.021 \pm .1781$	3.1257	2.1539	2.1050

TABLE 3.4
Two Node Serial Semi-open Network with $\lambda = 0.7$ (N varies)
 $\mu_1 = 1, \mu_2 = 1$

Average Total Number in the System				
N	SIMUL	$D/M_n/1$	$D/M/1$	$D/G/1$
9	$5.948 \pm .7297$	5.1687	4.1196	4.0893
10	6.012 ± 1.1240	5.3525	3.8037	3.7729
11	$5.864 \pm .6599$	5.2145	3.5732	3.5454
12	$5.312 \pm .4604$	4.9049	3.3993	3.3400

TABLE 3.5

Two Node Serial Semi-open Network with $\lambda = 0.8$ (N varies)

$$\mu_1 = 1, \mu_2 = 1$$

Average Total Number in the System				
Service	SIMUL	$D/M_n/1$	$D/M/1$	$D/G/1$
0.2 $\rightarrow$ 0.2	$0.4117 \pm .0044$	0.4200	0.2748	0.2722
0.4 $\rightarrow$ 0.2	$0.6771 \pm .0122$	0.7065	0.4954	0.4864
0.4 $\rightarrow$ 0.4	$0.9663 \pm .0195$	1.0558	0.7026	0.6696
0.6 $\rightarrow$ 0.2	$1.1480 \pm .0460$	1.2005	0.9386	0.9235
0.6 $\rightarrow$ 0.4	$1.5080 \pm .0569$	1.6801	1.2276	1.1557
0.6 $\rightarrow$ 0.6	$2.2920 \pm .1145$	2.7300	2.1541	1.9571
0.8 $\rightarrow$ 0.2	$1.4310 \pm .1902$	2.8505	2.2866	2.2548
0.4 $\rightarrow$ 0.8	$3.2830 \pm .2909$	3.5155	3.1457	2.9736
0.5 $\rightarrow$ 0.8	$4.6320 \pm .6615$	4.4680	4.4395	4.1423
0.9 $\rightarrow$ 0.2	$5.5650 \pm .9932$	4.7131	4.6165	4.7043

TABLE 3.6

Two Node Serial Semi-open Network with $N = 3$ (μ_1, μ_2 varies)

Average Total Number in the System			
N	SIMUL	$D/M_n/1$	$D/M/1$
4	6.062 ± 3.227	4.8867	3.6449
5	4.330 ± 0.478	3.8980	2.1541
6	3.549 ± 0.233	3.6490	1.6502
7	3.972 ± 0.227	3.2489	1.3978
8	3.876 ± 0.246	3.5361	1.2455

TABLE 3.7

Four Node Symmetric Semi-open Network with $\lambda = 0.5$

$$\mu_i = 1, i = 1, 2, 3, 4$$

Average Total Number in the System			
N	SIMUL	$D/M_n/1$	$D/M/1$
6	7.984 ± 4.213	6.5157	4.5733
7	6.889 ± 0.976	5.7615	3.1532
6	5.959 ± 0.828	4.5886	2.5115
11	5.567 ± 0.346	4.9726	1.7669
12	5.735 ± 0.528	5.2364	1.6504

TABLE 3.8

Four Node Symmetric Semi-open Network with $\lambda = 0.6$

$$\mu_i = 1, i = 1, 2, 3, 4$$

Average Total Number in the System			
N	SIMUL	$D/M_n/1$	$D/M/1$
9	14.200 ± 3.041	8.8195	6.4655
10	11.440 ± 2.021	8.6112	5.0562
11	10.890 ± 3.306	8.5523	4.1971
12	10.220 ± 3.686	8.3426	3.6429
13	9.146 ± 1.146	8.3241	3.2629

TABLE 3.9
Two Node Symmetric Semi-open Network with $\lambda = 0.7$
 $\mu_i = 1, i = 1, 2, 3, 4$

3.6 Approximations to Semi-Open Queueing Networks with Poisson Input

In this section, we consider a semi-open model with exponential service times and Poisson input. In order to analyse this model, we use the equivalent single server queue approximation described in Section 3.3. We recall that in order to apply this equivalent single server approach to semi-open networks, we first have to calculate the throughput of the associated closed queueing network for various n , where n is the number of jobs in the closed queueing network. Note that n takes on values from 1 to N , where N is the maximum number of jobs allowed in the semi-open queueing network. Let $Th(n)$ denote the throughput of the closed queueing network whenever there are n jobs in it. This throughput $Th(n)$ is then used as the service rate of the state dependent server in the equivalent single server queue.

In the cases of $D/M_n/1$ and $M/M_n/1$ equivalent single server queues, the knowledge of the $Th(n)$'s and hence of the service rates of the state dependent server were sufficient (due to the assumption of exponential service times, only the mean of the service times appeared in the formulae, see the equations in Section 1.2 and 3.4') to calculate the state probabilities and hence the expected number in the system of these queues. As we mentioned in Section 3.4, $Th(n)$, the throughput is by definition, the rate at which jobs leave the closed queueing network when there are n jobs in the network. Therefore, by assuming that the service times of the equivalent server are exponentially distributed with rate $Th(n)$, we are assuming that the exit times from the closed queueing network, when there are n jobs in it, are exponentially distributed with rate $Th(n)$. But this assumption is only an approximation. To see this, consider the semi-open model with $N = 1$. In this case, the $M/M_n/1$ equivalent single server approximation reduces to an $M/M/1$ queue. But a typical job in this semi-open network, would go through many

exponential service stages before it leaves the network. For example, consider a two node serial semi-open network with exponential service times. When $N = 1$, this semi-open network is actually a $M/E_2/1$ queue (see Section 3.5.3 for a similar argument), whereas the equivalent single server approximation reduces to an $M/M/1$ queue. This observation suggests that the assumption of exponential exit times from the closed network and hence exponential service times for the equivalent server is not correct.

In order to accommodate the idea that the exit times in the semi-open network are a mixture of exponential distributions, we let the exit times follow a general distribution and hence the state dependent service time of the equivalent server will be a random variable with general distribution. This in turn requires that we consider a $M/G_n/1$ approximation for the semi-open network. However, the analysis of $M/G_n/1$ queue, requires explicit representation (in terms of distribution or density function) of the state dependent service times in order to calculate the stationary distributions for the number in the system, i.e. the recursion relationships in Section 2.3 involves explicitly the distribution function of the service times (in terms of F_k).

In general, the characterisation of the state dependent general service times of the equivalent server and hence the exit time distribution from the closed queueing network for general semi-open queueing networks are very difficult. In this section, however we will first propose various approximate procedures to characterise the service times of the equivalent server for some special semi-open networks and then give examples to illustrate the use of the $M/G_n/1$ queue (developed in Chapter 2) as an approximation these semi-open queueing networks with Poisson input. We basically treat three types of network, namely

(1) Serial Semi-Open Network

(2) Feedback Semi-Open Network

(3) Central Server Semi-Open Network

3.6.1 Approximate Description of Exit times from Two Node Serial Semi-Open Network

This semi-open network consists of two servers which are connected in series. The service times at the two server are exponentially distributed with parameter μ_1 and μ_2 . We denote the random variables representing their service times as $E(1, \mu_1)$ and $E(1, \mu_2)$ respectively at the two servers. The notation $E(n, \mu)$ stands for the n stage Erlang random variable with parameter μ , whose mean is n/μ and the squared coefficient of variation is $1/n$. As usual, let $Th(n)$ be the throughput of the associated cyclic queue with two nodes when there are n jobs in the network. Note that since jobs can exit only from the second server in this semi-open network, this is our exit-entry node for the closed queueing network. See Fig. 3.3.

Figure 3.3

Two Node Serial Semi-Open Network

We now extend the arguments presented in Section 3.5.3 to characterize the service

times of the equivalent server. When there are n jobs in the semi-open network, let $p_1(n)$ be the probability that the second server is busy. As long as the second server is busy, the rate at which jobs leave the network is μ_2 and the exit times are distributed exponentially with parameter μ_2 or in other words, the exit times are distributed as $E(1, \mu_2)$. When the second server is not busy, with prob $(1 - p_1(n))$, the exit times are distributed as a sum of two exponential random variables with parameters μ_1 and μ_2 , i.e. as a $E(1, \mu_1) + E(1, \mu_2)$ random variable. Hence it is appropriate to approximate the exit time random variable (say X_n) and hence the service times of the equivalent server as

$$X_n = p_1(n)E(1, \mu_2) + (1 - p_1(n))\{E(1, \mu_1) + E(1, \mu_2)\}$$

with

$$E(X_n) = \frac{p_1(n)}{\mu_2} + (1 - p_1(n))\left\{\frac{1}{\mu_1} + \frac{1}{\mu_2}\right\}$$

and

$$E(X_n^2) = \frac{2p_1(n)}{\mu_2^2} + 2(1 - p_1(n))\left\{\frac{1}{\mu_1^2} + \frac{1}{\mu_2^2} + \frac{1}{\mu_1\mu_2}\right\}$$

Since $E(X_n) = 1/Th(n)$, $p_1(n)$, $n = 1, 2, \dots, N$ can be calculated easily as follows:

$$p_1(n) = \left\{\frac{1}{\mu_1} + \frac{1}{\mu_2} - \frac{1}{Th(n)}\right\}\mu_1$$

Once the values of $p_1(n)$ are known, the density of X_n can be calculated explicitly.

For example, when $\mu_1 = \mu_2 = \mu$, the exit time random variable X_n is distributed as a mixture of exponential with parameter μ and a two stage Erlang with parameters μ . When $\mu_1 \neq \mu_2$, X_n is distributed as a mixture of exponential with parameter μ_2 and generalized Erlang which is a sum of two exponential random variables with different parameters μ_1 and μ_2 . We note that the probabilities $p_1(n)$ are exactly the same as the utilization of the second node in a cyclic network when there are $(n - 1)$ jobs in the network. This is due to

the fact that in a closed queueing network with product form solutions, Poisson arrivals see time averages.

So far we have discussed the approximate calculation of the exit time distribution when the service times at the two nodes are exponentially distributed. This approach can be extended to the case where the service times are generally distributed. As an example, let us consider the case where the service times at the two nodes are $E(2, 2\mu_1)$ and $E(2, 2\mu_2)$ respectively. Then the exit times X_n is given by a mixture of $E(2, 2\mu_2)$ and $(E(2, 2\mu_1) + E(2, 2\mu_2))$ i.e.

$$X_n = p_1(n)E(2, 2\mu_2) + (1 - p_1(n))(E(2, 2\mu_1) + E(2, 2\mu_2))$$

when $\mu_1 = \mu_2 = \mu$, X_n simplifies to

$$\bar{X}_n = p_1(n)E(2, 2\mu) + (1 - p_1(n))E(4, 2\mu).$$

Once again $p_1(n)$'s can be calculated by equating $E(X_n)$ to $1/Th(n)$. In this case, $p_1(n)$'s are not equal to the utilization of the cyclic network with $(n-1)$ jobs in the network since the service times are not exponential.

3.6.2 Approximate Description of Exit times from Two Node Feedback Semi-Open Network

This semi-open network is the same as the one described previously except that, after completing service at server 2, a job leaves the network with probability θ and loops back to server one with prob $(1 - \theta)$. See Fig 3.4.

In order to calculate the exit time random variables, we argue as follows: Since the only difference between this network and the serial semi-open network is the looping

Figure 3.4

Two Node Feedback Semi-Open Network

phenomena, we first ignore this looping effect. Now the exit times X_n in the case of exponential servers for the serial semi-open network are

$$X_n = p_1(n)E(1, \mu_2) + (1 - p_1(n))(E(1, \mu_1) + E(1, \mu_2)), \quad n = 1, 2, \dots, N.$$

Since with probability θ jobs will exit from the original network, let X_n^θ be the exit times from this network. Also since the exit times from non-loop networks are split with probability θ and $1 - \theta$, and the exit times along the θ -branch correspond to the exit times from the feedback network, we use the splitting property of the renewal process to obtain

$$E(X_n^\theta) = \theta E(X_n)$$

$$C^2(X_n^\theta) = \theta C^2(X_n) + (1 - \theta)$$

where $C^2(X_n)$ and $C^2(X_n^\theta)$ denote the squared coefficient of variation of the non-loop and loop exit times respectively. So far, we have only obtained the two moments of the exit time random variables from the feed back network. In order to get a complete description of these random variables, we fit a distribution function matching these two moments. If the calculated $C^2(X_n^\theta)$ is less than 1, we either fit a shifted exponential or Erlang density,

otherwise, we fit a hyper exponential density. With this approach, for each n we have an explicit characterization of the exit time random variables.

3.6.3 Approximate Description of Exit times from Central Server Semi-Open Network.

In this network, there are three nodes. Node 1 is a central server and Nodes 2 and 3 are peripheral units. Jobs always enter the first node; after completing service at node 1, a job proceeds to node 2 and 3 with probabilities p_{12} and p_{13} respectively; with probability $p_{1,out}$ it leaves the network. After completing service at nodes 2 and 3, jobs always return to node 1. See Fig 3.5.

Figure 3.5

Central Server Semi-Open Network

Let μ_1 , μ_2 and μ_3 be the service rates at nodes 1, 2, and 3 respectively. In order to calculate the exit times, we first calculate the throughput $Th^*(n)$ of the associated closed

queueing network which is obtained by shorting the central server, i.e. we assume that the service rate at node 1 is infinite. This will give us the throughput of the subsystem consisting of nodes 2 and 3. In the original central server network, we replace nodes 2 and 3 by an equivalent server with state dependent service rate $Th^*(n)$. Now the transformed network consists of a central server, and a composite server (of nodes 2 and 3) with service rate $Th^*(n)$. This is a familiar network as section 3.6.2 with two nodes except that the composite server is a state dependent server. To use the approximate scheme developed in the Section 3.6.2, we replace the nodes 2 and 3 by a single server with service rate $Th^*(N)$ instead of replacing the nodes 2 and 3 by a state dependent server. For this network the approximate exit times X_n^* can be calculated easily as in Section 3.6.2. Once again, since jobs leave the original central server network with probability $p_{1,out}$, we obtain

$$E(X_n) = p_{1,out} E(X_n^*)$$

$$C^2(X_n) = p_{1,out} C^2(X_n^*) + (1 - p_{1,out})$$

As in Section 3.6.2, we fit distributions to the exit times by matching the two moments obtained.

3.6.4 Simulation Experiments

In this subsection, we investigate the effectiveness of the $M/G_n/1$ equivalent single server queue approximation for the three types of semi-open models introduced at the beginning of this section. The average number in the system for the $M/G_n/1$ is calculated via the recursion relationships given in Section 2.2 using the approximate descriptions of (the equivalent server) service times developed in sections 3.6.1 to 3.6.3. Note that F_k represents the distribution of the service times in Section 2.2. In order to compare the performance of this $M/G_n/1$ approximation, we simulated the semi-open queueing network using QNAP2. Statistics were collected using the replication method. In each replication,

the simulation time was taken long enough so that the steady state characteristics of this network were obtained. We also chopped off the data during the initial period so that no initial bias was introduced. In all the experiments, we report the average total number in the system with 95% confidence interval estimates. The tables provided in this section, only compare the average total number in the semi-open network obtained through simulation against various approximate procedures.

Apart from comparing the effectiveness of the $M/G_n/1$ approximation, we also compare three other approximations namely, the common $M/M_n/1$ approximation, $M/G/1$ and $M/M/1$ approximations. For the $M/M/1$ and $M/G/1$ approximations, we set the service rate of the equivalent server to the maximum throughput, $Th(N)$. In the case of $M/G/1$ approximation, the second moment of the equivalent service time is calculated along the same lines as the $D/G/1$ case which was given in the Section 3.5.3.

Tables 3.10 through 3.17 refer to the two node serial semi-open network. Here, we consider two cases namely, 1. the service times of both the servers are exponential, 2. the service times of both the servers are two stage Erlang. Tables 3.10 and 3.11 correspond to the two node serial semi-open network with equal exponential service rates. In Table 3.10, for a fixed arrival rate λ , the storage capacity N varies whereas in Table 3.3, N is fixed and the arrival rate λ varies. When λ is fixed and N varies, we would expect that as N becomes large relative to the arrival rate, this semi-open network would behave as an open model. Similarly, when N is fixed and λ is large relative to N , the semi-open network would have N jobs in the system most of the time and hence acts like a closed model. For $N = 10$, the expected number in the open model is 4.15 (calculated using the average number in the system for two $M/M/1$ queues in series) whereas the simulation result for the semi-open model is 4.144 (see Table 3.10). When N is fixed and λ varies,

the $M/G/1$ approximation performs poorly whereas when λ is fixed and N varies, the $M/G/1$ approximation is better than the others.

In Tables 3.12 and 3.13, the exponential service rates at the two servers are different. One server is a slow server relative to the other. In this case, we propose another approximation namely a slow server approximation, which is based on the fact that there will little or no queueing at the fast server and hence the semi-open model would behave as a single server queue with service time equal to that of the slow server. In this case, this is just an $M/M/1$ queue. As we observe from Tables 3.12 and 3.13, both the $M/M_n/1$ and the $M/G_n/1$ approximation over estimate the average total number in the system whereas the $M/M/1$, $M/G/1$ and the slow server approximations under estimates. Similarly, when the service times at the nodes are two stage Erlang (Tables 3.14 to 3.15), both the $M/M_n/1$ and the $M/G_n/1$ approximation over estimate the average total number in the system whereas the $M/M/1$, $M/G/1$ approximations under estimates. As expected (at least from the discussion for the case $N=1$ in Section 3.6), the increased information about the departure process from the closed queueing network and hence a detailed characterisation of the service time of the equivalent server does not seem to improve the quality of this $M/G_n/1$ approximation when $N > 1$.

Tables 3.18 and 3.19 refer to the feedback semi-open network with equal and unequal service rates respectively at the two nodes. Here, we keep N , λ fixed and vary θ which is the looping factor. When θ is small, this network should behave like a closed model and hence any approximation which is based on the maximum throughput alone should perform better. In Tables 3.18 and 3.19, when $\theta = 0.1$, all the approximations over estimate the average number in the semi-open model.

In Table 3.20, we report the comparisons for the central server semi-open model.

For the $M/G_n/1$ approximation, as we explained in Section 3.2.3, we fit either a shifted exponential or an Erlang distribution for the two moments of the interdeparture times. The approximation denoted by $M/G_n/1(1)$ is obtained through a shifted exponential fit and the $M/G_n/1(2)$ is obtained by an Erlang fit. Once again, the $M/M_n/1$ approximation is relatively better than all other approximations and the $M/G_n/1$ approximation is not performing as expected. As we mentioned in Section 3.5.4, the approximate procedures described in sections 3.6.1 to 3.6.3 to characterise the service times of the equivalent server are not very good. We expect this $M/G_n/1$ approximation to perform much better than the $M/M_n/1$ approximation when better procedures are developed to characterise the service times of the equivalent server.

N	SIMUL EN	$M M_n 1$ EN	$M G 1$ EN	$M G_n 1$ EN	$M M 1$ EN
3	8.373±.8232	9.4718	8.4937	9.5457	9.0000
4	5.421±.1623	6.0596	5.2177	6.4048	5.4000
5	4.695±.1706	5.0917	4.1681	5.5372	4.2630
6	4.381±.1750	4.6681	3.6464	5.1548	3.7050
7	4.319±.1935	4.4540	3.3343	4.9638	3.3750
8	4.295±.1706	4.3435	3.1388	4.8642	3.1560
9	4.237±.1244	4.2691	2.9775	4.7971	3.0000
10	4.144±.0970	4.2281	2.8659	4.7597	2.8830

Two Node Serial Semi-Open Network
with Equal Exponential Service Rates (N varies)

$$\lambda = 0.675, \mu_1, \mu_2 = 1$$

Table 3.10

λ	ρ	SIMUL EN	$M M_n 1$ EN	$M G 1$ EN	$M G_n 1$ EN	$M M 1$ EN
0.250	0.333	0.6717±.0162	0.6757	0.4895	0.7450	0.5
0.375	0.500	1.2130±.0258	1.2656	0.9677	1.4121	1.0
0.500	0.666	2.2320±.0456	2.3538	1.9167	2.5720	2.0
0.625	0.833	4.8200±.3016	5.4389	4.7395	5.6158	5.0
0.675	0.900	8.3730±.8232	9.4718	8.4937	9.5457	9.0

Two Node Serial Semi-Open Network
with Equal Exponential Service Rates (λ varies)

$$N = 3, \mu_1, \mu_2 = 1$$

Table 3.11

N	SIMUL EN	$M M_n 1$ EN	$M G 1$ EN	$M G_n 1$ EN	Slow Ser EN	$M M 1$ EN
3	10.233±.6311	10.5616	10.0072	10.6754	7.0	10.2188
4	8.196±.5615	8.4192	7.9255	8.706	7.0	7.9995
5	8.124±.5266	7.8429	7.3585	8.1758	7.0	7.3653
6	7.913±.4469	7.6512	7.1334	7.9977	7.0	7.1428
7	7.837±.5470	7.5803	7.0527	7.9321	7.0	7.0565
8	7.753±.2365	7.5542	7.0210	7.9087	7.0	7.0281
9	7.482±.1256	7.5461	7.0105	7.9012	7.0	7.0112
10	7.400±.3945	7.5423	7.0052	7.8977	7.0	7.0056

Two Node Serial Semi-Open Network
with Unequal Exponential Service Rates (N varies)

$$\lambda = 1.75, \mu_1 = 5.0, \mu_2 = 2.0$$

Table 3.12

λ	ρ	SIMUL EN	$M M_n 1$ EN	$M G 1$ EN	$M G_n 1$ EN	Slow Ser EN	$M M 1$ EN
0.50	0.260	0.447 ±.0080	0.4426	0.3496	0.4753	0.3333	0.3519
0.75	0.390	0.780 ±.0110	0.7827	0.6344	0.8443	0.6000	0.6405
1.00	0.521	1.271 ±.0144	1.2771	1.0720	1.3840	1.0000	1.0858
1.25	0.651	2.053 ±.0537	2.1033	1.8334	2.2597	1.6666	1.8629
1.50	0.781	3.882 ±.1089	3.8514	3.4953	4.0441	3.0000	3.5629
1.75	0.911	9.758 ±.7141	10.5616	10.0072	10.6754	7.0000	10.2339

Two Node Serial Semi-Open Network
with Unequal Exponential Service Rates (λ varies)

$$N=3, \mu_1 = 5.0, \mu_2 = 2.0$$

Table 3.13

N	SIMUL EN	$M M_n 1$ EN	$M G 1$ EN	$M G_n 1$ EN	$M M 1$ EN
3	4.062 ±.2466	5.4393	3.3159	4.9694	4.9090
4	3.636 ±.1575	4.2771	2.6060	4.1492	3.6018
5	3.291 ±.0333	3.9016	2.3309	3.8949	3.1134
6	3.277 ±.0975	3.7384	2.1889	3.7903	2.8602
7	3.241 ±.0982	3.6574	2.1004	3.7418	2.7132
8	3.235 ±.0991	3.6138	2.0403	3.7175	2.6032
9	3.232 ±.0913	3.5887	1.9964	3.7046	2.5282
10	3.203 ±.0844	3.5744	1.9637	3.6978	2.4725

Two Node Serial Semi-Open Network
with Equal Erlang Service Rates (N varies)

$$\lambda = 0.675, \mu_1 = 1.0, \mu_2 = 1.0$$

Table 3.14

λ	ρ	SIMUL EN	$M M_n 1$ EN	$M G 1$ EN	$M G_n 1$ EN	$M M 1$ EN
0.250	0.308	0.604 ±.0128	0.6460	0.3910	0.6945	0.6105
0.375	0.462	1.059 ±.0217	1.1601	0.7026	1.2487	0.8571
0.500	0.615	1.724 ±.0458	2.0011	1.2154	2.0771	1.6000
0.625	0.770	3.034 ±.1084	3.8278	2.0317	3.6650	3.3333
0.675	0.831	4.062 ±.2466	5.4393	3.3159	4.9694	4.9090

Two Node Serial Semi-Open Network
with Equal Erlang Service Rates (λ varies)

$$N=3, \mu_1 = 1.0, \mu_2 = 1.0$$

Table 3.15

N	SIMUL EN	$M M_n 1$ EN	$M G 1$ EN	$M G_n 1$ EN	$M M 1$ EN
3	6.006 ±.2956	8.0047	5.8885	6.6397	7.6754
4	6.166 ±.3777	7.4712	5.5395	6.2763	7.1138
5	5.875 ±.2862	7.3941	5.4863	6.2244	7.0281
6	5.512 ±.2067	7.3701	5.4688	6.2087	7.0000

Two Node Serial Semi-Open Network
with Unequal Erlang Service Rates (N varies)

$$\lambda = 1.75, \mu_1 = 5.0, \mu_2 = 2.0$$

Table 3.16

λ	ρ	SIMUL EN	$M M_n 1$ EN	$M G 1$ EN	$M G_n 1$ EN	$M M 1$ EN
0.50	0.253	0.414 $\pm$ 0.0046	0.4349	0.3158	0.4470	0.3383
0.75	0.379	0.706 $\pm$ 0.0104	0.7549	0.5498	0.7726	0.6107
1.00	0.506	1.096 $\pm$ 0.0148	1.2134	0.8865	1.2170	1.0225
1.25	0.632	1.725 $\pm$ 0.0288	1.9537	1.4314	1.8843	1.7170
1.50	0.758	3.014 $\pm$ 0.0655	3.4195	2.5119	3.1025	3.1381
1.75	0.885	6.600 $\pm$ 0.2956	8.0043	5.8849	6.6397	7.6754

Two Node Serial Semi-Open Network
with Unequal Erlang Service Rates (λ varies)

$$N=3, \mu_1 = 5.0, \mu_2 = 2.0$$

Table 3.17

θ	SIMUL EN	$M M_n 1$ EN	$M G 1$ EN	$M G_n 1$ EN	$M M 1$ EN
0.1	8.415 $\pm$ 2.044	14.4869	13.9183	10.5779	14.0000
0.2	1.103 $\pm$ 0.0409	1.1222	0.8699	1.3328	0.8750
0.3	0.634 $\pm$ 0.0202	0.6151	0.4489	0.6838	0.4516
0.4	0.441 $\pm$ 0.0195	0.4260	0.3026	0.4406	0.3043
0.9	0.170 $\pm$ 0.0056	0.1687	0.1150	0.1350	0.1157

Two Node Feedback Semi-Open Network
with Equal Exponential Service Rates (θ varies)

$$N=3, \lambda = 0.07, \mu_1 = 1.0, \mu_2 = 1.0$$

Table 3.18

θ	SIMUL EN	$M M_n 1$ EN	$M G 1$ EN	$M G_n 1$ EN	$M M 1$ EN
0.1	8.790 $\pm$ 1.684	10.5603	10.1994	9.1231	10.2279
0.2	1.010 $\pm$ 0.0300	1.0032	0.8345	1.1400	0.8364
0.3	0.549 $\pm$ 0.0155	0.5457	0.4351	0.5868	0.4361
0.4	0.380 $\pm$ 0.0101	0.3764	0.2942	0.3849	0.2949
0.9	0.149 $\pm$ 0.0030	0.1482	0.1124	0.1245	0.1126

Two Node Feedback Semi-Open Network
with Unequal Exponential Service Rates (θ varies)

$$N=3, \lambda = 0.175, \mu_1 = 5.0, \mu_2 = 2.0$$

Table 3.19

λ	SIMUL EN	$M M_n 1$ EN	$M G 1$ EN	$M G_n 1(1)$ EN	$M G_n 1(2)$ EN	$M M 1$ EN
0.01	0.355 ± 0.0388	0.3415	0.2048	0.3114	0.3173	0.2156
0.02	0.873 ± 0.0923	0.7944	0.5148	0.7980	0.8094	0.5498
0.03	1.444 ± 0.1276	1.4923	1.0397	1.5607	1.5779	1.1373
0.04	2.888 ± 0.2813	2.8987	2.1225	2.9175	2.9487	2.4420
0.05	7.247 ± 1.717	8.3867	5.6698	6.4401	6.5664	7.8370

Central Server Semi-Open Network

with Three Nodes (λ varies)

Exponential Service Times

$N=3, \mu_1 = 1.0, \mu_2 = 0.5, \mu_3 = 0.3$

$p_{12} = 0.7, p_{13} = 0.2, p_{21}, p_{31} = 1.0, p_{1,out} = 0.1$

Table 3.20

CHAPTER 4

HYDRODYNAMIC LIMIT FOR INFINITE SERIES OF EXPONENTIAL QUEUES

4.0 Introduction

One of the simplest open models which possesses the elegant *product form* solution is the serial network i.e., queues in series with Poisson input and exponential service times (single servers). The steady state analysis of this network shows that this network (with M nodes) can be decomposed and analysed separately as M single server $M/M/1$ queues. This result is due to the fact that the output process from each queue is Poisson. Similar to this result, most of the results on queueing networks are related to the steady state behavior. Since very few results exist about the time dependent or the transient behavior of these networks, it would be desirable to develop an analogue of the hydrodynamical study of the non-equilibrium flow of fluids. In recent years, considerable research has been carried out on the "hydrodynamic behavior" of certain infinite interacting particle systems (e.g. Liggett(1985)). With this motivation, the objective of this chapter is to describe the transient behavior of a simple queueing network namely, the flow of customers in the simple serial network using some recent developments from the theory of interacting particle systems..

The simple serial network we study here consists of single exponential servers. Service times at each server for each customer are exponentially distributed with rate one. There is unlimited waiting room in front of each server and the queue discipline is first come first served. In sections 4.1 through 4.4, we take the input process to be Poisson with rate one or less. When we do not specify the input process, we call this series of exponential

servers a series of exponential queues. In Section 4.5, we take the input process to be an arbitrary renewal process which satisfies the usual Strong Law of Large Numbers.

As we mentioned earlier, when a finite number of exponential queues in series receives a Poisson input with rate less than one, the equilibrium behavior of this network is very well understood and is known to have a *product form* solution for the steady state number of customers at each queue (Jackson(1963)). In order to study the flow of customers through this serial network, instead of following the traditional approach of queueing theory which studies the stochastic process directly connected to the characteristic of interest, we consider a model of an interacting particle system, namely, the asymmetric simple exclusion process which captures the essential characteristics of our serial network (see Kipnis(1986) for an earlier use of this model to queues in series). Also instead of studying the microscopic details of the network, we concentrate on the macroscopic behavior, i.e., we consider a very large serial network which has infinite number of single server exponential queues and study the flow of customers through this network for a suitable input process. We start this infinite series of queues in an empty state, and then let the network evolve in time according to the specified input process. For suitably scaled time and space (which will be explained later on), we concentrate on questions such as: How many queues have reached equilibrium?, With respect to the first customer who zips through the queues without any delay, how far back has one to go to find the queues in equilibrium? In providing answers to these questions, we also see that we can represent the macroscopic behavior of the flow of customers in terms of a partial differential equation which is common in many physical systems.

The first four sections are devoted to answering these questions when the infinite series of exponential queues receives a Poisson input. Here we isolate two cases namely,

the saturated exponential queues in series in which the traffic intensity is one and the unsaturated queues in series in which the traffic intensity is strictly less than one. In Section 4.1, we describe the asymmetric simple exclusion process, and the connection to the infinite series of exponential queues. In the Section 4.2, we explain the wave-like phenomena exhibited in the hydrodynamic limit of the saturated exponential queues (the traffic intensity is one) in series (Rost(1981)), and in the Section 4.3, we introduce an interacting particle system namely the asymmetric simple exclusion process with slow leading particle or simply a *Road Hog* process and provide the hydrodynamic limit and the density profile for the unsaturated exponential queues (traffic intensity is less than one) in series. In spite of the fact that the series of exponential queues with Poisson input is well understood, we obtain new insights into the behavior of customer flow through this network. In Section 4.4, we introduce another infinite particle system namely, the asymmetric zero range process as a model for the infinite series of unsaturated exponential queues. The results obtained in this section support the ones obtained in the Section 4.3. The Section 4.5 is devoted to studying the infinite series of exponential queues when the input process is arbitrary. We consider a special input process which satisfies the usual Strong Law of Large Numbers. In Section 4.5, we obtain some preliminary results necessary to derive the density profile for the asymmetric simple exclusion process which accommodates the special input process. In order to do this, we use coupling and stochastic ordering methods. In Section 4.6, we generalise the results obtained in the previous sections to a more complicated feedback queues in series which receives a non-Poisson input. The results obtained here also support the preliminary results in Section 4.5.

4.1 Asymmetric Simple Exclusion Process as a Model of Series of Exponential Queues

In what follows, we consider an infinite series of queues with identical service times which are exponential with rate one. The input process is Poisson with appropriate rate. This is the basic serial network we want to study. See Fig. 4.1.

Figure 4.1

Infinite Series of Exponential Queues

We know that when we have a finite number of queues in series, the steady state number of customers in the network satisfies the celebrated *product form* (Jackson(1963)). In order to study the flow of customers through these queues, we study a simple exclusion process with specified parameters which will describe our serial network. In this section, we will first explain the simple exclusion process and then identify our serial network which starts in the empty state.

4.1.1 Asymmetric Simple Exclusion Process

An interacting particle system consists of an infinite number of particles which evolve in the absence of interaction like independent Markov chains. As a result of interaction,

the evolution of a single particle is no longer Markovian but the system as a whole is a Markov process with a complex state space. Interactions can be used in many ways, of which we would like to consider the exclusion interaction. Particles move on the set of integers $\mathbb{Z}$ according to the following rules (we call $k \in \mathbb{Z}$ a site):

1. There is always at most one particle per site.
2. Particles stay at a site $k \in \mathbb{Z}$ for an exponential amount of time with rate one and at the end of this time they choose the site $k + 1$ with probability one.
3. If the site $k + 1$ is empty, the particle moves to $k + 1$, otherwise, the transition is suppressed and the particle remains at the site k .

The process $X(t)$ which describes the evolution of the particle system following the rules described above is an asymmetric simple exclusion process on $\mathbb{Z}$. The state space of this process is $E = \{0, 1\}^{\mathbb{Z}}$. The name exclusion is due to the fact that the transitions are allowed only to the empty site and all other transitions are suppressed. We denote the components of this process as $X(k, t)$, $k \in \mathbb{Z}$. Note that $X(k, t)$ assumes the value zero or one indicating at time t the site k is occupied or not. The set $\{k : X(k, t) = 1\}$ is the set of all occupied sites at time t . The generator of this Markov process is given by

$$Lf(X) = \sum_{k \in \mathbb{Z}} X(k)(1 - X(k+1)) [f(X^{k,k+1}) - f(X)]$$

where $X^{k,k+1}$ is obtained by permuting the coordinates of X at k and $k + 1$ and keeping the others fixed and f defined on E which depends only on finitely many coordinates. Let T_t be the semi group generated by L . Note that the process $X(t)$ has the property that the particles on $\mathbb{Z}$ always move to the right.

We know from Liggett(1985) that the product of Bernoulli measures with same parameter is an invariant measure for any simple exclusion process, i.e. with constant probability each site is occupied or not. This in turn implies that the interparticle distances are

geometrically distributed. This result is analogous to characterising the invariant distribution of the simple $M/M/1$ queue. A result similar to this one namely, characterising the invariant measures of the process is one of the main objectives of the study of interacting particle systems. In the next section, we will study this particle system in a different perspective which gives rise to the hydrodynamic limit behavior of the customers through the series of queues when traffic intensity is equal to one. The initial distribution $\nu^{1,b}$ of the process we are interested is a product measure on E such that

$$\nu^{1,b}\{X \in E : X(k) = 1\} = \begin{cases} 1, & \text{if } k \leq 0; \\ b, & \text{if } k > 0. \end{cases}$$

for a given $0 \leq b \leq 1$.

4.1.2 The Model

For the asymmetric simple exclusion process $X(t)$ defined above, consider the situation in which initially (at $t = 0$), all the sites to the left of one (excluding $k = 1$) are occupied and the sites to the right of zero (excluding $k = 0$) are empty. This situation corresponds to the initial distribution $\nu^{1,0}$ which is a product measure on E such that

$$\nu^{1,0}\{X : X(k) = 1\} = \begin{cases} 1, & \text{if } k \leq 0; \\ 0 & \text{if } k > 0. \end{cases}$$

Let $Z(k, t)$ denote the position of the particle at time t which was initially at $-k$ (i.e. at time zero). So initially, $Z(0, 0) = 0$, $Z(1, 0) = -1$ and $Z(2, 0) = -2$ etc. Now consider the interparticle distance $Y(k, t)$ defined as

$$Y(k, t) = Z(k-1, t) - Z(k, t) \quad \text{for } k = 1, 2, \dots$$

Initially $Y(k, 0) = 1$ for all $k = 1, 2, 3, \dots$. We interpret the interparticle distance $Y(k, t)$ minus one as the queue length at the k^{th} queue. Note that according to the description defined above, initially the number of customers at successive queues are identically

zero. Now label the particle which is initially at zero as the leading particle. The leading particle initially at zero under the initial distribution considered here, has no particle to its right and hence its motion is not at all affected by the particles to its left. This observation becomes very important in Section 4.3. When the leading particle moves to the right one step, the inter particle distance $Y(1, t)$ increases by one; this corresponds to an arrival in the first queue. Since the leading particle's holding time at a site is exponential with rate one, the leading particle's motion creates the input process which is Poisson to the first queue at rate λ . Note that the particle immediately behind the first particle can not move unless the leading particle moves to its right at least by one step. When the particle immediately behind the first particle moves to the right by one step, the interparticle distance $Y(1, t)$ decreases by one which corresponds to a service completion at the first queue; when this happens, $Y(2, t)$ increases by one and hence an arrival to the second queue. In general, when $Y(k, t)$ increases by one, there is a service completion at the $(k-1)^{th}$ queue and hence an arrival to the k^{th} queue. Similarly, when $Y(k, t)$ decreases by one, there is a service completion at the k^{th} queue and hence an arrival to the $(k+1)^{th}$ queue. Since the leading particle moves at a rate one which is equal to the rate of motion of other particles and hence the service rates, the situation described above represents the saturated serial network with traffic intensity equal to one. Therefore the asymmetric simple exclusion process explained above with the initial distribution prescribed completely characterises the saturated infinite series of exponential queues with Poisson input through the interparticle distances. Under the initial distribution given above, all the holes are initially to the right of the leading particle. As the process $X(t)$ evolves in time, the particles move to the right and the 'holes' move to the left.

In order to study the unsaturated exponential queues in series, we introduce an asym-

metric simple exclusion process with slow leader or simply a *Road Hog* process in which the leading particle moves at a slower rate when compared to all other particles. For example, when we assume that the holding time of the leading particle is exponential with rate say λ (< 1) (all other particle's holding times are exponential with rate one), as explained above, under the initial distribution $\nu^{1,0}$, the leading particle's motion creates the Poisson input process with rate λ to the first queue, and the interparticle distances describe the queue length at various queues. In a similar way, by changing the leading particle's holding time at a site to an independent and identically distributed arbitrary random variable, we can create the infinite series exponential queues which receives a non-Poisson input. The problem here is the process $X(t)$ is no longer Markovian, but, conditioned on the motion of the leading particle, $X(t)$ becomes Markovian. In the next section, we study the hydrodynamic limit behavior of the infinite series of saturated exponential queues with Poisson input.

4.2 Hydrodynamic Limit for the Infinite Series of Saturated Exponential Queues

In this section, we consider the asymmetric simple exclusion process in which the leading particle's holding time is exponential with rate one. The initial distribution $\nu^{1,0}$ is a product measure on E such that

$$\nu^{1,0}\{X : X(k) = 1\} = \begin{cases} 1 & \text{if } k \leq 0; \\ 0 & \text{if } k > 0. \end{cases}$$

In terms of the component process $\{X(k, t), t \geq 0, k \in \mathbb{Z}\}$, the initial distribution corresponds to $X(k, 0) = 1$ for $k \leq 0$ and $X(k, 0) = 0$ for $k > 0$. Let $EX(k, t)$ for $k \in \mathbb{Z}$ be the one point correlation or the occupancy number over the whole space E , then $EX(k, 0) = 1$ for $k \leq 0$ and $EX(k, 0) = 0$ for $k > 0$. Note that since $X(k, t)$ assumes the value zero or one, the quantity $EX(k, t)$ defines the probability with which each site is occupied. In this

respect, the initial distribution corresponds to the case in which the so called 'density' of the particles is one to the left of one and zero to the right of zero. We also point out that the simple exclusion process with the initial distribution $\nu^{(1-\lambda)}$ such that

$$\nu^{(1-\lambda)}(X : X(k) = 1) = (1 - \lambda), \quad \forall k \in \mathbb{Z}$$

will be called as an infinite series of exponential queue in equilibrium.

4.2.1 Preliminaries

Let $\mu(k, t)$ be the probability measure on E with respect to the initial distribution $\nu^{1,0}$ given by

$$\mu(k, t)\{Y : Y(k_1) = \dots = Y(k_n) = 1\} = P\{X : X(k + k_1, t) = \dots = X(k + k_n, t) = 1\}$$

Liggett(1985) has shown that

$$\text{weak } \lim_{t \rightarrow \infty} \mathcal{L}\mu(0, t) = \beta_{1/2}$$

where β_a is the Bernoulli measure of density a , i.e., for $0 \leq a \leq 1$

$$\beta_a(X(i_j) = 1, j = 1, 2, \dots, n) = a^n \quad \text{for all } n, i_1, i_2, \dots, i_n$$

This result implies that at any fixed set $\{i_1, i_2, \dots, i_n\}$ of sites, the random variables $X(i_j, t)$ for $j = 1, 2, \dots, n$ are independent with density $1/2$. Instead of studying the distribution of the process $X(t)$ at a fixed set of sites, in this section, we want to study the behavior of the process $X(t)$ near sites k_t for large times where $k_t \rightarrow \pm\infty$ as $t \rightarrow \infty$ at an appropriate rate. In other words, we want to study the weak convergence of $X([xt] + k, t)$ as $t \rightarrow \infty$ for every fixed $x \in \mathbb{R}$ not only for $x = 0$ as studied in Liggett(1985), i.e., we want to show

$$\text{weak } \lim_{t \rightarrow \infty} \mathcal{L}X([xt] + k, t) = \beta_{u(x,1)} \quad (4.1)$$

where $u(x, t)$ is the density profile defined as

$$u(x, t) = \lim_{h \rightarrow 0} EX(x/h, t/h) \quad (4.2)$$

The limit in (4.1) is usually called the hydrodynamic limit of the process $X(t)$. In order to show (4.1), the following steps are needed:

1. For any $x \in \mathcal{R}$, the limit $\lim_{h \rightarrow 0} EX(x/h, t/h)$ exists and is equal to $u(x, t)$ (the density profile).
2. The limit in step one is usually identified by considering the quantities $\sum_{m \leq l \leq n} X(l, t)$ which gives the total number of particles between the sites m and n and showing that the following Strong Law of Large Numbers holds:

$$\lim_{h \rightarrow 0} h \times \sum_{[x/h]+1}^{[y/h]} X\left(k, \frac{t}{h}\right) \stackrel{a.s.}{=} \int_x^y u(z, t) dz \quad \text{in } L^1.$$

3. The identification of limiting behavior of the process seen by a travelling observer, i.e., the weak limit distribution of $(X([xt] + k, t), k \in \mathcal{Z})$ as $t \rightarrow \infty$ for every fixed $x \in \mathcal{R}$.

Remarks: The density profile $u(x, t)$ can also be explained as follows: For the component process $\{X(k, t), k \in \mathcal{Z}\}$ define the rescaled process $X_t^h(x)$ for $x \in \mathcal{R}$ and for $h > 0$ as

$$X_t^h(x) = h \times \sum_{k \in \mathcal{Z}} X(k, t/h) I_{[hk, h(k+1))}(x)$$

Here $X_t^h(x)$ defines a family of measures on the real line. In other words, each wandering particle contributes a mass h to the random measure. Now the density profile $u(x, t)$ is the almost sure weak limit of the measures $X_t^h(x)dx$ as h goes to zero. For $t = 0$,

$X_0^h(x)dx$ converges weakly almost surely to the deterministic measure $1_{(-\infty, 0]}$ which is just the Lebesgue measure of the negative real line. By step (1), the existence of the density profile suggests that $X_t^h(x)dx$ converges to the deterministic measure $u(x, t)dx$ as h goes to zero. Also $u(x, t)$ satisfies the first order non-linear partial differential equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} u(1 - u) = 0 \quad (4.3)$$

with the initial condition $u(\cdot, 0) = 1_{(-\infty, 0]}$.

4.2.2 Density Profile and Hydrodynamic Limit

For the process $X(t)$, with the initial distribution $\nu^{1,0}$ on E , we have the following two theorems from Rost(1981).

Theorem 4.1 Density Profile: For any $x \in \mathcal{R}$, the limit

$$\lim_{h \rightarrow 0} EX(x/h, t/h)$$

exists and is equal to $u(x, t)$, where $u(x, t)$ is given by

$$u(x, t) = \begin{cases} 1 & \text{if } x \leq -t; \\ \frac{1}{2} \left(1 - \frac{x}{t}\right) & \text{if } -t \leq x \leq t; \\ 0 & \text{if } x \geq t. \end{cases} \quad (4.4)$$

and (Strong Law of Large Numbers)

$$\lim_{h \rightarrow 0} h \times \sum_{[x/h]+1}^{[y/h]} X\left(k, \frac{t}{h}\right) \stackrel{a.s.}{=} \int_x^y u(z, t) dz \quad \text{in } L^1$$

holds.

Theorem 4.2 For any $x \in \mathcal{R}$, the distribution of the $(X([xt] + k, t), k \in \mathcal{Z})$ as $t \rightarrow \infty$ converge weakly to $\beta_{u(x, 1)}$, the Bernoulli measure with density $u(x, 1)$.

The density profile is given in Fig. 4.2.

There are few points worth noticing about the profile given in the above figure.

Figure 4.2

Density Profile for the Saturated Queues

1. When compared to the initial distribution $\nu^{1,0}$, the density profile suggests that the process $X(t)$ expands linearly in time and admits $u(x, t)$ as its density profile.
2. The leading particle which was initially at zero, is at the site $k = t$. This is intuitively obvious because the leading particle's motion is Poisson with parameter t which satisfies the Strong Law of Large Numbers.
3. Near the leading particle, i.e., near $x = t$, the density is small and hence the interparticle distances are large. In other words, the queue lengths are large which is in quite agreement with the known results about the queue lengths when the traffic intensity is one.
4. None of the particles to the left of $x = -t$ have moved, and the density of the particles near this site is large implying that the interparticle distances are small. Note that $x = -t$ is the position of the first customer who has zipped through the queues without any queueing delay.
5. Since the traffic intensity is one, none of the queues would have reached equilibrium. But the Theorem 4.2 suggests that there are local equilibria, i.e., for every $x \in \mathcal{R}$, there is a block of particles which are at a fixed distance from $\{x\}$ for which the asymptotic independence of $X(k, t)$ holds, where $\lfloor \cdot \rfloor$ denotes the greatest integer less

than or equal to the quantity inside the brackets.

4.3. Hydrodynamic limit for the Infinite Series of Unsaturated Exponential Queues with Poisson input

In this section, we would like to construct the asymmetric simple exclusion process $X(t)$ which correspond to the series of exponential queues with identical exponential service rate one and a Poisson input with rate less than 1. This situation in turn corresponds to studying the *Road Hog* process $X^*(t)$ which starts with the initial distribution $\nu^{1,0}$ and the leading particle which is initially at zero moving slowly at a rate less than one and all other particle's holding times are exponential with rate one. Let the leading particle's holding time be exponential with rate $\lambda < 1$. Then under the initial distribution $\nu^{1,0}$, the initial queue sizes are zero and the leading particle's motion is Poisson at rate λt .

Thus in order to study the hydrodynamic limit of the unsaturated queues in series, we have to study this asymmetric simple exclusion process with slow leading particle. Note that when compared to the simple exclusion process described in the last section, here we are deliberately forcing the leading particle to move slowly. Though we have created an ideal simple exclusion process which corresponds to the unsaturated queues in series, we can not use any of the existing standard techniques such as coupling when we isolate a single particle to behave differently than others. In order to study this *Road Hog* process which represents our situation, we consider a different initial distribution for our original asymmetric simple exclusion process in which the leading particle's holding time is identical to that of other particles and then identify an embedded process which represents our situation.

4.3.1 Embedding Method

In this section, instead of studying the *Road Hog* process directly, we provide a method to identify an embedded process in the original simple exclusion process with special initial distribution which yields a version of the *Road Hog* process. Consider the original simple exclusion process $X(t)$ with the initial distribution $\nu^{1,(1-\lambda)}$ which is a product measure on E such that

$$\nu^{1,(1-\lambda)}\{X : X(k) = 1\} = \begin{cases} 1 & \text{if } k \leq 0; \\ (1-\lambda) & \text{if } k > 0. \end{cases} \quad (4.5)$$

This initial distribution corresponds to the case in which all the sites to the left of one are occupied and the sites to the right of zero are occupied with constant probability $(1-\lambda)$.

Let us assume that the particles initially at the positions $k \leq 0$ are colored red and $k > 0$ are colored black. Now we can think of the initial distribution as two simple exclusion processes corresponding to red and black particles, with initial distributions $\nu^{1,0}$ and $\nu^{0,(1-\lambda)}$ respectively being juxtapositioned. Note that in the absence of black particles, the red particles correspond to the simple exclusion process studied in Section 4.2. Since particles always move to the right, the red particles, do not affect the motion of the black particles, but black particles do slow down the motion of the red particles. Let $X_1(t)$ be the simple exclusion process which corresponds to the red particles. We call this process an embedded process in $X(t)$. We also call the red particle which was originally at zero a *tagged particle*. Note that the presence of black particles to the right of zero affect the motion of this *tagged particle* and we would like to know how much this particle gets slowed down.

Lemma 4.1 *The initial distribution $\nu^{1,(1-\lambda)}$ (given in the beginning of Section 4.2) for the process $X(t)$ corresponds to the infinite series of exponential queues in equilibrium with traffic intensity λ . Moreover, the position of the tagged particle is Poisson with parameter λt .*

Proof: Since Bernoulli measures with constant parameter $(1 - \lambda)$ are invariant for the simple exclusion process, according to this initial distribution, the interparticle distances are geometrically distributed with parameter λ and hence $\nu^{(1-\lambda)}$ correspond to the infinite series of exponential queues in equilibrium.

In equilibrium, the departure process from successive queues are Poisson and the queue lengths to the right of the *tagged particle* are independent of the departure process up to time t . Since the *tagged particle's* position is equal to the total increase of customers (holes) to its left, and hence to the departure process, the position of the *tagged particle* is Poisson with parameter λt . $\square$

Note: See page 398 and 414 of Liggett(1985) for another proof.

The following Embedding lemma provides the connection between the embedded process and the *Road Hog* process.

Embedding Lemma: Under the initial distribution $\nu^{1,(1-\lambda)}$, the leading red particle's position at time t in $X_1(t)$ (i.e., the *tagged particle's* position at time t in $X(t)$) is Poisson with parameter λt for all $t \geq 0$ and the embedded process $X_1(t)$ (which corresponds to the red particles) has the same probability law as the *Road Hog* process $X^*(t)$ with the initial distribution $\nu^{1,0}$.

Proof: The leading red particle's motion is only affected by the black particles which are distributed initially to the right of zero with constant probability $(1 - \lambda)$. Therefore, from Lemma 4.1, the position of the *tagged particle* at time t , and hence the position of the leading red particle (with respect to the initial distribution $\nu^{(1-\lambda)}$ to the right of zero) is Poisson with parameter λt for all $t \geq 0$.

In order to show the second part of the statement, note that the *tagged particle* in $X(t)$ is the same as the leading red particle in $X_1(t)$ and its position is Poisson with

parameter λt . Whereas in the *Road Hog* process $X^*(t)$, the leading particle moves at a rate λ and hence under the initial distribution $\nu^{1,0}$, its position is Poisson with parameter λt . Therefore, the embedded process $X_1(t)$ in $X(t)$ has the same probability law as the *Road Hog* process $X^*(t)$. $\square$

The purpose of identifying the embedded process in $X(t)$ with the *Road Hog* process is to use some of the existing hydrodynamic limit and density profile results for the process $X(t)$ and then deduce the corresponding results in our case.

4.3.2 Density Profile and Hydrodynamic Limit

Having identified the *Road Hog* process in which the leading particle moves slower than the other particles, in order to obtain the density profile and the hydrodynamic limit, we use the following theorem due to Benassi and Fouque (1987) for the process $X(t)$ under the initial distribution $\nu^{1,(1-\lambda)}$.

Theorem 4.3 For any $x \in \mathcal{R}$, the limit $\lim_{h \rightarrow 0} EX(x/h, t/h)$ exists and is equal to the density profile $u(x, t)$ which is given by

$$u(x, t) = \begin{cases} 1 & \text{if } x \leq -t; \\ \frac{1}{2} \left(1 - \frac{x}{t}\right) & \text{if } -t \leq x \leq (2\lambda - 1)t; \\ (1 - \lambda) & \text{if } x \geq (2\lambda - 1)t. \end{cases} \quad (4.7)$$

and (Strong Law of Large Numbers)

$$\lim_{h \rightarrow 0} h \times \sum_{[x/h]+1}^{[y/h]} X\left(k, \frac{t}{h}\right) \stackrel{a.s.}{=} \int_x^y u(z, t) dz \quad \text{in } L^1.$$

holds. Moreover for every fixed $x \in \mathcal{R}_x$

$$\text{weak } \lim_{t \rightarrow \infty} \text{law}\{X([xt] + k, t), k \in \mathcal{Z}\} = \beta_{u(x,1)}$$

The density profile for the process $X(t)$ given in the above theorem is illustrated in the Fig-4.3 for $\lambda > 1/2$. In the last section, when we studied the process $X(t)$ under

the initial distribution $\nu^{1,0}$, all the interparticle distances were initially one which corresponded to the empty queues in the serial network. Now in this section, under the initial distribution $\nu^{1,(1-\lambda)}$, not all the interparticle distances are identically one. But, initially, the interparticle distances corresponding to the red particles are one and only these particles which correspond to the embedded process and hence the *Road Hog* process relate to the empty queues. As explained earlier, the presence of black particles to the right of zero only serve the purpose of slowing down the *tagged particle*.

Figure 4.3

Density Profile (Benassi & Fouque)

Let us look at the density profile for the process $X(t)$ in Fig 4.3 carefully, $u(x, t) = (1 - \lambda)$ for $x \geq (2\lambda - 1)t$. Knowing that the *tagged particle's* position in $X(t)$ and hence the leading particle's position in $X^*(t)$ is Poisson with parameter λt , using the Strong Law of Large Numbers for the Poisson process, we locate the position of the *tagged particle* on the profile as λt . Note that in the case of infinite series of saturated queues, the leading particle's motion is Poisson with parameter t and $x = t$ is the position of this particle in the density profile. With this observation, the following Theorem states the results for the *Road Hog* process.

Theorem 4.4 Theorem 4.3 is true when we replace $X(t)$ by $X^*(t)$ with $u(x, t)$ defined as

$$u(x, t) = \begin{cases} 1 & \text{if } x \leq -t; \\ \frac{1}{2} \left(1 - \frac{x}{t}\right) & \text{if } -t \leq x \leq (2\lambda - 1)t; \\ (1 - \lambda) & \text{if } (2\lambda - 1)t \leq x \leq \lambda t; \\ 0 & \text{otherwise.} \end{cases} \quad (4.8)$$

The density profiles for the *Road Hog* process when $\lambda < 1/2$, $\lambda = 1/2$ and $\lambda > 1/2$ are illustrated in the Fig. 4.4.

Figure 4.4

Density Profile for the Unsaturated Queues

The following points are worth noting about these profiles.

1. In contrast to the saturated case, all the cases here exhibit a constant leading edge

where $u(x, t)$ is constant and is equal to $(1 - \lambda)$. Since in this case, the traffic intensity ($= \lambda$) is less than one, some of the queues from the front have reached equilibrium.

2. How many queues have reached equilibrium depends on the density of the particles $(1 - \lambda)$ to the right of zero in the initial distribution. Note that the length of the equilibrium stretch also depends on the value of λ .
3. Note that irrespective of the value of λ the leading customer who zips through the queues is always at queue numbered $x = -t$ on this time scale.
4. Since the product of Bernoulli measures with same parameter is invariant for the simple exclusion process, in the flat region $((2\lambda - 1)t \leq u \leq \lambda t)$, if we consider a fixed set of sites, the one point correlations are independent with the same density. This suggests that the interparticle distances are geometric which was expected for the $M/M/1$ queues in series at least for the first few queues.
5. **Theorem 4.5** *In the hydrodynamic time scale, the first $(1 - \lambda)^2 t$ queues are in equilibrium i.e., in the first $(1 - \lambda)^2 t$ queues, the queue lengths are independent, identical and geometrically distributed with parameter $(1 - \lambda)$.*

Proof: In the region $(2\lambda - 1)t \leq x \leq \lambda t$ of length $(1 - \lambda)t$, the density profile is constant and the interparticle distances are geometrically distributed with parameter $(1 - \lambda)$. Since each site in this region is occupied with probability $(1 - \lambda)$, the probability each site is a hole is λ . Therefore, the number of holes and hence the number of customers in this region is $\lambda(1 - \lambda)t$. Also, since the interparticle distances are geometrically distributed with parameter $(1 - \lambda)$, the number of queues which are in equilibrium are $(\lambda(1 - \lambda)t)/(\lambda/(1 - \lambda)) = (1 - \lambda)^2 t$.

6. With respect to the first customer who zipped through the queues without any delay, who is at the site $x = -t$, we have to go back to the queue numbered $(1 - \lambda)^2 t$ to see

the queues in equilibrium.

7. The density profile $u(x, t)$ also satisfies the same differential equation (4.3) given in Section 4.2.

4.4 Asymmetric Zero Range Process as a Model of Infinite Series of Unsaturated Exponential Queues

So far in the last three sections, we used the asymmetric simple exclusion process to derive the density profile and the hydrodynamic limit for the infinite series of exponential queues. In this section, supporting the results obtained in the last section, we illustrate another use of studying the infinite particle system by considering an asymmetric zero range process as a model for the infinite series of exponential queues in series.

4.4.1 Asymmetric Zero Range Process

The one dimensional asymmetric zero range process $N(t)$ can be explained intuitively as follows:

1. At each site $k \in \mathbb{Z}$, there are a certain number of particles.
2. Particles move from site k to $k + 1$ at rate $g(k)$. When the rate $g(k) = 1$, one of the particle chosen at random from a site moves to its right hand side neighbor whenever an exponential bell attached to the site rings. The parameter of the exponential bell is one.

Now corresponding to this description, there exists a unique Markov process $N(t)$ with state space $\mathcal{N}^{\mathbb{Z}}$. The generator of this process for $N \in \mathcal{N}^{\mathbb{Z}}$ and acting on a function f which depends only on finitely many coordinates is

$$Lf(N) = \sum_{k \in \mathbb{Z}} 1_{\{N(k) > 0\}} [f(N^{k, k+1}) - f(N)]$$

where $N^{k,k+1}$ is obtained by adding one particle to the site $k+1$ and subtracting a particle from the site k and also keeping the particles at other sites fixed from the configuration N . The set of extremal invariant measures for this process $N(t)$ is the set of product measures ν^λ such that

$$\nu^\lambda\{N(k) = l\} = \lambda^l(1 - \lambda)$$

for $k \in \mathbb{Z}$ and $0 \leq \lambda < 1$ (Andjel(1982)). This extremal invariant measure corresponds to distributing particles at each site according to a geometric distribution with parameter $(1 - \lambda)$: For the process $N(t)$, we consider the following initial distribution $\nu^{\lambda,0}$, i.e.,

$$\nu^{\lambda,0}\{N(k) = l\} = \begin{cases} \lambda^l(1 - \lambda) & \text{if } k \leq 0 \\ 0 & \text{if } k > 0. \end{cases} \quad (4.9)$$

4.4.2 The Model

Under the initial distribution $\nu^{\lambda,0}$, each site to the left of zero consists of particles which are geometrically distributed with parameter $(1 - \lambda)$ and to the right of zero there are no particles. We interpret each site $k \in \mathbb{Z}$ as a queue and the number of particles at each site as the number of customers in that queue. Recall that in the simple exclusion process, the interparticle distances or the number of holes corresponded to the customers in a queue. Here the queues correspond to $k = 1, 2, \dots$ are empty initially and these are the queues which correspond to our infinite series of queues. As was done in the last section, the purpose of distributing particles to the left of zero according to geometric distribution is to create the input process to the queue at $k = 1$. Since the geometric distribution is invariant for an $M/M/1$ queue, the queues at sites $k = -1, -2, \dots$ start with the equilibrium number of customers. Now Burke's output theorem for the exponential queues in series implies that the input process to the queue at $k = 1$ and hence the successive instants at which particles jump from the site zero to one under equilibrium is

2

of/de

2

1.0

1.1

1.25

1.4

1.6

1.8

2.0

2.2

2.5

2.8

3.2

3.6

4.0

4.5

5.0

5.6

6.3

7.1

8.0

9.0

10.0

11.2

12.5

14.0

16.0

18.0

20.0

22.5

25.0

MicroD

Poisson with parameter λ . Let $N(0, t/h)$ be the number of particle which have crossed the site zero at time t/h . Then we have the following lemma.

Lemma 4.2 Under the initial distribution $\nu^{\lambda,0}$,

$$\lim_{h \rightarrow 0} h N(0, t/h) = \lambda t.$$

Though we have not specified the input process to queue at $k = 1$, for large times, as the lemma suggests, the output from the queue at $k = 0$ and hence input to the queue at $k = 1$ is Poisson with parameter λ . In spite of the fact that we using known results such as Burke's theorem for the queues in series, we once again insist on the fact that the purpose of this chapter is study the flow of customers through the queues and hence to study how the equilibrium spreads in time. In the next section, we provide the hydrodynamic limit for this model.

4.4.3 Density Profile and Hydrodynamic Limit

Let T_t be the semi group for the process $N(t)$ and let τ_a be the translation operator (τ_a is the translation induced by a on the configuration $N \in \mathcal{N}^{\mathbb{Z}}$). The density profile $p(x, t)$ for the process $N(t)$ is defined as

$$\lim_{h \rightarrow 0} \tau_{[x/h]} \nu^{\lambda,0} T_{t/h} = \nu_{p(x,t)}$$

where $[x/h]$ denotes the greatest integer less than or equal to x/h . The following theorem is due to Andjel and Kipnis(1984) who shows the existence and the explicit identification of the density profile $p(x, t)$.

Theorem 4.6 For any $x \in \mathcal{R}$,

$$\lim_{h \rightarrow 0} \tau_{[x/h]} \nu^{\lambda,0} T_{t/h}$$

exists and is equal to $\nu^{p(x,t)}$ where $p(x,t)$ is given by

$$p(x,t) = \begin{cases} \lambda & \text{if } x < (1-\lambda)^2 t; \\ 1 - \sqrt{\frac{x}{t}} & \text{if } (1-\lambda)^2 t \leq x < t; \\ 0 & \text{if } x \geq t \end{cases} \quad (4.10)$$

The Fig.4.5 illustrates the density profile.

Figure 4.5

Density Profile (Zero Range Process)

The density profile suggests the following points:

1. Initially, $p(x,t)$ was λ to the left of zero and 0 to the right of zero, i.e. particles were distributed to the left of zero according to the geometric distribution with parameter $(1-\lambda)$ and no particles to the right of zero.
2. As the density profile indicates, $p(x,t) = \lambda$ for $x \leq (1-\lambda)^2 t$, suggesting that the equilibrium phenomena has spread up to the queue $x = (1-\lambda)^2 t$ (see Theorem 4.5). In other words, for sufficiently large t , the queues at $x = 1, 2, \dots, (1-\lambda)^2 t$ which were initially empty have 'reached equilibrium' with the same parameter $(1-\lambda)$.
3. The queues in between $x = (1-\lambda)^2 t$ and $x = t$ are in local equilibrium where each queue in this region has geometric number of customers with parameter $1 - \sqrt{x/t}$.

4.4.4 Comparison of the Hydrodynamic Density Profiles

Now, we show that the density profile obtained for the zero range process are consistent with that of the simple exclusion process. For this purpose, we consider the density profile picture for the simple exclusion process in which the leading particle's rate is $\lambda > 1/2$ given in Fig 4.4.

As we know for the simple exclusion process, the density profile $u(x, t)$ is the occupancy number or the probability of a site being occupied. In the equilibrium region, i.e., $(2\lambda - 1)t \leq x \leq \lambda t$, the probability that each site is occupied is $(1 - \lambda)$ and hence the probability each site is a hole λ . Since the number of holes between two occupied sites correspond to the number of customer in a queue, the total number of customers in the equilibrium region is $\lambda(1 - \lambda)t$. Now for the zero range process, in the equilibrium region, the density profile $p(x, t) = \lambda$ which implies that the number of customers at each queue in the region $0 \leq x \leq (1 - \lambda)^2 t$ is geometric with parameter $(1 - \lambda)$. Therefore, the total number of customer in this region is $\lambda(1 - \lambda)t$ which is the same as obtained by $u(x, t)$. Similarly, the total number of customers in the region $-t \leq x \leq (2\lambda - 1)t$ for the simple exclusion process and the total number of customers in the region $(1 - \lambda)^2 t \leq x \leq t$ for the zero range process is $\lambda^2 t$. This observation also confirms the fact that in the simple exclusion process, the leading particle is at λt and hence the number of holes to the left of this particle is λt which is the total number of customers who have arrived to the infinite series of queues. For the zero range process, the number of particles which have crossed the bond $(0, 1)$ is λt which is the total number of customers who have arrived to the queues $x = 1, 2, \dots$

4.5 Density Profile for Infinite Series of Exponential Queues with non-Poisson Input

So far, in the last four sections, we studied the infinite series of exponential queues with Poisson input. In this section, we will let the input process to be arbitrary but one which satisfies the Strong Law of Large Numbers. Once again, we consider the asymmetric simple exclusion process with a slow leader or simply the *Road Hog* process as a model for our queueing network. As explained earlier, in order to accommodate our input process to the exponential queues in series, we let the holding time of the leading particle to be arbitrary. In other words, if T_i for $i \geq 1$ denotes the random variables representing the successive holding times of the leading particle at different sites, then we assume that with probability one,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n T_i = 1/\lambda \quad (*)$$

where λ is the inverse of the expected value of T_1 . We will also assume that $\lambda < 1$, so that the leading particle moves at a slower rate than that of the other particles and we have unsaturated exponential queues in series. The exponential queues in series with Poisson input treated in Section 4.3, is a special case of this section since the exponential holding times obviously satisfies the Strong Law of Large Numbers.

Though under the assumption of arbitrary holding time, there is no Markov process which represents the intuitive description of the simple exclusion process, conditioned on the motion of the leading particle, the process $X(t)$ is Markov. Recall that in Section 4.3, when the leading particle's holding time was exponential with rate less than one, instead of dealing with the *Road Hog* process directly, we used the embedded process to derive the density profile. Here we can not use the same methods since the leading particle's holding time is arbitrary. In order to derive the density profile in this case, we introduce a coupling technique which is quite different from the classical one, and develop some monotonicity properties for this process. Before we discuss the coupling, we first explain

the intuitive reasoning behind the results of this section by considering the exponential holding time for the leading particle.

4.5.1 Motivation Behind the Stochastic Comparison Method

Consider two simple exclusion processes $X(t)$ and $Y(t)$ in which the holding times of the leading particles are exponential with rate λ_1 and λ_2 respectively ($\lambda_1 > \lambda_2$ and $\lambda_1, \lambda_2 < 1$). As explained in the previous sections, the processes $X(t)$ and $Y(t)$ correspond respectively to two infinite series of unsaturated exponential queues which receives Poisson inputs with rate λ_1 and λ_2 respectively. The density profiles $u_X(x, t)$ and $u_Y(x, t)$ superimposed on each other is given in Fig.4.6.

Figure 4.6

Density Profiles $\lambda_1 > \lambda_2$

From the figure, we have the following observations about these two profiles:

1. Since $\lambda_1 > \lambda_2$, the position of the leading particles in the density profiles for the two process satisfies the natural ordering $\lambda_1 t > \lambda_2 t$.
2. For the process $X(t)$, the number of particles to the right of $x = \lambda_2 t$ ($< \lambda_1 t$) is larger than that for the process $Y(t)$. Actually, there are no particles to the right of $x = \lambda_2 t$ for the process $Y(t)$.

3. In the region $(2\lambda_2 - 1)t \leq x \leq \lambda_2 t$, the density of the particles for the process $Y(t)$ is strictly larger than that for $X(t)$, i.e. there are more holes in this region for the process $X(t)$ than that of $Y(t)$.
4. All these observations amount to saying that the number of particles to the right of $y > (2\lambda_2 - 1)t$ for the process $X(t)$ is greater than that for the process $Y(t)$.
5. Also the number of particles to the right of $x = (2\lambda_2 - 1)t$ is the same for the two processes and is equal to $(1 - \lambda_2)^2 t$.

In the following subsections, we will try to provide probabilistic statements of these observations for all time t instead of just for the limiting case.

4.5.2 Coupling for Multi-Coloured Particle Systems

Let $X_1(t)$ and $X_2(t)$ be two simple exclusion processes in which the leading particles move slowly when compared to other particles. Let λ_1 and λ_2 be the rate at which these particles move and $\lambda_1 > \lambda_2$. We consider the coupled process $\bar{X}(t) = (X_1(t), X_2(t))$ which is intuitively explained as follows: There are four kinds of particles distributed along the sites of $k \in \mathbb{Z}$. The particles which correspond to the process $X_1(t)$ are coloured *red* and *black*. The colour of the leading particle in this process is *black* and all other particles are coloured *red*. For the process $X_2(t)$, the leading particle's colour is *white* and all other particles are coloured *green*. Therefore the *red* and *black* pair correspond to the process $X_1(t)$ and the *green* and *white* pair correspond to the process $X_2(t)$.

The state space for the process $\bar{X}(t)$ is $E \times E$. There are at most two colours of particles at each site (*red* and *green*, *red* and *white*, *black* and *green* and *black* and *white* pairs). If at site k there is a *red* and *green* pair, they try to move simultaneously to the site $k + 1$ after an exponential amount of time with rate one if they can. If on the other hand, only one (*red* or *green*) can move then it does so after the exponential amount of

time. Note that the coupling of this pair correspond to the classical coupling. We denote this coupling as a *red-green* coupling. When a *red and white* or *green and black* or *black and white* pairs appears at a site, no coupling is invoked, and they move independently according to the simple exclusion interaction.

We can not enforce the normal coupling for these pairs because their holding times are completely different. However, when the leading particles (*black* and *white*) holding times are exponential with different rates, we will introduce a *black-white* coupling for this pair in addition to the *red-green* coupling introduced for the *red and green* pair in the following subsection which is quite different from the classical one. This special case will be referred to as the exponential case and the situation in which the holding time of the *black* particle is exponential and that of the *white* particle is general will be referred to as the non-exponential case.

4.5.3 Definitions and Notations

As usual let $E = \{0, 1\}^{\mathbb{Z}}$ be the state space of the asymmetric simple exclusion process. For $X \in E$, define $n_0(X)$ and $m_0(X)$ as

$$n_0(X) = \max\{k : X(k) = 1\} \quad (4.11)$$

and

$$m_0(X) = \min\{k : X(k) = 0\}. \quad (4.12)$$

Note that for a configuration $X \in E$, $n_0(X)$ defines the position of the leading particle and $m_0(X)$ defines the position of the left most hole. All the simple exclusion processes considered in this section are assumed to have the property that $n_0(X) < \infty$ and $m_0(X) > -\infty$ so that no extremities are allowed in the configurations. Now consider the reduced

space E^* of E where E^* is defined as

$$E^* = \{X \in E : n_0(X) < \infty \text{ and } m_0(X) > -\infty\}.$$

Also the position of the k^{th} particle, $n_k(X)$ is iteratively defined as

$$n_k(X) = \max \{l : l < n_{k-1}, X(l) = 1\}, \quad k = 1, 2, \dots \quad (4.13)$$

Note that the position of the leading particle $Z(0, t)$ which was used in the previous sections corresponds here to $n_0(X(t))$ and the position of the k^{th} particle $Z(k, t)$ corresponds to $n_k(X(t))$. For the two simple exclusion process $X_1(t)$ and $X_2(t)$, consider the coupled process $(X_1(t), X_2(t))$ with state space $E^* \times E^*$. Let $\bar{k}_0$ be a function from $E^* \times E^* \rightarrow Z \cup \{-\infty\} \cup \{\infty\}$ defined as

$$\bar{k}_0 = \inf \left\{ k \in Z : \sum_{l \geq k'} X_1(l) \geq \sum_{l \geq k'} X_2(l) \quad \forall k' \geq k \right\} \quad (4.14)$$

Intuitively, in the coupled process, for any given configuration $\{X_1, X_2\}$ $\bar{k}_0$ defines the position of the site in Z from where the number of particles to the right of every site in X_1 is greater than or equal to the number of particles in X_2 . Also, we say $\bar{k}_0 = -\infty$ or $\bar{k}_0 = \infty$ when this relation holds for all $k \in Z$ or for no $k \in Z$. Similarly, we define k_0^* as

$$k_0^* = \sup \{k \in N : n_l(X_1) \geq n_l(X_2) \quad \forall l \leq k\}. \quad (4.15)$$

From the definition of $n_l(X)$, k_0^* identifies the number of particles which are ordered through labels. We say $k_0^* = \infty$ when all the particles are ordered and $k_0^* = -\infty$ when none of the particles are ordered. We also define $\hat{k}_0$ as follows

$$\hat{k}_0 = n_{k_0^*+1}(X_2) + 1 \quad (4.16)$$

which is the location of the site to the right of which all the particles are ordered. Also $S(X, k) = \sum_{n \geq k} X(n)$ is defined as the number of particles to the right of k in the configuration X .

4.5.4 Stochastic Comparisons

Theorem 4.7: Let $X_1(t)$ and $X_2(t)$ be two simple exclusion processes coupled via red- & green coupling and in addition assume that for all $t \geq 0$, $n_0(X_1(t)) \geq n_0(X_2(t))$ then for all $t \geq 0$, with probability one,

$$\overline{k_0}(t) = -\infty; \quad k_0^*(t) = \infty$$

whenever $X_1^0 = X_2^0$.

In order to prove this theorem, it is enough to show the following lemma which is stated in terms of the labelled process $n_k(X(t))$.

Lemma 4.3: Under the assumption of Theorem 4.7 (i.e., $n_0(X_1(t)) \geq n_0(X_2(t))$ for all $t \geq 0$), we have

$$n_k(X_1(t)) > n_k(X_2(t)) \quad \forall k = 0, 1, \dots, \quad \forall t \geq 0$$

whenever $X_1^0 = X_2^0$. Note that initially $n_k(X_1(0)) = n_k(X_2(0))$ for all $k = 0, 1, \dots$

From the definitions of $\overline{k_0}$ and k_0^* , this lemma implies the theorem is obvious.

Proof of Lemma 4.3: Under the assumption given in the lemma, we have for all time $t \geq 0$, $n_0(X_1(t)) \geq n_0(X_2(t))$. Now conditioned on this statement, we will show that $n_1(X_1(t)) \geq n_1(X_2(t))$. Note that these are the particles which were initially at -1 , i.e. a red and green pair and this pair is coupled in the standard way. The following are the options of jumping for this pair:

the pair jump together, or

either red or green jumps to the right, or

none of them jump

Since $n_0(X_1(t)) \geq n_0(X_2(t))$, the position $n_1(X_1(t))$ will be either greater than or equal to $n_0(X_2(t))$ or less than $n_0(X_2(t))$.

Case 1 $n_1(X_1(t)) \geq n_0(X_2(t))$: In this case, since $n_1(X_2(t)) < n_0(X_2(t))$, we have $n_1(X_1(t)) > n_1(X_2(t))$.

Case 2 $n_1(X_1(t)) < n_0(X_2(t))$: In this case we claim that $n_1(X_1(t)) = n_1(X_2(t))$. Suppose $n_1(X_1(t)) < n_1(X_2(t))$ then this would imply that $n_1(X_2(t))$ has increased by one before $n_1(X_1(t))$ did. But the coupling for this *red and green* pair which was initially at -1 is the classical coupling and only way this could happen is when $n_0(X_2(t)) > n_0(X_1(t))$. Due to the coupling and the conditional argument, this is not true and hence $n_1(X_1(t)) = n_1(X_2(t))$. Under this case, $n_1(X_1(t))$ could be greater than $n_1(X_2(t))$. Again using the same arguments as above we can show that this can not happen.

Using the same arguments as above, conditioned on the statement $n_1(X_1(t)) \geq n_1(X_2(t))$, we can show that $n_2(X_1(t)) \geq n_2(X_2(t))$ and so on. $\square$

Complete Coupling in the Exponential case

In this subsection, the leading particles holding times are exponential with parameters λ_1 and λ_2 respectively for the two processes ($\lambda_1 > \lambda_2$), let the initial distribution in terms of the configuration for $i = 1, 2$ be

$$X_i^0 = \begin{cases} 1 & \text{if } k \leq 0; \\ 0 & \text{if } k > 0. \end{cases} \quad (4.17)$$

Now, we use a special *black-white* coupling (in addition to the *red-green* coupling explained before for other particles) for the leading particles when their holding times are exponential with rate λ_1 and λ_2 respectively to show that the Theorem 4.7 holds in this case.

Lemma 4.4 When the leading particles holding times are exponential with parameter λ_1 and λ_2 ($\lambda_1 > \lambda_2$) respectively, there exists a coupled version of the processes so that for all $t \geq 0$,

$$n_0(X_1(t)) > n_0(X_2(t)). \quad \text{with prob 1}$$

Proof: Firstly, in addition to the classical coupling introduced for the *red and green* pair, now we develop a coupling for the *black and white* pair. We once again emphasize the fact that we do not couple the *red and white* and *black and green* pairs. We will also observe soon that due to the coupling which is to be discussed, the *black and green* pair would not appear at any site.

Let $\{t_i\}$ and $\{s_i\}$ for $i \geq 1$ be two sequences of independent and identically distributed exponential random variables which are independent of each other, representing the successive holding times of the leading particles for the process $X_1(t)$ and $X_2(t)$ respectively with parameter λ_1 and λ_2 . Since t_i and s_i are independent of each other, even though $\lambda_1 > \lambda_2$, in any realization of these sequences, with positive probability $\sum s_i < \sum t_i$, and hence the position of the leading particle $n_0(X_2(t))$ will be ahead of $n_0(X_1(t))$. In order to avoid this, we couple the jump times through a common exponential stream.

Let $\{e_i\}$ be a sequence of independent and identically distributed exponential random variables with rate one. Now construct the holding times $\{t_i^*\}$ and $\{s_i^*\}$ for the leading particles in the respective processes $X_1(t)$ and $X_2(t)$ from the common sequence $\{e_i\}$ such that $t_i^* = e_i/\lambda_1$ and $s_i^* = e_i/\lambda_2$. Since $\lambda_1 > \lambda_2$, with probability one $t_i^* < s_i^*$ and hence the number of renewals or the number of occurrences in time t for the process $X_1(t)$ is larger than that for the process $X_2(t)$ which in turn implies that the position of the leading particle $n_0(X_1(t))$ for the process $X_1(t)$ is always greater than $n_0(X_2(t))$. $\square$

We mention here that the *black-white* coupling is possible only for the case in which

the leading particles holding times are exponential. From Theorem 4.7 and Lemma 4.4, it is easy to conclude the following result which confirms the intuitive reasoning in the beginning of this section about the density profiles

Theorem 4.8 *Let $(\bar{X}_1, \bar{X}_2)$ be the coupled process obtained by using both the red-green and black-red coupling, then under the initial condition $X_1^0 = X_2^0$,*

$$S(X_1(t), k) \geq S(X_2(t), k) \quad \forall t \geq 0, \quad \forall k \in \mathbb{Z}$$

and the hydrodynamic limit versions of the S for the process $X_1(t)$ and $X_2(t)$ defined as

$$S_1(x, t) \stackrel{a.s.}{=} \lim_{h \rightarrow 0} S \left(X_1 \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) \times h$$

$$S_2(x, t) \stackrel{a.s.}{=} \lim_{h \rightarrow 0} S \left(X_2 \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) \times h$$

also satisfies the property

$$S_1(x, t) \geq S_2(x, t) \quad \forall t \geq 0, \quad \forall x \in \mathbb{R}. \quad (4.18)$$

Non-Exponential Case

Let $X_1(t)$ and $X_2(t)$ be two simple exclusion processes with exponential and arbitrary holding times respectively for the leading particles at rate λ_1 and λ_2 ($\lambda_1 > \lambda_2$). Starting with the initial configuration X_1^0 and X_2^0 , we would like to develop monotonicity properties as done before for the coupled process $\bar{X}(t)$. Since the leading particles holding times are differently distributed, we can not use the *black-white* coupling introduced in the exponential case for the leading particles. Even though $\lambda_1 > \lambda_2$, we can not expect $n_0(X_1(t))$ to be larger than $n_0(X_2(t))$ for all $t \geq 0$. With positive probability, we could

have $n_0(X_1(t)) < n_0(X_2(t))$ for some t . Since $n_0(X_1(t))$ is a Poisson process with parameter λ_1 and $n_0(X_2(t))$ is a renewal process with parameter λ_2 , we can use the Strong Law of Large Numbers to conclude that there exists a t_0 such that for all $t \geq t_0$, we have $n_0(X_1(t)) > n_0(X_2(t))$. This is obvious since $\lambda_1 > \lambda_2$.

As in the exponential case (Theorem 4.7), it is not true that $\bar{k}_0(t) = -\infty$ and $k_0^*(t) = \infty$ for all $t \geq t_0$. The main difficulty here is that even though for all $t \geq t_0$, $n_0(X_1(t)) > n_0(X_2(t))$, in contrast to the exponential case, we may not have the ordering for the other particles, i.e., $n_k(X_1(t))$ for $k > 0$ and $t > t_0$ may not be ordered at all. For example, $n_1(X_1(t))$ could be behind $n_1(X_2(t))$, i.e., $n_1(X_1(t)) < n_1(X_2(t))$. This is due to the fact that $n_0(X_1(t))$ could be less than $n_0(X_2(t))$ with positive probability in the time interval $[0, t_0)$. Note that this situation was not allowed to happen in the exponential case due to the coupling of the leading particles.

What we would like to show is that even though under this situation, we could still have $n_k(X_1(t)) < n_k(X_2(t))$ for some $k = 2, 3, \dots$ (note that n_k denote the location of the k^{th} particle), in hydrodynamic limit time scale,

$$\lim_{h \rightarrow 0} S \left(X_1 \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) \times h \geq \lim_{h \rightarrow 0} S \left(X_2 \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) \times h$$

i.e. the density profiles are ordered in terms of the areas. We have the following theorem:

Theorem 4.9 *Let $X_1(t)$ and $X_2(t)$ be two asymmetric simple exclusion processes with exponential (λ_1) and arbitrary (λ_2) holding times for the leading particles respectively, where the arbitrary holding times ($\lambda_2 < \lambda_1$) satisfy the Strong law of Large numbers given in (*). Also let $(\bar{X}_1, \bar{X}_2)$ be the coupled process obtained by the red-green coupling alone, then under the initial distribution $X_1^0 = X_2^0$,*

$$P \{S_1(x, t) \geq S_2(x, t)\} = 1 \quad (4.19)$$

Proof: Let E_T be the event that the leading particles in the two processes namely, $X_1(t)$ and $X_2(t)$ respectively are ordered, i.e., $n_0(X_1(t)) > n_0(X_2(t))$ for all $t \geq T$. Then by the Strong Law of Large Numbers,

$$\lim_{h \rightarrow 0} P(E_{\log(1/h)}) = 1. \quad (4.20)$$

Note that as we mentioned earlier, $n_k(X_1(t))$ could be less than $n_k(X_2(t))$ for some k at this time. Therefore, it is not true that

$$S(X_1(t), k) \geq S(X_2(t), k) \quad \forall k \in \mathcal{Z} \quad \text{and} \quad t \geq \log(1/h).$$

However,

$$\begin{aligned} & P \left\{ S \left(X_1 \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) \times h \geq S \left(X_2 \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) \times h - \epsilon \right\} \\ & \geq P \left\{ S \left(X_1 \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) \times h \geq S \left(X_2 \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) \times h - \epsilon \mid E_{\log(1/h)} \right\} \\ & \quad \times P \{ E_{\log(1/h)} \} \end{aligned} \quad (4.21)$$

Let $\hat{S}(\hat{X}_1(\cdot), \cdot)$ and $\hat{S}(\hat{X}_2(\cdot), \cdot)$ be the corresponding quantities for the approximating processes $\hat{X}_1^h(t)$ and $\hat{X}_2^h(t)$ respectively, where $\hat{X}_1^h(t)$ and $\hat{X}_2^h(t)$ are obtained by filling up all the holes the left of the leading particles at time $\log(1/h)$.

Since the leading particle in $X_1(t)$ is renewal (Poisson when $i = 1$) with mean $\lambda, \log(1/h)$, at most $\#_h(i)$ ($\leq n_0(X_1(\log(1/h)))$) are needed to fill all the holes to the left of the leading particles. The processes $\hat{X}_1(t)$ and $\hat{X}_2(t)$ are exactly the similar to the situation in Theorem 4.7 and 4.8 from time $\log(1/h)$ onwards, and therefore, as $h \rightarrow 0$,

$$P \left\{ \hat{S} \left(\hat{X}_1 \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) \times h \geq \hat{S} \left(\hat{X}_2 \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) \times h \mid E_{\log(1/h)} \right\} = 1 \quad (4.22)$$

But

$$\hat{S} \left(\hat{X}_i \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) = S \left(\hat{X}_i \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) + \#_h(i) \quad (4.23)$$

Therefore, for any $\epsilon > 0$, the R.H.S. of (4.21) using (4.23) becomes

$$\begin{aligned}
& P \left\{ S \left(X_1 \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) \times h \geq S \left(X_2 \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) \times h - \epsilon \middle| E_{\log(1/h)} \right\} P \{ E_{\log(1/h)} \} \\
& = P \left\{ \hat{S} \left(\hat{X}_1 \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) \times h \geq \hat{S} \left(\hat{X}_2 \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) \times h + h \#_h(1) - h \#_h(2) - \epsilon \middle| E_{\log(1/h)} \right\} \\
& \quad P \{ E_{\log(1/h)} \} \\
& \geq P \left\{ \hat{S} \left(\hat{X}_1 \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) \times h \geq \hat{S} \left(\hat{X}_2 \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) \times h, h \#_h(1) - h \#_h(2) \leq \epsilon \middle| E_{\log(1/h)} \right\} \\
& \quad P \{ E_{\log(1/h)} \} \\
& \geq P \left\{ \hat{S} \left(\hat{X}_1 \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) \times h \geq \hat{S} \left(\hat{X}_2 \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) \times h \middle| E_{\log(1/h)} \right\} P \{ E_{\log(1/h)} \} \\
& \quad - P \left\{ h \#_h(1) - h \#_h(2) > \epsilon \middle| E_{\log(1/h)} \right\} P \{ E_{\log(1/h)} \} \tag{4.24}
\end{aligned}$$

Now note that because

$$\lim_{h \rightarrow 0} h \times n_0(X_i(\log(1/h))) = 0 \quad \text{in Prob.}$$

so does

$$\lim_{h \rightarrow 0} h \times \#_h(i) = 0 \quad \text{in Prob.}$$

Since

$$S_i(x, t) = \lim_{h \rightarrow 0} S \left(X_i \left(\frac{t}{h} \right), \left[\frac{x}{h} \right] \right) \times h,$$

therefore,

$$P \{ S_1(x, t) \geq S_2(x, t) \} = 1$$

□

Having established the stochastic ordering property in terms of the density profiles (areas) for the processes $X_1(t)$ and $X_2(t)$ when the leading particles holding times are exponential and one of the leading particle's holding time is exponential and other one is renewal, now consider three asymmetric simple exclusion processes $X_1(t)$, $X_2(t)$ and

$X_3(t)$ with exponential (rate ϵ), arbitrary renewal (rate λ) and exponential (rate $\lambda + \epsilon$) holding times for the leading particles respectively (all other particle's holding times in the three processes are exponential with rate one). The following theorem is an immediate consequence of Theorem 4.8 and 4.9 for all $\epsilon > 0$.

Theorem 4.10 Under the initial distribution $X_1^0 = X_2^0 = X_3^0$, with probability one,

$$S_1(x, t) \geq S_2(x, t) \geq S_3(x, t)$$

where S_i corresponds to the process X_i .

From Theorem 4.10, it is easy to see that the density profile for the asymmetric simple exclusion process with arbitrary holding time for the leading particle is bounded above and below (in terms of the areas) by the known density profiles for the *Road Hog* processes. Even though, we have not shown directly the existence of the density profile in this case, we conclude at this point that the density profile is the same as that of the *Road Hog* process. This in turn implies that the density profile for the infinite series of exponential queues with non-Poisson input is the same as the one proved in Section 4.3.

4.6 Density Profile and Hydrodynamic Limit for a Feedback Exponential Queues in Series

In the last few sections, we considered the extreme (totally asymmetric) simple exclusion process as a model for the exponential queues in series. In this section, we modify this simple exclusion process (see Section 4.1) by letting the particles jump to the right or to the left (with exclusion) after mean one exponential holding time with probability $1/2 < p \leq 1$ instead of $p = 1$, i.e. particles only move to the right. This simple exclusion process corresponds to feedback queues in series which will be explained in Section 4.6.1. In Section 4.6.2, we provide the density profile and hydrodynamic results for this network.

4.6.1 The Feedback Queues in Series

Consider an infinite number of exponential queues which are in series. Customers arrive to the first queue (initially all the queues are empty) according to an arbitrary arrival process. A customer, after completing service (exponential with rate one) at the first queue, proceeds to the second queue with probability p and leaves the network with probability $(1 - p)$. The customer who arrived at the second queue, proceeds to the third queue after completing its service (again exponential with rate one) at the second queue with probability p or goes back to the first queue with probability $(1 - p)$. Similarly, a customer after completing its service at the third queue proceeds to the fourth queue with probability p or goes back to the second queue with probability $(1 - p)$. This kind of feedback continues at each queue. Note that the arrival process to the first queue includes the internal arrivals from the second queue. Also note that customers can only leave from the first queue after completing their service. See Fig.4.7 for the illustration of this feedback mechanism.

Figure 4.7

Feedback Queues in Series

Let $X(t)$ be the original simple exclusion process with the transition probability

$1/2 < p \leq 1$. We start this process under the initial distribution $\nu^{1,(1-\lambda)}$ which is defined as

$$\nu^{1,(1-\lambda)} = \begin{cases} 1 & \text{if } k \leq 0; \\ (1-\lambda) & \text{if } k > 0 \end{cases}$$

As in section 4.3, let the particles initially in positions $k \leq 0$ be colored red and $k > 0$ be colored black. Let $X_1(t)$ corresponding to the red particles be the *embedded process* in $X(t)$. The red particle initially at zero is the *tagged particle* in $X(t)$ which is also the leading red particle in $X_1(t)$. Since we are considering feedback queues in series, the leading particle describes part of the input process to the first queue which is initially empty.

As before, let $Y(k, t)$ be the inter particle distances between the k^{th} and $(k+1)^{\text{th}}$ red particles. The queue length at the first queue which is given by $Y(1, t)$ increases due to arrival when the tagged particle (red) jumps to the right with probability p after an exponential amount of time with rate one. See Fig 4.8.

Figure 4.8

Flow Through Queues Represented by Particles

Unlike the simple exclusion process considered in the last sections, $Y(1, t)$ can decrease either when the tagged particle jumps to the left with probability $(1-p)$ or when the

particle which was originally at (-1) jumps to the right with probability p . The former case corresponds to a departure from the first queue and the later to an arrival at the second queue with probability p . Note that initially, unless the second particle moves to the right with probability p , it can not go backwards (exclusion interaction). In general, the inter particle distance $Y(k, t)$ and hence the queue length at k increases either due to an arrival from the $(k-1)^{th}$ queue or the $(k+1)^{th}$ queue. Similarly, $Y(k, t)$ decreases due to a service completion and hence an arrival to either $(k+1)^{th}$ queue or $(k-1)^{th}$ queue.

Once again, unlike the $p = 1$ case, the movement of the individual particles, does not necessarily represent the service completion epochs. For example, when $p = 1$, the motion of the second particle (to its right) always represented a service completion at the first queue; whereas here, the motion of the tagged particle to its left, as well as the motion of the second particle to its right can represent service completion at the first queue after exponential rate one holding times.

As explained in Section 4.3, the purpose of distributing particles to the right of zero with probability $(1 - \lambda)$ is to create the arrival process to the first queue. Here, the presence of red particles (to the left of black particles) does affect the motion of the black particles. Especially, the red tagged particle motion is affected by the black particles as well as the red particles. This is due to the fact that $1/2 < p \leq 1$. Unlike the case $p = 1$, the motion of the tagged particle is not Poisson (as it was in the $p = 1$ case), it is much more complicated. For example, when we distribute particles initially with probability $(1 - \lambda)$ at all sites except the site $k = 0$ where the tagged particle is present, the motion of the tagged particle is only known to satisfy the following Strong Law of Large Numbers (assuming X_t denotes the position of the tagged particle) (Kipnis(1986)):

$$\frac{X_t}{t} \xrightarrow{a.s.} (2p - 1)\lambda.$$

Therefore, the motion of the *tagged particle* (not Poisson) to the right as well as the motion of the second particle to the left constitute the arrival process to the first queue. This is the reason, why we mentioned in the beginning of this section that the arrival process to the first queue is general. In the next section, we use the results of Benassi and Fouque(1987) to obtain the density profile and the hydrodynamic limit for this network.

4.6.2 Density Profile and Hydrodynamic Limit

We first state the results for the process $X(t)$ (red and black particles) which is due to Benassi and Fouque(1987) and then identify the density profile for the embedded process $X_1(t)$.

Theorem 4.11 *The hydrodynamic limit result stated in Theorem 4.3 holds for the process $X(t)$, when $1/2 \leq p \leq 1$ with the following density profile:*

$$u(x, t) = \begin{cases} 1 & \text{if } x \leq -(2p-1)t \\ \frac{1}{2}(1 - \frac{x}{(2p-1)t}) & -(2p-1)t \leq x \leq (2p-1)(2\lambda-1)t \\ (1-\lambda) & x \geq (2p-1)(2\lambda-1)t \end{cases}$$

For the embedded process (corresponding to the red particles) $X_1(t)$ in $X(t)$, with $1/2 < p \leq 1$, under the initial distribution $\nu^{1, (1-\lambda)}$, we have the following result.

Theorem 4.12 *The hydrodynamic limit result stated in Theorem 4.3 holds for the embedded process $X_1(t)$ when $1/2 \leq p \leq 1$ with the following density profile:*

$$u(x, t) = \begin{cases} 1 & \text{if } x \leq -(2p-1)t \\ \frac{1}{2}(1 - \frac{x}{(2p-1)t}) & -(2p-1)t \leq x \leq (2p-1)(2\lambda-1)t \\ (1-\lambda) & (2p-1)(2\lambda-1)t \leq x \leq (2p-1)\lambda t \\ 0 & x \geq (2p-1)\lambda t \end{cases}$$

Proof: As in Section 4.3, the Benassi and Fouque(1987) result corresponds to the process $X(t)$ (red and black particles) where the density profile $u(x, t) = (1-\lambda)$ for all $x \geq$

$(2p-1)(2\lambda-1)t$. The result follows from the next Lemma 4.5 which identifies the position of the leading red particle. $\square$

Lemma 4.5 *The limiting position of the tagged particle under the initial distribution $\nu^{1,(1-\lambda)}$, when $1/2 \leq p \leq 1$ is $(2p-1)\lambda t$.*

Proof: We first note that due to exclusion interaction, the order of the particles are preserved in time, i.e., the red particles are to the left and the black particles are to the right. In Theorem 4.11, the density profile is linear in the region $-(2p-1)t \leq x \leq (2p-1)(2\lambda-1)t$. But initially, the density profile was one to the left of zero and $(1-\lambda)$ to the right of zero. Moreover, the position of the constant density profile is at $(2p-1)(2\lambda-1)t$. Since the number of red particles which have moved to the right of zero is $((2p-1)t)/4$ and this must be equal to the number of red particles to the right of zero, the position of the leading red particle is at $(2p-1)\lambda t$ (Note that $((2p-1)t)/4 = ((2p-1)\lambda t - (2p-1)(2\lambda-1)t) \times (1-\lambda) + (3-2\lambda) \times (2p-1)(2\lambda-1)t$).

Remark: Lemma 4.5 implies that the position of the tagged particle under the initial distribution $\nu^{1,(1-\lambda)}$ satisfies the same Strong Law of Large Numbers mentioned above for the tagged particle under the initial distribution $\nu^{(1-\lambda)}$. Since $1/2 \leq p \leq 1$, the particles to the left of zero do not influence the motion of the tagged particle.

Fig.4.9 illustrates this density profile for $p > 1/2$ and $\lambda > 1/2$.

1. When $p = 1$, the profile obtained here is the same as the one given in Section 4.3 and when $p = 1$ and $\lambda = 1$ we get the profile given in Section 4.2. In the case of $\lambda = 1$ and $1/2 < p \leq 1$, the density profile is very similar to the one obtained in Section 4.2 but the leading particle is at $(2p-1)t$ instead of at t .
2. The leading front (where the profile is constant) in this case is much shorter than the one given in Section 4.3. Note that the stretch of the leading front when $p = 1$ is

Figure 4.8

Density Profile for the Feedback Queues in Series

$(1 - \lambda)t$ whereas here the stretch is $(2p - 1)(1 - \lambda)t$

3. **Theorem 4.13** *In the hydrodynamic scale, the first $(2p - 1)(1 - \lambda)^2 t$ queues are in equilibrium i.e., in the first $(2p - 1)(1 - \lambda)^2 t$ queues, the queue lengths are independent, identical and geometrically distributed with parameter $(1 - \lambda)$.*

Proof: Similar to Theorem 4.5.

4. Though the tagged particle's motion does not explicitly represent the arrival process to the first queue, as we mentioned earlier, the arrival process is quite general. Even for this complex arrival process, with the feedback mechanism, the density profile is very similar to the one obtained in Section 4.3.

REFERENCES

- Andjel, E.D., "Invariant Measures for the Zero-Range Process", *Ann. Prob.*, Vol. 10, 1982, pp 525-547.
- Andjel, E.D. and C. Kipnis, "Derivation of the Hydrodynamical Equation for the Zero-Range Interaction Process", *Ann. Prob.*, No. 2, 1984, PP 325-334.
- Avi-Itzhak, B. and D.P. Heyman, "Approximate Queueing Models for Multi-programming Computer Systems", *Oper. Res.*, Vol. 21, 1973, pp 1212-1230.
- Benassi, A. and J.P. Fouque, "Hydrodynamical Limit for the Asymmetric Simple Exclusion Process", *The Annals of Prob.*, Vol. 15, No. 2, 1987, pp 546-560.
- Buzacott, J.A. and J.G. Shanthikumar, "On Approximate Models of Dynamic Job Shops", *Mgmt. Sci.*, Vol. 31, No. 7, 1985, pp 870-886.
- Chaudhry, M.L. and J.G.C. Templeton, *A First Course in Bulk Queues*, New York: John Wiley & Sons, 1983
- Cohen, J.W., *The Single Server Queue*, Amsterdam: North Holland, 1969.
- Conolly, B.W., *Lecture Notes on Queueing Systems*, Sussex: Ellis Horwood Ltd., 1975.
- Cooper, R.B., *Introduction to Queueing Theory*, Amsterdam: North-Holland, 1981.
- Cox, D.R. and D.V. Hinkley, "Some Properties of Multi-Server Queues with Appointments", *J.R. Stat. Soc. Ser. A*, Vol. 133, 1970, pp 1-13.
- Daley, D.J., "Queueing Output Process", *Adv. Appl. Prob.*, Vol. 8, 1976, pp 395-415.
- Dubois, D., "A Mathematical Model of a Flexible Manufacturing System with Limited In-Process Inventory", *Eur. Jnl. of Oper. Res.*, Vol. 14, No. 1, 1983, pp 66-78.
- Federgruen, A. and H.C. Tijms, "Computation of the Stationary Distribution of the Queue Size in an $M/G/1$ Queueing System with Variable Service Rate", *J. Appl. Prob.*, Vol. 17, 1980, pp 515-522.

Gorden, W.J. and G.F. Newell, "Closed Queueing Systems with Exponential Servers", *Oper. Res.*, Vol. 15, No. 2, 1967, pp 254-265.

Gross, D. and C.M. Harris, *Fundamentals of Queueing Theory*, 2nd Edition, New York: Wiley, 1985.

Hillier, F.S. and O.S. Yu, *Queueing Tables and Graphs*, New York: North-Holland, 1981.

Jackson, J.R., "Job-Shop like Queueing Systems", *Mgmt. Sci.*, Vol. 10, No. 1, 1963, pp 131-142.

Kapur, J.N. and V. Kumar, "A Comparison of Performance Measures for Closed and Semi-Open Network of Queueing Models of Flexible Manufacturing Systems", WPS 84-16, School of Business, Carleton University, 1984.

Kelly, F.P., *Reversibility and Stochastic Network*, New York: Wiley Interscience, 1976.

Kipnis, C., "Central Limit Theorems for Infinite Series of Queues and Applications to Simple Exclusion", *The Annals of Prob.*, Vol. 14, No. 2, 1986, pp 397-408.

Konheim, A.G. and M. Reiser, "Finite Capacity Queueing Systems with Applications in Computer Modelling", *SIAM. J. Computing*, Vol. 7, No. 2, 1978, pp 210-229.

Kramer, W. and M. Langenbach-Belz, "Approximate Formula for the Delay in the Queueing System $GI/G/1$ ", *Congresbook*, 8th Internat. Teletraffic Congress, Melbourne, 1976, pp 235.1-235.8.

Liggett, T.M., *Interacting Particle Systems*, New York: Springer Verlag, 1985.

Page, E., *Queueing Theory in O.R.*, *Opnl. Res. Series*, 1972.

Prabhu, N.U., *Queues and Inventories*, New York: John Wiley & Sons, 1965.

Potier, D., *QNAP2: Introduction to QNAP2*, INRIA, 1984.

Rost, H., "Non-Equilibrium Behaviour of a Many Particle Process: Density Profile and Local Equilibria", *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, Vol. 58, 1981, pp 41-53.

Shanthikumar, J.G. and R.G. Sargent, "A Hybrid Simulation/Analytic Model of a Computerised Manufacturing System", *Proce. Internat. Conf, Oper. Res.*, 1981, pp 901-915.

Stidham, S., JR., "Regenerative Processes in the Theory of Queues with Applications to the Alternating Priority Queue", *Adv. Appl. Prob.*, Vol. 4, 1972, pp 542-577.

Tijms, H.C., *Stochastic Modelling and Analysis: A Computational Approach*. New York: John Wiley & Sons, 1986.

Whitt, W., "Open and Closed Models for Networks of Queues", *AT & T Bell. Labs. Tech. Jnl.*, Vol. 63, No. 9, 1984, pp 1911-1979.

Wolf, R.W., "Poisson Arrivals See Time Averages", *Oper. Res.*, Vol 30, 1982.

END

06|07| 89|