PROBLEMS IN FEEDBACK QUEUEING SYSTEMS WITH SYMMETRIC QUEUE DISCIPLINES

by

Georgia-Ann Klutke

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Industrial Engineering and Operations Research

APPROVED:

R. L. Disney, Chairman

I. M. Besieris

R. D. Foley

T. L. Herdman

Y. D. Mittal

August, 1986

Blacksburg, Virginia

PROBLEMS IN FEEDBACK QUEUEING SYSTEMS WITH SYMMETRIC QUEUE DISCIPLINES

by

Georgia-Ann Klutke

Committee Chairman: Ralph L. Disney
Department of Industrial Engineering and Operations Research

(ABSTRACT)

In this paper we study properties of a queue with instantaneous Bernoulli feedback where the service discipline is one of two symmetric disciplines. For the processor sharing queue with exponentially distributed service requirements we analyze the departure process, imbedded queue lengths, and the input and output processes. We determine the semi-Markov kernel of the internal flow processes and compute their stationary interval distributions and forward recurrence time distributions. For generally distributed service times, we analyze the output process using a continuous state Markov process. We compare the case where service times are exponentially distributed to the case where they are generally distributed. For the infinite server queue with feedback, we show that the output process is never renewal when the feedback probability is non-zero. We compute the time until the next output in three special cases.

ACKNOWLEDGEMENTS

I wish to express my gratitude and appreciation to Professors Ioannis Besieris, Terry Herdman, Robert Foley and Yashaswini Mittal for serving on my committee. In their courses and in subsequent contacts during my graduate career, they have always given me useful and instructive criticism.

I would like to thank both Professors Peter Kiessler and Donald McNickle for the many conversations on this work and the intellectual "jump starts" they provided.

typed this dissertation. I thank her for producing a fine document on a very tight schedule. Moreover, I thank her for making it appear that the hard part of the day was deciding where to go for lunch.

The problem analyzed in this dissertation was suggested by the chairman of my committee, Professor Ralph Disney. Professor Disney has an international reputation, but perhaps only his students know the amount of time and effort he spends in nurturing young researchers in the field of applied probability. I am fortunate to have had the opportunity to work closely with him and thank him for his contribution to my professional development.

My family has always been a source of encouragement and inspiration to me. I have been hard put to compete with their accomplishments. This work is dedicated to my mother and father,

, , , and , and to , whose spirit for life has convinced me that any hurdle can be overcome.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Background

Queueing networks have by now become standard models for analyzing the random behavior of complicated systems. They are widely used in computer applications, flexible manufacturing systems, biological and ecological studies, teletraffic engineering and data transmission, and complex machine repair problems, to give an incomplete list. Operations management in many midsize companies today routinely involves some queueing network expertise, be it in the form of a software package for network analysis or staff trained in analysis techniques. If merely measured by its broad practical usage, queueing network theory has proved itself a valuable tool.

From an analytical point of view, however, the picture is somewhat different. Theoretical results for networks are still quite sparse. While more work is being done on approximating systems, many open problems remain in describing the probabilistic behavior of general queueing networks.

1.2 Simple Queues and Queueing Networks

Basically, a queueing network is a collection of queues (called nodes) through which customers (or units) travel. It will therefore be useful to begin with a description of a simple queue before we put these simple queues together into a network.

A queue is a service system. It consists of one or more servers who satisfy the demands of arriving units. The system has space for units that are being serviced and space for those whose demand cannot

be satisfied immediately. Each of these spaces may hold a finite or infinite number of units, and the waiting room for those not in service may even have zero capacity. The system has a rule (a service discipline) to decide which of the demanding units are chosen for service and at what rate these units are serviced. The most commonly analyzed service discipline is first-come, first-served, which selects the earliest arriving unit among those present to receive service. Other systems have priorities for determining how units wait in line. In some systems, all units receive a fraction of the server's effort simultaneously.

Most of queueing theory assumes that units arrive for service at random points in time and carry with them random service requirements. We call the sequence of arrival times $\{T_m, m = 1,2,\ldots\}$ the (random) arrival process and the sequence of service times (or service requirements) $\{S_m, m = 1,2,\ldots\}$ the service process. In this work we will assume that the arrival process is a point process with independent and identically distributed interarrival intervals, and that the service process is a sequence of independent and identically distributed random variables. The particular assumptions we make on the distributions of these random variables are given in Chapter 2.

From these basic processes, we are interested in studying many derived processes in the queue. The queue length process charts the number of units in the system at all points in time during which the system operates. The output process is the sequence of times at which a service completion occurs. The departure process is the sequence of times at which units leave the system. There are many other processes

of interest; these are the ones with which we will be most concerned in this paper. These processes may be studied as functions of time, for example, when the system starts empty at time zero. They may also be studied as time becomes infinitely large. Under certain conditions the processes of interest may possess limiting probability distributions. These are called equilibrium or stationary distributions. Most of queueing theory is concerned with the limiting properties of the processes of interest.

A queueing network is a collection of simple queueing systems linked together in some fashion by arcs via which units travel without delay. To the network belongs a switching rule that routes units between nodes in some prescribed (possibly random) order. Each node may have a stream of arrivals from outside the network (external arrivals), and because of the switching mechanism, a stream of arrivals from inside the network (internal arrivals). In addition, the stream of units completing service at a node consists of units travelling to other nodes in the network (internal departures) and units leaving the system entirely (external departures). Units are allowed to return immediately for more service at a node from which they just departed; the stream of units returning immediately is called the instantaneous feedback stream.

The processes of interest in queueing networks are related to but obviously not equivalent to the processes of interest in simple queueing systems. The queue length process is a vector-valued process that charts the number of units in each queue over time. The output process at each node is the sequence of times at which customers

complete service at a node, and the departure process at a node is the sequence of times at which units complete service and simultaneously leave the network. The output process of the network is the sequence of times at which a service completion occurs at any node and is thus the (reordered) superposition of the output process at each node. Likewise, the departure process of the network is the sequence of times at which a customer completes service and leaves the network at any node.

1.3 Jackson Networks

J. R. Jackson (1957) is usually credited with instigating the study of the properties of queueing networks. The Jackson network consists of $J < \infty$ nodes, some of which are fed by exogenous Poisson arrival streams independent of the state of the network and of each other. Service requirements at each node are assigned according to a exponential distribution function, and the service requirements of a unit at a node is assigned independently of the service requirements of other units at any node in the network. A unit leaving a node enters another node or leaves the network according to a fixed matrix of probabilities, and these probabilities are independent of the state of the network at that instant.

Jackson's results were among the first to examine properties of a network that were different from properties of a node in isolation. He found that the limiting vector-valued queue length distribution behaved as if the queue lengths at each node were independent random variables, each of which being the queue length of a simple s-server queue with an appropriately scaled arrival parameter. This result was indeed

surprising as it became clear that although the limiting joint queue length had a relatively simple structure, the internal flow processes were indeed complicated; in general, they did not even possess independent intervals. Much work has subsequently been done to describe these internal traffic processes.

1.4 Service Disciplines

Most of the early queueing work was done with systems that operated on a first-come, first-served basis. With the increase in computer and teletraffic technology, different methods of assigning the service resource to units needing service were employed. The goal here was to optimize certain performance measure of the system, such as the expected wait in the queue conditioned on the required service time. One early discipline used in time sharing computer networks is known as the round robin discipline. In the round robin discipline, units arrive at the service center with a service requirement and take their place at the tail of the line. The server works at the head of the line, and spends a fixed quantum of time servicing the unit in the first position. At the end of this quantum of time, the unit has either exhausted its service requirement or needs more service. If its service need is met, it leaves the system. If it needs more service it is placed at the tail of the queue and all other units move up one space. The server then services the new unit at the head of the queue for a fixed quantum of time, and the process continues. An advantage of this discipline is that a unit requiring only a small amount of service is not slowed down as much as in first-come, first-served by units with large service requirements that arrived earlier than it.

When the quantum of time spent on each unit is allowed to shrink to zero, all units in the queue are effectively receiving service simultaneously, but are sharing the server equally. This discipline is known as processor sharing and is used in many computer applications as a model for time sharing of a device. Although it is not a physically realizable discipline, it does provide useful approximations to round robin performance measures.

A great advantage of the processor sharing service discipline is that it appears to be robust to changes in the distribution of service requirements when the mean service requirement is held constant. Many performance measures, including the queue length distribution and the expected length of time in the system conditioned on the required amount of service depend on the service requirement distribution only through its mean. Moreover, this discipline in equilibrium transforms a Poisson arrival stream into a Poisson departure stream irrespective of the service time distribution. These results are useful from a practical point of view, since they rely little on any assumptions about the service requirement distribution. These properties of the system are said to be insensitive to the service time distribution given its mean.

Another class of queueing systems that exhibits many of the same insensitivity properties as the processor sharing queue is the class that operates as if each unit has its own dedicated server. These are known as infinite server queues, and seem to have been used primarily in telephony and biological models.

Mathematical biologists often model the major organs and systems

of the body as compartments (each of which acts as an infinite server queue) that may exchange material among themselves. Service requirements are thought of as holding times of material in the different compartments. With a transfer function (switching process) defined between the individual compartments (nodes), the compartmental system can be thought of as a queueing network. If we think of an external arrival process as modelling the injection into the blood stream of a chemical dye or radioactive isotope, the queue length process would measure the amount of material residing in each compartment (eg. blood, liver, brain, etc.) over time.

Analytically, these infinite server queues have limiting queue length distributions that depend on the service time distribution only through the mean service time (i.e., they have the insensitivity property). They also have Poisson departure processes in equilibrium when the arrival process is Poisson. The first-come, first-served discipline does not have these insensitivity properties when there are only a finite number of servers available.

1.5 Purpose of This Research

The purpose of this paper is to examine the properties of the processor sharing and infinite server queueing disciplines in a network context. We are primarily concerned with analyzing the internal flow processes between nodes in a network, and in particular, in determining how the flow processes depend on the service requirement distribution. Our main goal is to understand how networks of processor sharing queues and infinite server queues are similar to or different from a collection of isolated processor sharing or infinite server nodes.

This paper analyzes a simple queueing network with an instantaneous feedback loop. Although this network is not as general as could be, it possesses the important features whose properties we wish to study. We can define distinct flow processes of external arrivals, internal arrivals, returning units, internal outputs and external outputs. We can define queue length processes fixed at the times at which a unit travels in any of the streams above. And, perhaps most importantly, the feedback queue is, for our purposes, a tractable system. Section 2.4 explicitly defines the system we discuss here and the processes of interest to us. From the results of this analysis, we provide a beginning for the study of traffic processes in more general networks of processor sharing and infinite server queues.

1.6 Organization

This paper consists of 7 chapters, a table of contents, bibliography and various supporting pages. Each chapter is divided into sections. Within a chapter, theorems, lemmas and corollaries are numbered consecutively, as are figures. The chapter number is included in this numbering scheme; eg. Figure 4.3 is the third figure that appears in Chapter 4. When necessary, equations are referenced consecutively by a number in parentheses to the right of the equation. These numbers use the same scheme as for figures or theorems. Throughout the paper, references to material appearing elsewhere will include the chapter number and, when necessary, section and subsection number of the relevant passages. Bibliographic references include the surname of the author(s) and the publication date of the reference; the bibliography is arranged alphabetically.

CHAPTER 2

DEFINITIONS, NOTATION AND PRELIMINARIES

2.1 Introduction

In this section, we introduce the structure of the system we wish to analyze and provide the notation we will use. We begin with a discussion of order relations between random variables. We illustrate these orderings with several examples. We discuss the ideas of reverse processes and reversibility which will be of use to us in succeeding chapters. We then introduce the particular network that we will analyze, beginning with an informal description and then making the description formal. Finally, we discuss the concept of a symmetric queueing discipline and formally define the processor sharing discipline. Throughout this paper, we will use the notation A/B/C/D to refer to a particular queueing system. As is standard, A describes the arrival process, B describes the service process and C is the number of servers. We will use D to refer to the service discipline. For the arrival process, A = M when the interarrival times are independent exponentially distributed random variables. For the service process, B = M or GI according to whether service times are independent and exponentially distributed or independent and generally distributed, respectively. When the discipline is processor sharing, D = PS; when it is first-come, first-served, we will suppress the descriptor D. In addition we will denote instantaneous feedback by tacking on "-IBF" to the descriptor.

2.2 Partial Order Relations Among Random Variables

One of our goals in this research is to compare random processes

in queueing systems with different parameters. At the least, such a comparison involves comparing two or more random variables. In this section, we will discuss two partial orderings which provide the tools for making such comparisons, namely stochastic ordering and convex ordering. Our principle reference for this section is Stoyan (1983).

Definition 2.1. The random variable X is stochastically smaller (or smaller in distribution) than the random variable Y, written

$$X <_d Y$$

if their respective distribution functions F and G satisfy

$$F(x) \geq G(x) \quad \text{for all } x \in \mathbb{R} .$$

Stoyan has shown that stochastic order as given in Definition 2.1 is equivalent to requiring $E(f(X)) \leq E(f(Y))$ for all nondecreasing real functions f on $\mathbb{R}$. Taking $f(x) = x^n$ we have that $X <_d Y$ implies that all moments around zero of X are smaller than the corresponding moments of Y. The converse is not true.

Example 2.1. Let X and Y have distribution functions F and G, where

$$F(x) = 1 - e^{-\lambda x}, \quad G(x) = 1 - e^{-\mu x}, \quad x \geq 0$$

and $0 < \lambda < \mu$. Then $X <_d Y$.

The order of the moments is not sufficient to insure that the random variables are ordered in distribution, as the following example indicates.

Example 2.2. Let X and Y have distribution functions F and G, where (for $\lambda > 0$)

$$F(x) = \int_{-\infty}^{x} \frac{e^{-\frac{(t-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}\,dt \qquad G(x) = \begin{cases} 1 - e^{-\lambda x} & x > 0 \\ 0 & x < 0 \end{cases}$$

Now for $\mu = 0$, $\sigma^2 = 1$ and $\lambda < 1$, $E(X^n) < E(Y^n)$ for all n, but $X \not\leq_d Y$.

Definition 2.2. The random variable X is convexly smaller than the random variable Y, written

$$X <_c Y$$

if their respective distribution functions F and G satisfy

$$\int_x^\infty (1 - F(t))dt < \int_x^\infty (1 - G(t))dt \quad \text{for all } x,$$

or equivalently, if

$$E(\max(0,X-x)) < E(\max(0,Y-x)) \quad \text{for all } x$$

provided these integrals (equivalently, expectations) are finite.

Definition 2.2 is equivalent to requiring $E(f(X)) < E(f(Y))$ for all nondecreasing convex real functions f on $\mathbb{R}$.

Example 2.3. (Stoyan (1983)). For any random variable X with finite mean E(X),

$$\max(x,E(X)) < E(\max(x,X))$$

since $\phi(\cdot) = \max(x,\cdot)$ is convex. Thus

$$E(X) <_c X.$$

Clearly $X \leq_d Y$ implies $X \leq_c Y$ provided $E(\max(0,Y)) < \infty$. The converse is not true, as the following example indicates.

Example 2.4. Let X and Y have distribution functions F and G, where

$$F(x) = \begin{cases} 0 & x < 1 \\ 1 & x > 1 \end{cases} \qquad G(x) = \begin{cases} 0 & x < \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} < x < \frac{3}{2} \\ 1 & x > \frac{3}{2} \end{cases}$$

Then

$$\int_x^\infty \bar{F}(t)dt = \begin{cases} 1 - x & 0 < x < 1 \\ 0 & x > 1 \end{cases} \qquad \int_x^\infty \bar{G}(t)dt = \begin{cases} 1 - x & 0 < x < \frac{1}{2} \\ \frac{3}{4} - \frac{1}{2}x & \frac{1}{2} < x < \frac{3}{2} \\ 0 & x > \frac{3}{2} \end{cases}.$$

Hence $X \leq_c Y$, but $X \not\leq_d Y$.

Thus stochastic order is stricter than convex order. Both $X \leq_d Y$ and $X \leq_c Y$ imply that $E(X) < E(Y)$. Clearly the converse is false in either case.

We can define strict stochastic order and strict convex order by replacing "less than or equal to" by "less than" in Definitions 2.1 and 2.2. If $X \leq_d Y$ and $Y \leq_d X$ then we write $X =_d Y$. This implies that $F(x) = G(x)$ for all $x \in \mathbb{R}$, where F and G are the respective distribution functions for X and Y. The generalization to many random variables is straightforward; $(X_1,X_2,\ldots,X_n) =_d (Y_1,Y_2,\ldots,Y_n)$ means that the joint distribution functions for $(X_1,X_2,\ldots,X_n)$ and $(Y_1,Y_2,\ldots,Y_n)$ are equal at every point $\underset{\sim}{x}$ in $\mathbb{R}^n$.

In this paper we will be interested in comparing stationary interevent times in certain traffic processes in queueing networks. We prove the following result which we will make use of in Chapter 4.

Theorem 2.1. Let X_1 and X_2 be two stationary interevent times, and Y_1 and Y_2 the corresponding stationary backward recurrence times (time since the last event). If $E(X_1) = E(X_2)$ and $Y_1 <_d Y_2$, then $X_1 <_c X_2$.

Proof: Let F_1 and F_2 be the distribution functions of X_1 and X_2, and G_1 and G_2 the distribution functions of Y_1 and Y_2. Then (using the notation $\bar{G}(x) = 1 - G(x)$)

$$\bar{G}_i(x) = \frac{1}{E(X_i)} \int_x^\infty \bar{F}_i(t)dt \quad i = 1,2, \quad x \in \mathbf{R}^+$$

$$Y_1 <_d Y_2 \Longrightarrow G_1(x) > G_2(x)$$

$$\Longrightarrow \frac{1}{E(X_1)} \int_x^\infty \bar{F}_1(t)dt < \frac{1}{E(X_2)} \int_x^\infty \bar{F}_2(t)dt$$

$$\Longrightarrow X_1 <_c X_2.$$

2.3 Reverse Processes and Reversibility

In this paper we will use some ideas on the reversibility of Markov processes to demonstrate many of our assertions. Therefore, we present some preliminaries on reversibility and reverse processes. Additional information is given in Kelly (1979) and Kiessler (1983).

Our discussion here will be limited to stationary Markov and Markov renewal processes. We will assume that our systems are such that a stationary probability distribution exists for the processes in question.

From a stationary Markov process $\mathcal{N} = \{N(t), t \in \mathbf{R}\}$ on a

countable state space E with generator $Q = \langle\langle q(i,j)\rangle\rangle$ and stationary distribution $\underline{\pi}$, we can construct a Markov process $\mathcal{N}^r = \{N^r(t),\ t \in \mathbb{R}\}$ with generator $Q^r = \langle\langle q^r(i,j)\rangle\rangle$ and stationary distribution $\underline{\pi}$ where

$$q^r(i,j) = \frac{\pi_j}{\pi_i} q(j,i) \quad \text{for all } i,j \in E.$$

Definition 2.3. $\mathcal{N}^r$ is the reverse process of $\mathcal{N}$.

Definition 2.4. A Markov process $\mathcal{N}$ is said to be reversible if for each $n \in \mathbb{I}^+$, $t_1, t_2, \ldots, t_n, \tau \in \mathbb{R}$

$$(N(t_1), N(t_2), \ldots, N(t_n)) =_d (N(\tau - t_1), N(\tau - t_2), \ldots, N(\tau - t_n))$$

or equivalently

$$(N(t_1), N(t_2), \ldots, N(t_n)) =_d (N^r(t_1), N^r(t_2), \ldots, N^r(t_n)).$$

At this point, let us explain our definitions. In most of our work on the feedback queue, we work with random processes defined on $\mathbb{R}^+$ or $\{0, \mathbb{R}^+\}$. Our definition of reversibility, however, is given for processes on $\mathbb{R}$. To avoid confusion we will rely on the following construction. From a stationary Markov process $\mathcal{N} = \{N(t),\ t \in \mathbb{R}^+\}$, we can produce the unique extension $\mathcal{N}' = \{N(t),\ t \in \mathbb{R}\}$, where $\mathcal{N}$ and $\mathcal{N}'$ have the same finite dimensional distributions on $\mathbb{R}^+$. Note that a Markov process that is stationary at time zero can be thought of as having started in any state at time $-\infty$. Then the extension $\mathcal{N}'$ includes the states of the process at any finite negative t, whose finite dimensional distributions we know. The reverse process $\mathcal{N}'^r$ of

$\mathcal{N}'$ is then constructed as above. By the reverse process $\mathcal{N}^r$ of $\mathcal{N}$, we then mean the process $\mathcal{N}'^r$ restricted to R^+. We say that $\mathcal{N}$ is reversible if $\mathcal{N}'$ is reversible. We will use an analogous construction for the Markov renewal processes below.

If $\mathcal{N}$ is a reversible Markov process, Kelly (1979) has shown that

$$\pi_i \, q(i,j) = \pi_j \, q(j,i) \quad \text{for all } i,j \in E \tag{2.1}$$

and hence $q(i,j) = q^r(i,j)$ for all $i,j \in E$. Conversely, if Equations 2.1 are satisfied the stationary Markov process $\mathcal{N}$ is reversible. Equations 2.1 are known as the detailed balance equations and provide a means of demonstrating reversibility.

The global balance equations are obtained by summing the detailed balance equations over all j; i.e. they are

$$\sum_{j \in E} \pi_i \, q(i,j) = \sum_{j \in E} \pi_j \, q(j,i) \quad \text{for all } i \in E$$

or

$$\pi_i \sum_{j \in E} q(i,j) = \sum_{j \in E} \pi_j \, q(j,i) \quad \text{for all } i \in E.$$

A solution $\{\pi_i\}$ to the detailed balance equations is a solution to the global balance equations, but the converse is not true.

Similarly, we can define a reverse process and reversibility for stationary Markov renewal processes. A time-homogeneous Markov renewal process $(\mathcal{N},\mathcal{T}) = \{N_n, T_n, n = 0,1,2,\ldots\}$, where N_n takes values on a countable state space E and T_n takes values in R^+ (with $T_0 < T_1 < \ldots < T_n$), is a two dimensional process with the property that

$$\Pr(N_n=j, T_n-T_{n-1} \leq t \mid N_0, N_1, \ldots, N_{n-1}=i, T_0, T_1, \ldots, T_{n-1})$$

$$= \Pr(N_n=j, T_n-T_{n-1} \leq t \mid N_{n-1}=i) = k(i,j,t) \quad \text{for all } n \in \mathbb{I}^+,\ t \in \mathbb{R}^+.$$

The matrix $K = K(t) = \langle\langle k(i,j,t) \rangle\rangle$ is the semi-Markov kernel of $(\mathcal{N},\mathcal{T})$, and $k(i,j,t)$ $(i,j \in E)$ are the transition functions of the process. Çinlar (1975) showed that if $(\mathcal{N},\mathcal{T})$ is a Markov renewal process, then $\mathcal{N} = \{N_n,\ n \in \mathbb{I}^+\}$ is a Markov chain with state space E and transition matrix P whose elements are $p(i,j) = \lim_{t\to\infty} k(i,j,t)$.

Given a Markov renewal process $(\mathcal{N},\mathcal{T})$ on E with semi-Markov kernel K and stationary distribution ν of $\mathcal{N}$ (we will later give conditions under which the processes we work with are irreducible, aperiodic and recurrent non-null, and hence possess a limiting (stationary) probability distribution), we can construct a Markov renewal process $(\mathcal{N}^r,\mathcal{T}^r)$ on E with transition functions

$$k^r(i,j,t) = \frac{\nu_j}{\nu_i} k(j,i,t) \quad \text{for all } i,j \in E,\ t \in \mathbb{R}.$$

The stationary distribution of $\mathcal{N}^r$ is also ν.

Definition 2.5. $(\mathcal{N}^r,\mathcal{T}^r)$ is the reverse process of $(\mathcal{N},\mathcal{T})$.

Definition 2.6. A Markov renewal process $(\mathcal{N},\mathcal{T})$ is said to be reversible if $K(t) = K^r(t)$ for all $t \in \mathbb{R}$.

Equations 2.1 have an analogue for Markov renewal processes. The Markov renewal process $(\mathcal{N},\mathcal{T})$ will be reversible if and only if the detailed balance equations are satisfied:

$$\nu_i \, k(i,j,t) = \nu_j \, k(j,i,t) \quad \text{for all } t \in \mathbb{R} . \tag{2.2}$$

Equations 2.2 are equations for functions of t; Equations 2.1 for a Markov process are equations for constants.

We will use properties of the reverse process to relate traffic processes corresponding to inputs and outputs at a node in a queueing network. As above, we quote many results without proof. The proofs are available in Kiessler (1983).

2.4 The Queue with Instantaneous Bernoulli Feedback

In Chapters 4 and 5, we will analyze a queueing system in which units arrive for service from "the outside" according to some random arrival process. Each unit is assigned a service requirement according to some distribution function. Depending on the order of arrival to the server, each unit receives service at a non-negative rate. When the unit's service requirement is met, the unit leaves the server and encounters a switch where with some (fixed) probability the unit leaves the system for "the outside" (it departs) and with the complementary probability the unit is independently assigned another service requirement and returns immediately to the server (it feeds back). This system is illustrated in Figure 2.1. We shall label the units entering the service mechanism (server and queue) as inputs and those leaving the server as outputs. Those units entering the server from the outside will be called arrivals and those leaving the system for the outside will be called departures.

Let $\{A_n, n = 0,1,2,\ldots\}$ be the sequence of interarrival times to this queue, $\{S_k, k = 0,1,2,\ldots\}$ the sequence of service requirements,

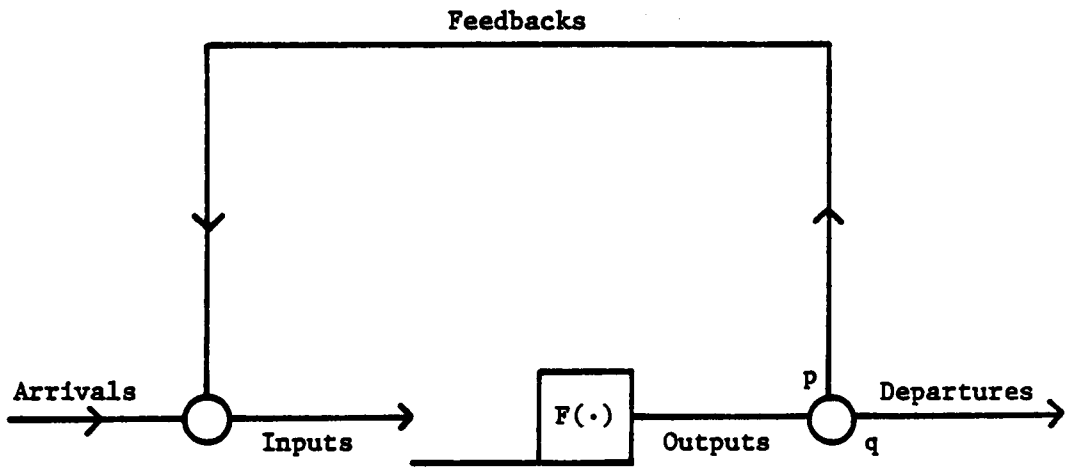

Figure 2.1 Queue with instantaneous Bernoulli feedback

and $\{Y_j,\ j = 0,1,2,\ldots\}$ the sequence of states of a switching mechanism. The n^{th} arrival does not necessarily receive the n^{th} service requirement or the n^{th} state of the feedback switch. The values $\{Y_0,Y_1,\ldots\}$ determine the feedback sequence, with the interpretation that Y_j is zero if the j^{th} output feeds back and one if it departs. For each n, A_n, S_n and Y_n are real valued random variables on a complete probability space $(\Omega,\mathcal{F},P)$. We assume that the sequences $\{A_n\}$, $\{S_n\}$ and $\{Y_n\}$ are mutually independent, i.i.d. random sequences with the following properties:

$$\Pr(A_n < t) = 1 - e^{-\lambda t}, \quad 0 < \lambda < \infty,\ t > 0,$$

$$\Pr(S_k < t) = F(t), \quad t > 0,$$

$$\text{with } 0 < E(S_n) < \infty,\ F(0^+) = 0$$

$$\Pr(Y_j = i) = \begin{cases} p & i = 0 \\ q = 1 - p & i = 1 \\ 0 & \text{otherwise.} \end{cases} \qquad 0 < q < 1$$

From our description above, the arrival times form a Poisson process with rate λ, the service requirements have a general distribution, and the feedback (switching) mechanism is a Bernoulli process.

Now, define the counting processes associated with arrivals (a), inputs (i), feedbacks (f), outputs (o) and departures (d). For $x = a,i,f,o$ or d, define a random variable $C^x(t)$ that counts the number of x-events in $(0,t]$. For example, for each $\omega \in \Omega$, $C^f(t;\omega)$ is the number of feedbacks in the interval $(0,t]$. For each x, $C^x(t)$ is a right continuous step function from $(\Omega,\mathcal{F},P)$ into $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$. Clearly

$$C^i(t) = C^a(t) + C^f(t) \text{ and } C^o(t) = C^d(t) + C^f(t).$$

From these counting processes, we can determine the sequences of event epochs $\mathcal{T}^x = \{T_n^x,\ n = 1,2,\ldots\}$, where T_n^x is the time of the n^{th} x-event (i.e. the n^{th} jump point of $C^x(t)$). Then

$$T_n^x = \begin{cases} \inf\{t > 0:\ C^x(t^-) \neq C^x(t)\} & n = 1 \\ \inf\{t > T_n^x:\ C^x(t^-) \neq C^x(t)\} & n = 2,3,4,\ldots. \end{cases}$$

Thus the arrivals and the feedbacks comprise the inputs, and the departures and feedbacks comprise the outputs; that is,

$$\mathcal{T}^f \subset \mathcal{T}^i,\ \mathcal{T}^a \subset \mathcal{T}^i,\ \mathcal{T}^f \cup \mathcal{T}^a = \mathcal{T}^i$$

and

$$\mathcal{T}^f \subset \mathcal{T}^o,\ \mathcal{T}^d \subset \mathcal{T}^o,\ \mathcal{T}^f \cup \mathcal{T}^d = \mathcal{T}^o.$$

$\mathcal{D}^x = \{D_n^x,\ n = 1,2,\ldots\}$ is the <u>interevent time process</u> for events of type x. That is

$$D_1^x = T_1^x$$

$$D_n^x = T_n^x - T_{n-1}^x \quad n = 2,3,4,\ldots.$$

Since there is a one-to-one correspondence between D_n^x and T_n^x, it makes no difference whether we discuss the process $\mathcal{D}^x$ or $\mathcal{T}^x$. we will examine whichever is most convenient for a particular situation. For x = a,i,f,o or d, $C^x(t)$, T_n^x , and D_n^x are all measurable functions of $(\Omega,\mathcal{F},P)$ into $(R^+,\mathcal{B}(R^+))$.

Throughout this work, we will be concerned with the <u>queue length process</u> and certain <u>traffic processes</u> in our feedback queue. Informally, we think of the queue length process as the number of units

in the queue (in service and waiting for service) at any time t. Since we will deal with systems in which units begin receiving service upon their arrival to the queue, the term queue may seem a little strange, since units never wait for service to begin. This terminology is entrenched in the queueing literature, so we will use it here. At any point in time, the queue length Q(t) is determined by

$$Q(t) = Q(0) + C^a(t) - C^d(t)$$

where Q(0) is the number in the system initially. A traffic process will refer to one of the sequences $\mathcal{T}^x$, but may include other information as well, depending on context. Informally, a traffic process will be made up of the times at which a particular event (e.g. an output) occurs. Our main goal will be to describe the times between events in such a process.

We shall use $\mathcal{N} = \{N(t),\ t \in \mathbb{R}^+\}$ to denote the <u>state process</u> of the system. In general the state of the system N(t) will include the queue length at time t, but in some situations it will include more. We shall always use the term state process to refer to a Markov process describing the evolution of the queue length. Sometimes it will be necessary to supplement the queue length with continuous random variables (e.g. with a vector of remaining service times), and hence the state process may be defined on an uncountable state space. In such cases we will refer to a <u>limiting probability density</u> for the state process instead of a limiting probability distribution. We will, however, provide marginal results for the queue length itself.

We refer to the traffic processes corresponding to inputs, outputs and feedbacks as <u>internal</u> flows, and those corresponding to arrivals

and departures as external flows. In order to analyze the internal flows in the network, we shall often make use of a joint process made up of the state of the system at a particular point in time (e.g. the time of the n^{th} output) and the corresponding event epoch. This pair will be denoted by (N_n^x, T_n^x) for x = i,f,o where

$N_n^x = N^x(T_n)$ is the state of the system at the n^{th} x-event

and

T_n^x is the time of the n^{th} x-event.

The process $\{(N_n^x, T_n^x),\ n = 1,2,3,\ldots\}$ will be denoted by $(\mathcal{N}^x, \mathcal{T}^x)$.

2.5 Symmetric Queueing Disciplines; The Processor Sharing Queue

Kelly (1979) has used the term "symmetric" queue to identify a particular class of service disciplines. We will use Kelly's definition with the exception that we consider only one class of customer.

Definition 2.7. A service discipline is called symmetric if it operates in the following manner:

(i) a total service effort is provided at rate $\phi(n)$ when n units are in the queue;

(ii) a proportion $\gamma(\ell,n)$ of this service effort is dedicated to the unit in position ℓ ($\ell = 1,2,\ldots,n$); when this unit leaves the queue, units in positions $\ell+1, \ell+2, \ldots, n$ move to positions $\ell, \ell+1, \ldots, n-1$;

(iii) when a unit arrives at the queue it moves into position ℓ ($\ell = 1,2,\ldots,n+1$) with probability $\gamma(\ell,n+1)$; units in

positions $\ell,\ell+1,\ldots,n$ previously move to positions $\ell+1,\ell+2,\ldots,n+1$ respectively.

We analyze a special type of symmetric discipline known as processor sharing. For this discipline $\gamma(\ell,n) = 1/n$, $\ell = 1,2,\ldots,n$, $n = 1,2,\ldots$; all units in service are receiving service at the same rate (i.e. they share the processor equally). If we let $\phi(n) = 1$, the queue behaves as a single server processor sharing queue. Then each unit receives service at rate $\gamma(\ell,n) \cdot \phi(n) = 1/n$. If we let $\phi(n) = n$, the queue behaves as an infinite server queue and each unit receives service at rate $\gamma(\ell,n) \cdot \phi(n) = 1$. In this case each unit acts as if it has its own server.

Other symmetric disciplines include stacks (last in, first out with preemption) and overflow queues (queues with no waiting room). While we have studied only processor sharing queues here, we discuss possible extensions to other symmetric disciplines in Chapter 6.

The class of symmetric queues possesses certain properties that make them quite appealing from a practical modelling standpoint. These are related to a property called insensitivity, and we give a definition of insensitivity here in the form in which we will use it. Let $\{\Sigma_i(M_\alpha), i \in \Theta\}$ (where Θ is some index set) designate a collection of queueing systems with the property that the governing sequences are probabilistically indistinguishable (i.e. have the same finite dimensional distributions) except perhaps for governing sequence M, which has fixed parameter α. Let $A(\Sigma)$ be a property of the queueing system Σ.

Definition 2.8. If for any two systems $\Sigma_1, \Sigma_2 \in \{\Sigma_i(M_\alpha), i \in \Theta\}$ in equilibrium, $A(\Sigma_1) = A(\Sigma_2)$, A is said to be insensitive to governing sequence M given parameter α.

As an example, suppose $A(\cdot)$ is the stationary queue length distribution of $(\cdot)$ and M the sequence of service requirements with mean α. Then for all processor sharing queues with Poisson arrivals (rate λ), A is insensitive to M given α. Insensitivity is an important property in that it allows us to use results from simpler systems to analyze more complicated systems. We will explore the notion of insensitivity of interval distributions in Chapters 4, 5 and 6.

2.6 Direction of This Research

We will use the ideas in this chapter to analyze a feedback queue where the service discipline is processor sharing. Our main goal is to study the effect of the service time distribution on the queueing properties of this system. We are interested in insensitivity properties of various imbedded and arbitrary time stationary queue length distributions. We would like to describe the properties of internal traffic processes in this queue and, where possible, to make comparisons between interinput and interoutput distributions in comparable queueing systems. In Chapter 4 we analyze the single server processor sharing queue with exponential service requirements, in Chapter 5 we analyze the same system with general service requirements and in Chapter 6 we analyze the infinite server queue. For the special case of processor sharing considered in Chapter 6, we are able to obtain explicit formulas for the interoutput distribution for certain

service time distributions.

CHAPTER 3

LITERATURE REVIEW

3.1 Introduction

This chapter reviews the literature on queueing networks as it pertains to our study of feedback queues. The study of queueing networks, of which the feedback queue is a special case, begins with the papers of Jackson (1957,1963). Most of the current queueing network literature deals with networks studied by Jackson or generalizations of them. Disney and König (1985) give a more general overview of these topics than we will give here. We will discuss only the literature that provides a background for the systems we analyze in this paper. Unless otherwise noted, the results we cite are equilibrium results.

3.2 Traffic Processes in Queueing Networks

While feedback queues are analyzed as systems in their own right, we begin our literature review with some results that have appeared in the queueing network literature on traffic processes in more general networks. This literature has focused primarily on determining which flows in queueing networks are Poisson processes.

Beutler and Melamed (1978) and Melamed (1979) gave necessary and sufficient conditions for the flows in a Jackson network of ·/M/1 nodes to be a Poisson process. The conditions are the so-called "loop criteria" and can be stated as follows: a flow along an arc from node i to node j is a Poisson process if and only if once a unit leaves node j it can never return to node i; further, the output process from a node is a Poisson process if and only if once a unit leaves the node it

can never return to it. These criteria have been extended to networks of quasi-reversible queues with bounded service rates (Walrand (1982)). Clearly, feedback queues violate the loop criteria.

The study of the relationship between input and output processes at a single node in a network also gives insight into the feedback queue. A result of Walrand and Varaiya (1980) states that in the M/M/1-IBF queue, the input and the output processes are different as point processes (i.e. they do not have the same joint interval distribution). Disney, McNickle and Simon (1980) showed that when these processes were considered as Markov renewal processes, the input and output processes are different processes, although single intervals have the same distribution. Kiessler (1983) and Disney and Kiessler (1987) gave general results on the relationship between input and output processes at a node using the idea of reversing a Markov renewal process. For networks with a reversible state process (the M/M/1-IBF queue, for instance) they showed that the Markov renewal input process is the reverse of the Markov renewal output process. This result gives a way to compute the kernel of one process from the kernel of the reverse process, and thus a way to determine the finite dimensional distributions of either process. This has proved to be a useful tool in our analysis.

3.3 Feedback Queues

Queues with instantaneous feedback have long been used to model random phenomena, most importantly in computer systems; Kleinrock (1976), Chapter 4 and Wyszewianski and Disney (1974) discuss applications of these models. All results of which we are aware are

for stationary, first-come, first-served systems.

The earliest analytical results on the M/GI/1 queue with instantaneous Bernoulli feedback (M/GI/1-IBF) were developed by Takács (1963). He found the stationary distribution of the queue size and the first two moments of the sojourn time (time spent in the system) of a unit. Montazer-Haghighi (1977) extended this analysis to multi-server feedback queues.

D'Avignon (1974) and D'Avignon and Disney (1976,1977) considered the M/GI/1 queue with a feedback rule that could depend on the increment in the queue length between two successive service completions, the length of the service received, and whether or not the previous unit fed back. For this system, the authors found the stationary queue length distribution and characterized the output and departure processes as marked point processes (specifically, as Markov renewal processes). They analyzed the busy period as well.

Disney, König and Schmidt (1985) studied the stationary imbedded and arbitrary time queue length distributions for the M/GI/1-IBF queue. They also studied the stationary waiting time for one pass of a unit through the system and the time stationary virtual waiting time.

The internal flows in feedback queues are much more complicated than might be expected, as several papers have indicated. Burke (1976) studied the internal traffic processes in the M/M/1-IBF queue and showed that, in this simple single-server system, the input process to the server is not a Poisson process. This paper provided an important counterexample to the conjecture that all flows in a Jackson network are Poisson processes, indeed making Jackson's results all the more

surprising and demonstrating the need to analyze the internal flows in these networks.

Disney, McNickle and Simon (1980) gave the first detailed analysis of the traffic processes in the M/GI/1-IBF queue. They showed that while the output process is a Markov renewal process, it is never a renewal process unless the server is exponential and the feedback probability is zero (i.e. M/GI/1-IBF reduces to M/M/1). Similarly, the input process is a Markov renewal process and is equivalent (in the sense of Simon and Disney (1984)) to a renewal process if and only if GI = M and the feedback probability is zero, in which case it is a Poisson process.

The papers of Hunter (1983,1984,1985) further exploit the Markov renewal structure of the traffic processes to derive many results on the equality of various imbedded stationary distributions. He studied the relationships between flows in Markovian single server feedback queues. His results on properties of the flow processes particularize the results of Kiessler (1983) and point out that these flows are very complicated indeed. Chandramohan and Disney (1982) computed correlations between flows in the feedback queue and showed that intervals in traffic processes are not only autocorrelated but cross-correlated as well.

3.4 Symmetric Service Disciplines: The Processor Sharing and Infinite Server Queues

The concept of a symmetric queueing discipline was introduced by Kelly (1976) as a generalization of certain disciplines considered by, among others, König, Matthes and Nawrotzki (1967). For the processor

sharing queue (a member of this class), Baskett and Palacios (1972) first established the insensitivity of the stationary queue length distribution to the distribution of service requirements when the mean service requirement is held fixed. This property was then extended to general symmetric disciplines by a number of authors (e.g., Brumelle (1978), Schassberger (1977,1978a,1978b,1978c), Jansen and König (1980), Burman (1980)).

The processor sharing queue was introduced as a limiting model for the round-robin discipline (cf. Kleinrock (1976) and references therein) and appears to have found application in computer science. Muntz (1972) first gave a proof of the important result that the departure process from a stationary processor sharing queue with Poisson input is Poisson irrespective of the service time process. Kelly (1979), using the tools of reversibiltiy, extended this result to symmetric disciplines with Poisson input. In a series of papers, Yashkov (1980,1981a,1981b) and Kitayev and Yashkov (1979) analyzed the processor sharing queue and gave an alternative proof of the Poisson departure property. Cohen (1979) also considered a network of processor sharing nodes and, also using reversibility arguments, showed that the departure process is a Poisson process.

The infinite server queue was discussed by Takács (1960), who obtained the stationary distribution of the queue length process (Exercise 8, pp. 39 and 86). For the M/GI/∞ queue, Mirasol (1963) first showed that the stationary queue length distribution depends on the service time distribution only through its mean and that the departure process is Poisson with the same parameter as the arrival

process. Foley (1982) extended this result in two directions. He showed that for a nonstationary Poisson arrival process, the departure process is a (possibly nonstationary) Poisson process. He also showed that the departure process from a tandem sequence of $\cdot/GI/\infty$ queues with an external Poisson arrival process is a Poisson process.

We are unaware of any literature dealing specifically with processor sharing or infinite server queues with feedback, or any analysis of internal flows in networks of symmetric queues. Indeed, the following chapters were strongly motivated by Kelly's (1979) construction of a network of quasi-reversibility queues. Kelly defines such a network as a collection of nodes that in isolation would be quasi-reversible. He gives a definition of quasi-reverisbility on pp. 65-66; since symmetric queues are quasi-reversible, we could substitute symmetric for quasi-reversible in this discussion. On p. 69, Kelly states, "Note that the j^{th} queue of the network will not in general satisfy the conditions required for it to be quasi-reversible...". This statement begs a question. Exactly which properties of a quasi-reversible node in isolation are inherited when that node is put in a network? In Chapters 4, 5 and 6, we attempt to ascertain to what extent the insensitivity properties of a node in isolation carry over to a node in a network.

3.5 Comparison Methods for Queues

Most of the applications literature in queueing theory has emphasized measures of system effectiveness as a way to compare different models. These measures commonly include the stationary average queue length, average waiting time and coefficient of variation

of the queue length. There are major problems with limiting a comparison to mean values, and even comparing higher moments of distributions may hide different tail probabilities. In the last 15 years, the ideas of partial order relationships have been used to compare entire probability distributions in queueing systems. These results are sometimes more difficult to obtain but are stronger than moment results. Often one avoids the problem of actually computing moments as well, for techniques have been developed whereby stochastic inequalities are verified as pointwise inequalities on a suitably chosen probability space (cf. O'Brien (1975) and Sonderman (1980)). For the order relationships used in this paper, Stoyan (1983) provides a complete reference. We also mention Whitt (1980) which studied order properties of certain congestion measures (such as the stationary waiting time process) in the GI/G/s queue.

3.6 Summary

Feedback queues with the first come, first served discipline have been well studied. The internal traffic processes in these queues have been shown to be quite complicated. In symmetric queues, the queue length process and the departure process has been well studied. These queues possess insensitivity properties that simplify the analyses considerably.

As we mentioned in Section 3.4, Kelly's comments about networks of quasi-reversible queues need clarification. To this end, the next three chapters will study two symmetric queueing disciplines at a queue with instantaneous feedback. Our purpose is to study in detail properties of the internal traffic processes and the corresponding

imbedded queue length distributions so as to see what effect queue discipline has on these processes. Specifically, we expose the insensitivity properties of internal flows in these networks and thereby clarify Kelly's comments.

CHAPTER 4

THE SINGLE SERVER PROCESSOR SHARING QUEUE WITH FEEDBACK: EXPONENTIAL SERVICE REQUIREMENTS

4.1 Introduction

In this chapter we discuss some properties of the symmetric queueing discipline known as processor sharing. As in Chapter 2, arrivals to the queue occur according to a Poisson process with rate λ. Service requirements are assigned independently according to a distribution function $F(\cdot)$. All units in the system share the single processor equally, so that when n units are in service, the service requirement of each decreases at rate 1/n. In the notation of Section 2.5, $\phi(n) = 1$ for $n = 1,2,\ldots$ ($\phi(0) = 0$) and $\gamma(\ell,n) = 1/n$ for $\ell = 1,2,\ldots,n$ and $n = 1,2,\ldots$. Upon completion of the service requirement, a unit departs with probability q and feeds back with probability p, with $p + q = 1$. In this chapter we will describe the stationary queue length distributions and traffic processes in this network for exponential service times.

Let

$$F(t) = 1 - e^{-\mu t}, \quad t > 0.$$

That is, the service requirements are exponentially distributed random variables with parameter $\mu > 0$. Service requirements are assigned independently. We refer to Section 2.4 for a complete description of the network.

4.2 Queue Length Processes

For this system, the queue length process $\mathcal{N} = \{N(t), t \in \mathbb{R}^+\}$ is a Markov process with generator Γ, where

$$\gamma(i,i+1) = \lambda \qquad i = 0,1,2,\ldots$$

$$\gamma(i,i-1) = q\mu \qquad i = 1,2,3,\ldots \quad .$$

The transition rate diagram is given in Figure 4.1. When n units are in the queue, the instantaneous departure rate of each is $q\mu/n$, so the rate at which any unit departs is $q\mu/n \cdot n = q\mu$. Notice that the generator is the same as for an M/M/1 queue with service requirement parameter $q\mu$.

Theorem 4.1. $\mathcal{N}$ has a limiting distribution iff $\lambda/q\mu < 1$.

Proof: Let us transform the M/M/1/PS-IBF queue into a queue without feedback that stochastically has the same queue length. This "equivalent" queue works in the following manner. Units arrive according to a Poisson process with rate λ and are independently assigned a total service requirement S' from the distribution U(·) where

$$U(s) = \begin{cases} F_1(s) & \text{with probability } q \\ F_2(s) & \text{with probability } qp \\ \vdots & \\ F_n(s) & \text{with probability } qp^{n-1} \\ \vdots & \end{cases}$$

and for each j, $F_j(s)$ is the distribution function of the sum of j independent exponential (μ) random variables. The new queue now operates as an M/GI/1/PS queue (without feedback); that is, it provides uninterupted service to each unit. It operates as if in the original queue each unit is assigned upon arrival a total service time which

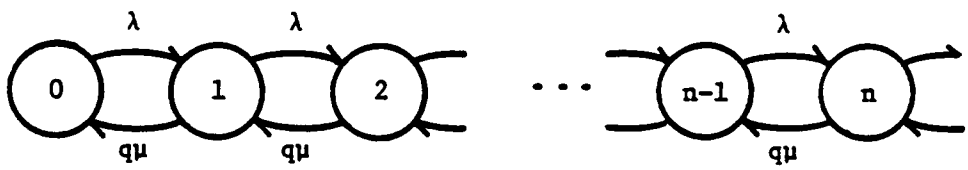

Figure 4.1 Transition rate diagram for the queue length process

takes into account the number of times it will feed back and its service requirement on each pass.

Both queues are irreducible for $\lambda,\mu > 0$. Clearly, the new queue will be empty if and only if the original queue is empty. Hence the condition for stationarity (namely that the state $\{N(t) = 0\}$ occurs for infinitely many t) will be the same for both queues. For the M/GI/1/PS queue the condition is

$$\lambda\, E(S') < 1.$$

Now $E(S') = (q\mu)^{-1}$. So $\lambda\, E(S') < 1 \iff \lambda/q\mu < 1$.

The queue length process in the feedback queue will be stationary if and only if a limiting queue length distribution exists and is the initial distribution. Let $\pi_k^o = \Pr(N(0) = k)$ and $\pi_k = \lim_{t\to\infty} \Pr(N(t) = k)$.

Theorem 4.2. The M/M/1/PS-IBF queue length process is stationary iff $\lambda/q\mu < 1$ and $\pi_k^o = \pi_k = (1 - \lambda/q\mu)(\lambda/q\mu)^k$ for all $k \in E$.

Proof: The global balance equations for $\{\pi_k\}$ are

$$\lambda\pi_0 = q\mu\pi_1$$

$$(\lambda + \mu)\pi_k = \lambda\pi_{n-1} + q\mu\pi_{n+1} \qquad n = 1,2,\ldots.$$

Then $\pi_k = (\lambda/q\mu)^k\pi_0$, and $\pi_0 = [1 - \sum_{k=1}^{\infty} \pi_k]^{-1} \Longrightarrow \pi_k = (1 - \lambda/q\mu)(\lambda/q\mu)^k$.

In the remainder of this chapter we will assume that $\mathcal{N}$ is stationary. To describe the various imbedded queue length distributions, we use the notation π_k^x to mean $\lim_{n\to\infty} \Pr(N_n^x = k)$ for

$x = a,i,f,o$ or d. The random variables N_n have been defined in Section 2.4.

Lemma 4.3. The stationary probabilities π_k^a and π_k are equal.

Proof: See Wolff (1981). This result is a special case of the result that Poisson arrivals to Markov processes see time averages.

Now in order to study the queue length at departure epochs, we need the following characterization of the queue length process.

Lemma 4.4. $\mathcal{N} = \{N(t),\ t \in \mathbb{R}^+\}$ is a reversible process.

Proof: We need to demonstrate that the detailed balance equations (2.1) are solved by the $\{\pi_k\}$ in Theorem 4.2. These equations are

$$\pi_i\ \gamma(i,j) = \pi_j\ \gamma(j,i) \quad \text{for all } i,j \in E$$

or in our case

$$\pi_i\ \gamma(i,i+1) = \pi_{i+1}\ \gamma(i+1,i) \quad \text{for all } i = 0,1,2,\ldots.$$

Specifically,

$$\pi_i \cdot \lambda = \pi_{i+1} \cdot q\mu$$

$$\frac{\pi_{i+1}}{\pi_i} = \frac{\lambda}{q\mu}.$$

Clearly, the equations are satisfied by $\pi_k = (1 - \lambda/q\mu)(\lambda/q\mu)^k$.

We are now in a position to characterize $\{\pi_k^d\}$, the stationary queue length distribution imbedded just after a departure.

Theorem 4.5. The stationary probabilities π_k^d and π_k are equal.

Proof: Consider the sample paths of the process $\{N(t), t \in \mathbb{R}^+\}$. A realization of the process is given in Figure 4.2. Let time run "forward" when we observe the process from left to right and "backward" when we observe the process from right to left. In forward time, upward jumps correspond to arrivals and downward jumps to departures. In reverse time, upward jumps corrrespond to departures and downward jumps to arrivals. Since $\{N(t), t \in \mathbb{R}^+\}$ is reversible, and upward jumps in the forward process form a Poisson process with rate λ, so do upward jumps in the reverse process. Moreover, if t_o is a fixed point, the sequence of downward jumps prior to t_o and the queue length at t_o are independent. Hence the reverse process $\mathcal{N}^r$ is also the queue length proces of an M/M/1/PS-IBF queue with arrival parameter λ, and therefore $\pi_k^d = \pi_k$.

To summarize, we have shown that the stationary queue length distributions at an arbitrary time, just before an arrival and just after a departure coincide. In Section 4.4 we will look at these distributions imbedded at input and output times, and in Section 5.4 we will study the insensitivity of these distributions to the distribution of service requirements with fixed mean.

4.3. The Departure Process

The only external flow in this system that we have not described is the departure process $\mathcal{T}^d = \{T_n^d, n = 1,2,\ldots\}$. From our discussion of the proof of Theorem 4.5, which follows along the lines of Theorem 2.1 of Kelly (1979), we will state the following result without a separate proof. In Section 5.3 we will give a proof due to Kitayev and

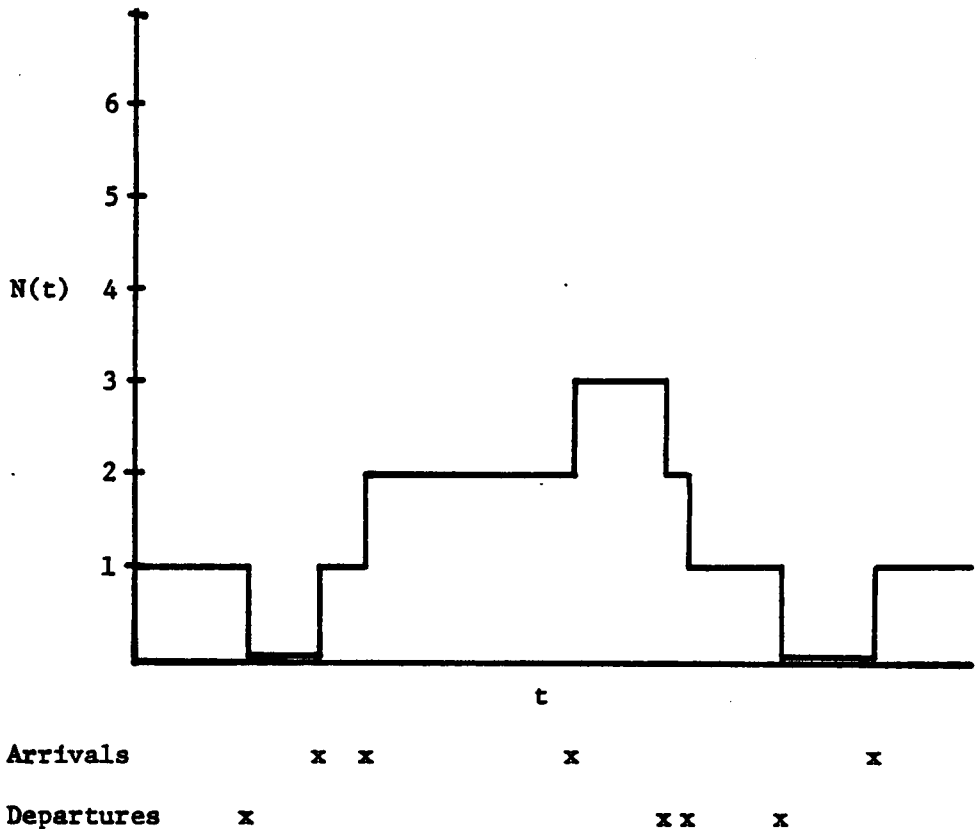

Figure 4.2 A sample path of the queue length process

Yashkov (1979) for general service distributions that we find more illuminating.

Theorem 4.6. (Kelly (1979)). $\mathcal{T}^d$ is a Poisson process with rate λ.

4.4. Input and Output Processes and Imbedded Queue Lengths

We now describe some of the internal processes in this feedback queue. It is easiest to analyze the input process first, and from this analysis we will be able to give results about the output process using reversibility arguments. In all imbedded processes, the unit making the jump is not included in the state descriptor (e.g. the queue length).

Let T_n^i be the time of the n^{th} input and let $N_n^i = N(T_n^i-)$ be the queue length just before the n^{th} input. We call $(\mathcal{N}^i,\mathcal{T}^i) = \{N_n^i, T_n^i,\ n = 1,2,3,\ldots\}$ the marked input process.

Theorem 4.7. The marked input process $(\mathcal{N}^i,\mathcal{T}^i)$ is a Markov renewal process on $E = \{0,1,2,\ldots\}$. Its kernel is given by the functions

$$Q^i(j,k,t) = \begin{cases} \int_0^t pq^j L_0(y)dS^{(j+1)}(y) + \int_0^t q^{j+1} \sum_{n=j+1}^{\infty} M_n(y)dA(y) & \begin{array}{l} j = 0,1,2,\ldots \\ k = 0 \end{array} \\ \int_0^t pq^{j-k} L_0(y)dS^{(j-k+1)}(y) + \int_0^t q^{j-k+1} M_{j-k+1}(y)dA(y) & \begin{array}{l} j = 1,2,3,\ldots \\ k = 1,2,\ldots,j \end{array} \\ \int_0^t M_0(y)dA(y) & \begin{array}{l} j = 0,1,2,\ldots; \\ k = i+1 \end{array} \\ 0 & \text{otherwise} \end{cases}$$

where

$$L_n(y) = \frac{(\lambda y)^n e^{-\lambda y}}{n!} \qquad M_n(y) = \frac{(\mu y)^n e^{-\mu y}}{n!}$$

$$dS^{(n)}(y) = \frac{\mu(\mu y)^{n-1} e^{-\mu y}}{(n-1)!}\, dy.$$

Proof: Case i) $j = 0,1,2,\ldots;\ k = 0$. In order for the queue to be empty before the next input, when the current input finds i units in the system, one of two things must occur. Either there are no arrivals during the service of all j+1 units, all but the last unit depart, and the last unit feeds back (the first term), or "at least" all j+1 units complete service and depart before the next external arrival (the second term).

Case ii) $j = 1,2,3,\ldots;\ k = 1,2,\ldots,j$. Again, one of two cases must occur. Either there are no arrivals during service of j-k+1 units, the first j-k of which depart and the last of which feeds back, or exactly j-k+1 units complete service and depart during an interarrival interval.

Case iii) $j = 0,1,2,\ldots;\ k = j+1$. This can only happen if no customers complete service during an interarrival interval.

The transition probabilities for the Markov chain $\mathcal{N}^i = \{N_n^i,\ n = 1,2,\ldots\}$ are obtained by evaluating the kernel as $t \to \infty$. We obtain the following matrix:

$$P^i = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ j \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} \frac{\mu}{\lambda+\mu} & \frac{\lambda}{\lambda+\mu} & 0 & 0 \\ \frac{q\mu^2}{(\lambda+\mu)^2} & \frac{\mu(\lambda+p\mu)}{(\lambda+\mu)^2} & \frac{\lambda}{\lambda+\mu} & 0 \\ \frac{q^2\mu^3}{(\lambda+\mu)^3} & \frac{q\mu^2(\lambda+p\mu)}{(\lambda+\mu)^3} & \frac{\mu(\lambda+p\mu)}{(\lambda+p\mu)^2} & \frac{\lambda}{\lambda+\mu} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{q^j\mu^{j+1}}{(\lambda+\mu)^{j+1}} & \frac{q^{j-1}\mu^j(\lambda+p\mu)}{(\lambda+\mu)^{j+1}} & \frac{q^{j-2}\mu^{j-1}(\lambda+p\mu)}{(\lambda+\mu)^j} & \frac{q^{j-3}\mu^{j-2}(\lambda+p\mu)}{(\lambda+\mu)^{j-1}} \quad \cdots \\ \vdots \end{bmatrix} \end{array}$$

Theorem 4.8. The stationary distribution of the queue length imbedded just before input times is given by

$$\pi_k^i = (1 - \lambda/q\mu)(\lambda/q\mu)^k \qquad k = 0,1,2,\ldots \quad .$$

provided $\lambda/q\mu < 1$.

Proof: The global balance equations for the Markov chain $\{N_n^i,\ n = 1,2,\ldots\}$ are given by

$$\sum_{j=0}^{\infty} \pi_j^i P^i(j,k) = \pi_k^i .$$

So for π_0^i, for example, we have

$$\pi_0^i = \pi_0^i(\frac{\mu}{\lambda+\mu}) + \pi_1^i(\frac{q\mu^2}{(\lambda+\mu)^2}) + \pi_2^i(\frac{q^2\mu^3}{(\lambda+\mu)^3}) + \ldots \quad .$$

Clearly, $\pi_k^i = (1 - \frac{\lambda}{q\mu})(\frac{\lambda}{q\mu})^k$ solves this equation provided $\lambda/q\mu < 1$,

since

$$\pi_0^i = \frac{\lambda+\mu}{\lambda}\left(1 - \frac{\lambda}{q\mu}\right)\left[\left(\frac{\lambda}{q\mu}\right)\left(\frac{q\mu^2}{(\lambda+\mu)^2}\right) + \left(\frac{\lambda}{q\mu}\right)^2\left(\frac{q^2\mu^3}{(\lambda+\mu)^3}\right) + \ldots\right]$$

$$= \left(1 - \frac{\lambda}{q\mu}\right)\left(\frac{\mu}{\lambda+\mu}\right)\left[1 + \frac{\lambda}{\lambda+\mu} + \frac{\lambda^2}{(\lambda+\mu)^2} + \ldots\right]$$

$$= \left(1 - \frac{\lambda}{q\mu}\right)\left(\frac{\mu}{\lambda+\mu}\right)\left(\frac{\lambda+\mu}{\mu}\right) = \left(1 - \frac{\lambda}{q\mu}\right),$$

and we can similarly verify that each π_k solves $\sum_{j=0}^{\infty} \pi_j^i P^i(j,k) = \pi_k^i$.

We are now in a position to characterize the output process from this system. Recall that the output process $\{N_n^o, T_n^o,\ n = 1,2,3,\ldots\}$ contains the queue length imbedded just after the n^{th} output. From Kiessler (1983) we know that the output process is a Markov renewal process and is the reverse of the input process provided that the queue length process is reversible. But by Lemma 4.4, $\mathcal{N}$ is reversible. Since we have the kernel of the input process and the stationary distribution of its imbedded Markov chain, we can easily determine the kernel of the output process.

Theorem 4.9. The marked output process $(\mathcal{N}^o, \mathcal{T}^o)$ is a Markov renewal process on $E = \{0,1,2,\ldots\}$. Its kernel is given by the functions

$$Q^{o}(j,k,t) = \begin{cases} \int_0^t p\, L_k(y)dS(y) + \int_0^t q\, L_{k+1}(y) \sum_{n=k+1}^{\infty} M_n(y)dA(y) & j = 0,\ k = 0,1,2,\ldots \\ \int_0^t p\, L_{k-1}(y)dS(y) + \int_0^t q\, L_{k-j+1}(y)dS(y) & j = 1,2,3,\ldots,\ k = j+1, j+2,\ldots \\ \int_0^t q\, L_0(y)dS(y) & j = 1,2,3,\ldots,\ k = j-1 \\ 0 & \text{otherwise} \end{cases}$$

Proof: The kernel is obtained from $Q^{i}(j,k,t)$ using the relationship

$$Q^{o}(j,k,t) = \frac{\pi_k^i}{\pi_j^i} Q^{i}(k,j,t).$$

So, for example, for $j = 1,2,3,\ldots$, $k = j+1, j+2,\ldots$

$$Q^{o}(j,k,t) = \frac{(1 - \frac{\lambda}{q\mu})(\frac{\lambda}{q\mu})^k}{(1 - \frac{\lambda}{q\mu})(\frac{\lambda}{q\mu})^j}\left[\int_0^t pq^{k-j} L_0(y)dS^{(k-j+1)}(y) + \int_0^t q^{k-j+1} M_{k-j+1}(y)dA(y)\right]$$

$$= \frac{\lambda^{k-j}}{(q\mu)^{k-j}}\left[\int_0^t pq^{k-j} e^{-\lambda y} \frac{\mu^{k-j+1} y^{k-j} e^{-\mu y}}{(k-j)!}\, dy + \int_0^t q^{k-j+1} \frac{(\mu y)^{k-j+1} e^{-\mu y}}{(k-j+1)!} \lambda e^{-\lambda y} dy\right]$$

$$= \int_0^t p \frac{(\lambda y)^{k-j} e^{-\lambda y}}{(k-j)!} \mu e^{-\mu y} dy + \int_0^t q \frac{(\lambda y)^{k-j+1} e^{-\lambda y}}{(k-j+1)!} \mu e^{-\mu y} dy$$

$$= \int_0^t p\, L_{k-1}(y)dS(y) + \int_0^t q\, L_{k-j+1}(y)dS(y).$$

The other terms follow by similar algebraic manipulation.

Corollary 4.10. The stationary probabilities π_k^i and π_k^o are equal.

Proof: From Kiessler (1983), since $(\mathcal{N}^o, \mathcal{T}^o)$ is the reverse of $(\mathcal{N}^i, \mathcal{T}^i)$, both $\mathcal{N}^o$ and $\mathcal{N}^i$ have the same stationary distribution.

From the generators Q^i and Q^o we can determine the stationary distribution of an interinput interval $D_n^i = T_n^i - T_{n-1}^i$ and an interoutput interval $D_n^o = T_n^o - T_{n-1}^o$. Since we are assuming stationarity we will drop the subscript n, since $D_n^i =_d D_{n+k}^i$ and $D_n^o =_d D_{n+k}^o$ for any $k \in \Pi^+$.

Theorem 4.11. If D^i is a stationary interinput time and D^o a stationary interoutput time, then

$$\Pr(D^i < t) = \Pr(D^o < t) = 1 - \frac{q\mu-\lambda}{\mu-\lambda} e^{-\lambda t} - \frac{p\mu}{\mu-\lambda} e^{-\mu t}, \quad t > 0,$$

$$E(D^i) = E(D^o) = \frac{q}{\lambda},$$

$$\mathrm{Var}(D^i) = \mathrm{Var}(D^o) = \frac{q\mu(1+p) - 2\lambda p}{\lambda^2 \mu}.$$

Proof: The distribution of D^i is given by

$$\Pr(D^i < t) = \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} Q^i(j,k,t)\pi_j^i$$

and the distribution of D^o is given by

$$\Pr(D^o < t) = \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} Q^o(j,k,t)\pi_j^o.$$

By computing these sums using Theorem 4.8 and Corollary 4.10 we obtain the result stated in the Theorem. The mean and variance follow from standard principles.

Theorem 4.11 states that single intervals in the input and output processes have the same distribution. This result is typical of these feedback queues (cf. Disney, McNickle and Simon (1980), for a similar result in the M/M/1/FCFS-IBF queue). However, the joint distributions of multiple intervals in the input and output processes are <u>not</u> the same; we can demonstrate this by noting that (in matrix notation)

$$\Pr(D_n^i \leq t_1, D_{n+1}^i \leq t_2) = \underset{\sim}{\pi}\, Q^i(t_1)Q^i(t_2)U$$

and

$$\Pr(D_n^o \leq t_1, D_{n+1}^o \leq t_2) = \underset{\sim}{\pi}\, Q^o(t_1)Q^o(t_2)U$$

where U is a row vector of ones. By performing this matrix multiplication we find that $\Pr(D_n^i \leq t_1, D_{n+1}^i \leq t_2) \neq \Pr(D_n^o \leq t_1, D_{n+1}^o \leq t_2)$. This tells us that as processes $\mathscr{D}^i$ and $\mathscr{D}^o$ are not equivalent (in the sense of Simon and Disney (1984)); they do not have the same finite dimensional distributions, even though single intervals in each process have the same distribution.

For $t \in R^+$, define V(t) to be the forward recurrence time (time until the next event) in the output process. That is,

$$V(t) = T_{n+1}^o - t \quad \text{for} \quad T_n^o \leq t \leq T_{n+1}^o.$$

Since the system is stationary at time zero, for simplicity let us use $V \equiv V(0)$ to be the time until the first output after time zero. From Theorem 4.11 we can determine the distribution of V.

<u>Theorem 4.12</u>. The stationary distribution of V is given by

$$\Pr(V \leq t) = 1 - \frac{q\mu-\lambda}{q(\mu-\lambda)} e^{-\lambda t} - \frac{p\mu}{q(\mu-\lambda)} e^{-\mu t}, \quad t > 0.$$

Proof: $\Pr(V > t) = \frac{1}{E(D^o)} \int_t^\infty \Pr(D^o > y)dy$

$$= \frac{\lambda}{q} \int_0^\infty \left(\frac{q\mu-\lambda}{\mu-\lambda} e^{-\lambda y} + \frac{p\mu}{\mu-\lambda} e^{-\mu y}\right)dy$$

$$= \frac{\lambda}{q(\mu-\lambda)} \left(\frac{q\mu-\lambda}{\lambda} e^{-\lambda y} \Big|_t^\infty + pe^{-\mu y} \Big|_t^\infty \right)$$

$$= \frac{q\mu-\lambda}{q(\mu-\lambda)} e^{-\lambda t} + \frac{p\lambda}{q(\mu-\lambda)} e^{-\mu t}, \quad t > 0.$$

$$\Pr(V < t) = 1 - \frac{q\mu-\lambda}{q(\mu-\lambda)} e^{-\lambda t} - \frac{p\lambda}{q(\mu-\lambda)} e^{-\mu t}, \quad t > 0.$$

Corollary 4.13. The mean and variance of V are given by

$$E(V) = \frac{q\mu^2 - \lambda\mu + p\lambda^2}{q\lambda\mu(\mu-\lambda)},$$

$$\mathrm{Var}(V) = \frac{2(q\mu-\lambda)}{q\lambda^2(\mu-\lambda)} + \frac{2p\lambda}{q\mu^2(\mu-\lambda)} - \left(\frac{q\mu^2 - \lambda\mu + p\lambda^2}{q\lambda\mu(\mu-\lambda)}\right)^2.$$

We will use Theorem 4.12 and Corollary 4.13 to compare the distribution of output intervals in queues with non-exponential service requirements that have mean $1/\mu$. Table 4.1 compares the means and variances of D^o and V for some settings of λ and q with μ held fixed at 1. When the stationary queue length distribution exists (i.e. $\lambda/q\mu < 1$) and q is small, both the mean and the variance of V (the forward recurrence time) are much larger than the respective parameter of D^o (the interoutput interval). For any fixed λ, as q increases to 1, both the means and variances of V and D^o get closer together. This convergence reflects the fact that for $q = 1$, both V and D^o have the same

Table 4.1

Means and variances of D^o and V, $\mu = 1$

λ \ q		.2	.3	.5	.7	.9
.1	$E(D^o)$	2.00	3.00	5.00	7.00	9.00
	E(V)	6.00	7.67	9.00	9.57	9.89
	$Var(D^o)$	20.00	37.00	65.00	85.00	97.00
	Var(V)	76.00	89.89	97.00	98.96	99.76
.2	$E(D^o)$		1.50	2.50	3.50	4.50
	E(V)		2.67	4.00	4.57	4.89
	$Var(D^o)$		5.75	13.75	19.75	23.75
	Var(V)		14.89	22.00	23.96	24.77
.3	$E(D^o)$			1.67	2.33	3.00
	E(V)			2.33	2.90	3.22
	$Var(D^o)$			5.00	8.11	10.33
	Var(V)			8.11	10.07	10.88
.4	$E(D^o)$			1.25	1.75	2.25
	E(V)			1.50	2.07	2.39
	$Var(D^o)$			2.19	4.19	5.69
	Var(V)			3.25	5.21	6.02
.5	$E(D^o)$				1.40	1.80
	E(V)				1.57	1.89
	$Var(D^o)$				2.44	3.56
	Var(V)				2.96	3.77
.6	$E(D^o)$				1.17	1.50
	E(V)				1.24	1.56
	$Var(D^o)$				1.53	2.42
	Var(V)				1.74	2.54
.7	$E(D^o)$					1.29
	E(V)					1.32
	$Var(D^o)$					1.74
	Var(V)					1.81
.8	$E(D^o)$					1.13
	E(V)					1.14
	$Var(D^o)$					1.30
	Var(V)					1.33

an output is almost surely a departure, and since the departures form a Poisson process with rate λ (Theorem 4.6) the interoutput interval is exponentially distributed, and therefore so is the forward recurrence time.

4.5 Summary

In this chapter we have obtained results for the queue length and traffic processes in the M/M/1 processor sharing queue with feedback. We gave conditions under which a stationary queue length distribution exists and computed the distribution. Using the idea of reversibility, we showed that the stationary queue lengths imbedded at arrival points, departure points and at an arbitrary time are equal. We showed that the input and output processes are Markov renewal processes and determined their semi-Markov kernels. We were then able to compute the stationary distribution of a single interinput and interoutput interval and the forward recurrence times in the input and output processes. The analysis in this chapter relies heavily on the Markovian structure of the queue length, which is the case only when service requirements are independent exponential random variables.

CHAPTER 5

THE SINGLE SERVER PROCESSOR SHARING QUEUE WITH FEEDBACK: GENERAL SERVICE REQUIREMENTS

5.1 Introduction

The Markovian structure of the queue length process when the service requirements are exponentially distributed as in Section 4.2 allowed us to obtain a number of queue length and traffic process results. Now we keep the same structure as in Section 4.2 but relax the assumption of exponential service requirements.

Assume that service requirements are assigned independently according to distribution function $F(\cdot)$ with density $f(\cdot)$. The queue length process in this system is no longer Markov. We can, however, define a vector valued process $(\mathcal{N}, \underset{\sim}{\mathcal{Y}}) = \{N(t), \underset{\sim}{Y}(t), t \in R^+\}$ where

$$\underset{\sim}{Y}(t) = (y_1(t), y_2(t), \ldots, y_{N(t)}(t))$$

and

$N(t)$ = number of units in the system at time t

$y_i(t)$ = accumulated service time of i^{th} unit in the system.

We shall assume that $y_1(t) < y_2(t) < \ldots < y_n(t)$, so elements of the vector $\underset{\sim}{Y}(t)$ are ordered. Then the <u>state process</u> $(\mathcal{N}, \underset{\sim}{\mathcal{Y}})$ is a Markov process on $E = \Pi^+ \times (\emptyset \cup R^1 \cup R^2 \cup \ldots)$. We will use $\underset{\sim}{y} = (y_1, y_2, \ldots, y_n)$ to denote a generic vector of accumulated service times. Thus in state $(n, \underset{\sim}{y})$, n units are in the system with accumulated service times $y_1 < y_2 < \ldots < y_n$. The state where the server is idle is designated by $\emptyset$. We denote by $e_i(\underset{\sim}{y})$ the vector obtained when, from the vector $\underset{\sim}{y}$, the i^{th} job instantaneously completes service; i.e. $e_i(\underset{\sim}{y}) = (y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$. We will use $\pi_k(\underset{\sim}{y})$ and $\pi(\emptyset)$ to

denote the limiting probability density of $(n,\underset{\sim}{y})$ and the limiting probability of an empty system, respectively, and $\{\pi_k\}$ to denote the marginal limiting queue length distribution.

5.2 Queue Length Processes

We use the approach of Ross (1983) to study the queue length process at an arbitrary point in time. As in Section 4.2, we let $U(\cdot)$ be the distribution of the entire service time experienced by a unit before it leaves the system and $u(\cdot)$ its density. The instantaneous failure rate (departure rate) of a unit that has completed z units of service is

$$\mu(z) = \frac{u(z)}{\bar{U}(z)}, \quad z > 0.$$

Now define the revese process $(\mathcal{N}^r, \underset{\sim}{\mathcal{Y}}^r)$ of $(\mathcal{N}, \underset{\sim}{\mathcal{Y}})$ to be a two-dimensional vector valued process $\{N^r(t), \underset{\sim}{Y}^r(t),\ t \in \mathbb{R}^+\}$ where

$$\underset{\sim}{Y}^r(t) = (y_1^r(t), y_2^r(t), \ldots, y_{N(t)}^r(t))$$

and

$N(t)$ = number of customers in the system at time t

$y_i^r(t)$ = remaining service time of i^{th} customer in the system.

Kelly (1979) calls $(\mathcal{N}^r, \underset{\sim}{\mathcal{Y}}^r)$ the dynamic reverse of $(\mathcal{N}, \underset{\sim}{\mathcal{Y}})$. We can solve for $\{\pi(\emptyset), \pi_k(\underset{\sim}{y})\}$ by solving the detailed balance equations for this system. The possible transitions and their rates in the forward and reverse process are as follows:

Arrivals (F) $(n,\underset{\sim}{y}) \longrightarrow (n+1,(0,\underset{\sim}{y}))$ at rate λ

(R) $(n+1,(0,\underset{\sim}{y})) \longrightarrow (n,\underset{\sim}{y})$ at rate $1/n+1$

Departures (F) $(n,\underset{\sim}{y}) \longrightarrow (n-1,e_i(\underset{\sim}{y}))$ at rate $\mu(y_i)/n$

(R) $(n-1),e_i(\underset{\sim}{y})) \longrightarrow (n,\underset{\sim}{y})$ at rate $\lambda u(y_i)$.

Let us explain these rates. In the forward process, arrivals occur at rate λ. With the system in state $(n,\underset{\sim}{y})$ the i^{th} unit will depart at rate $\mu(y_i)/n$, since it receives service at rate $1/n$ and instantaneously completes its service requirement at rate $\mu(y_i)$ when it has had y_i service units already. Similarly, in the reverse process the state changes from $(n+1,(0,\underset{\sim}{y}))$ to $(n,\underset{\sim}{y})$ when the unit whose service requirement is depleted departs; this happens at rate $\mu(0)/n+1 = 1/n+1$. The transition $(n-1,e_i(\underset{\sim}{y})) \longrightarrow (n,\underset{\sim}{y})$ occurs when a unit with remaining service requirement y_i arrives; this happens at rate $\lambda u(y_i)$.

Theorem 5.1. The limiting density for the M/GI/1/PS-IBF state process is given by

$$\begin{aligned} \pi(\emptyset) &= (1 - \lambda/q\mu) \\ \pi_k(\underset{\sim}{y}) &= \lambda^k(1 - \lambda/q\mu)k! \prod_{i=1}^{k} \bar{U}(y_i). \end{aligned} \tag{5.1}$$

Proof: It suffices to demonstrate that $\{\pi(\emptyset),\pi_k(\underset{\sim}{y})\}$ above solves the detailed balance equations. From the rates of the forward and reverse processes, the equations are

$$\begin{aligned} \pi_k(\underset{\sim}{y})\lambda &= \pi_{k+1}(0,\underset{\sim}{y})\,\frac{1}{k+1} \\ \pi_k(\underset{\sim}{y})\,\frac{\mu(y_i)}{k} &= \pi_{k-1}(e_i(\underset{\sim}{y}))\lambda u(y_i) \\ \pi(\emptyset)\lambda &= \pi_1(y)\mu(y_i) \end{aligned} \qquad k = 1,2,\ldots$$

and the normalizing condition is

$$1 = \pi(\emptyset) + \sum_{k=1}^{\infty} \int_0^{\infty} \cdots \int_0^{\infty} \pi_k(\underline{y})d\underline{y}.$$

Hence

$$\pi_k(\underline{y}) = \pi_{k-1}(e_i(\underline{y}))\frac{\lambda u(y_i)k}{\mu(y_i)} = \pi_{k-1}(e_i(\underline{y}))\lambda\bar{U}(y_i)k$$

$$= \pi_{k-2}(e_j(e_i(\underline{y})))\lambda^2\bar{U}(y_j)\bar{U}(y_i)k(k-1) \qquad (i \neq j)$$

$$\vdots$$

$$= \pi(\emptyset)\lambda^k\, k! \prod_{i=1}^{k} \bar{U}(y_i)$$

$$\Longrightarrow 1 = \pi(\emptyset) + \sum_{k=1}^{\infty} \pi(\emptyset)\lambda^k k! \int_{y_1<y_2<\cdots<y_k} \cdots \int \prod_{i=1}^{k} \bar{U}(y_i)dy_1dy_2\cdots dy_k.$$

Since each of the k orderings $y_1 < y_2 <\cdots< y_k$ are equally likely with probability $\frac{1}{k!}$, we have

$$\pi(\emptyset) = [1 + (\sum_{k=1}^{\infty} \lambda^k \int_{y_1} \int_{y_2} \cdots \int_{y_k} \prod_{i=1}^{k} \bar{U}(y_i)dy_1dy_2\cdots dy_k)]^{-1}$$

$$= [\sum_{k=0}^{\infty} \lambda^k(E(S'))^k]^{-1} = 1 - \lambda/q\mu$$

$$\pi_k(\underline{y}) = \lambda^k(1 - \lambda/q\mu)k! \prod_{i=1}^{k} \bar{U}(y_i).$$

Corollary 5.2. The limiting distribution for the queue length process is given by

$$\pi_k = (1 - \lambda/q\mu)(\lambda/q\mu)^k \qquad k = 0,1,2,\ldots$$

and, in equilibrium, given that there are $k > 0$ units in the system, their accumulated service requirements are independent and identically

distributed with distribution function

$$\Pr(y_i \leq y) = q\mu \int_0^y \bar{U}(t)dt \qquad y > 0,\ i = 1,2,\ldots,k.$$

Proof: $\pi_k = \int_{y_1<y_2<\cdots<y_k} \pi_k(\underline{y})d\underline{y}$

$$= \frac{1}{k!} \int_0^\infty \int_0^\infty \cdots \int_0^\infty \lambda^k(1 - \lambda/q\mu)k! \prod_{i=1}^{k} \bar{U}(y_i)d\underline{y}$$

$$= (1 - \lambda/q\mu)(\lambda/q\mu)^k.$$

From (5.1) it follows that given $k > 0$, $(y_1,y_2,\ldots,y_k)$ are i.i.d., and

$$\Pr(y_i \leq y|N(0) = k) = \frac{\Pr(y_i \leq y,\ N(0) = k)}{\Pr(N(0) = k)}$$

$$= \frac{\int_{y_1<y_2<\cdots<y_{i-1}<0<y_{i+1}<\cdots<y_k} \int_0^y \pi_k(\underline{y})dy_1dy_2\cdots dy_k}{\pi_k}$$

$$= \frac{\int_0^y \int_0^\infty \cdots \int_0^\infty (1 - \frac{\lambda}{q\mu})\lambda^k \frac{k!}{k!} \bar{U}(t) \prod_{\substack{j=1 \\ j\neq i}}^{k} \bar{U}(y_j)dy_jdt}{(1 - \frac{\lambda}{q\mu})(\frac{\lambda}{q\mu})^k}$$

$$= \frac{(1 - \frac{\lambda}{q\mu})\lambda^k(\frac{1}{q\mu})^{k-1} \int_0^y \bar{U}(t)dt}{(1 - \frac{\lambda}{q\mu})(\frac{\lambda}{q\mu})^k}$$

$$= q\mu \int_0^y \bar{U}(t)dt.$$

Independence follows because

$$\Pr(N(0) = k, y_1 \leq x_1,\ldots,y_k \leq x_k) = \pi_k \prod_{i=1}^{k} \Pr(y_i \leq x_i).$$

Notice that $\{\pi_k\}$ in Corollary 5.2 exhibits the insensitivity of the stationary queue length distribution to changes in the service requirement distribution since it depends only on the mean service

requirement. We can do a little more at this point. In Theorem 5.1 we demonstrated the reversibility of the state process $(\mathcal{N}, \underset{\sim}{\mathcal{Y}})$. Now arrivals in the forward process form a Poisson process with rate λ independent of the state of the system and correspond to times at which the N(t) part of the state vector increases by one. Therefore, times at which the reverse process changes state by increasing N(t) by one also form a Poisson process with rate λ independent of the system state. But these are the departure times of the forward process. Hence arguing analogously to Theorems 4.3 and 4.5 we give the following result.

Theorem 5.3. The stationary probabilities π_k^a, π_k^d and π_k are equal.

Thus both the stationary queue length seen by arrivals and the stationary queue length left behind by departures are insensitive to the service requirement distribution given its mean.

5.3. The Departure Process

The following proof is due to Kitayev and Yashkov (1979). We present it here because it is an illuminating proof of the Poisson departure property. We will use their notation in this section only.

Theorem 5.4. In equilibrium $\mathcal{T}^d$ is a Poisson process.

Proof: (Kitayev and Yashkov (1979)). Consider a random process $(\mathcal{N}, \underset{\sim}{\mathcal{X}}, \underset{\sim}{\mathcal{Y}}) = \{N(t); x_j(t), 1 \leq j \leq m(t); y_i(t), 1 \leq i \leq N(t)\}$, $m(t) > 0$, $t > 0$ with the following interpretation. N(t) is the queue length at time t, $y_i(t)$ is the accumulated service time on unit i at

time t, and the vector $\underset{\sim}{X}(t) = (x_1(t), x_2(t), \ldots, x_{m(t)}(t))$ is the vector of backward recurrence times in the departure process where m(t) is the number of departures up to time t. That is, $x_1(t)$ is the time since the last departure, $x_2(t)$ is the time since the next-to-last departure, and so on. Yashkov works with an unordered vector $\underset{\sim}{Y}(t)$.

Now $(\mathcal{N}, \underset{\sim}{\mathcal{X}}, \underset{\sim}{\mathcal{Y}})$ is a Markov process. We denote by $\pi_{k,m}(\underset{\sim}{x}, \underset{\sim}{y})$, $k = 0,1,2,\ldots$, $m = 0,1,2,\ldots$, $x_i > 0$, $i = 0,1,2,\ldots,m$, $0 < y_1 < \ldots < y_k$ the stationary probability density of the state process $(\mathcal{N}, \underset{\sim}{\mathcal{X}}, \underset{\sim}{\mathcal{Y}})$. The stationary density can be obtained by considering the evolution of the Markov process over a very small time interval Δ. For simplicity, let $q_{k,m}(\underset{\sim}{x}, \underset{\sim}{y}) = \dfrac{\pi_{k,m}(\underset{\sim}{x}, \underset{\sim}{y})}{\prod_{i=1}^{k} (\bar{U}(y_i))}$. Letting $\Delta \to 0$, the equations that the $q_{k,m}(\underset{\sim}{x}, \underset{\sim}{y})$ must satisfy are given by:

$$\left[\frac{1}{k} \sum_{i=1}^{k} \frac{\partial}{\partial y_i} + \sum_{j=1}^{m} \frac{\partial}{\partial x_j} + \lambda\right] q_{k,m}(x_1, \ldots, x_m; y_1, \ldots, y_k) = 0 \qquad (k > 1),$$

$$\left[\sum_{j=1}^{m} \frac{\partial}{\partial x_j} + \lambda\right] q_{0,m}(x_1, \ldots, x_m) = 0 \qquad (k = 0),$$

$$q_{k+1,m}(x_1, \ldots, x_m; y_1, \ldots, y_k, 0) = \lambda q_{km}(x_1, \ldots, x_m; y_1, \ldots, y_k), \qquad (5.2)$$

$$q_{k,m}(0, x_1, \ldots, x_{m-1}; y_1, \ldots, y_k)$$

$$= \int_{y=0}^{\infty} \int_{x_m = x_{m-1}}^{\infty} q_{k+1,m}(x_1, \ldots, x_m; y_1, \ldots, y_k, y)\, dx_m\, dU(y).$$

The first two equations specify the behavior of the system between jump points (i.e. the continuous evolution of the process); the third equation specifies the behavior at an arrival time, and the fourth equation specifies the behavior at a departure time.

The boundary conditions for m = 0 are

$$[\sum_{j=1}^{m} \frac{\partial}{\partial x_j} + \lambda]q_k(y_1,\ldots,y_k) = \int_0^\infty q_{k+1}(y_1,\ldots,y_k,y)dU(y)$$

$$q_{k+1}(y_1,\ldots,y_k,0) = \lambda q_k(y_1,\ldots,y_k). \tag{5.3}$$

Now the departure process will be a Poisson process if the density

$$\pi_{k,m}(x_1,\ldots,x_m,y_1,\ldots,y_k) = \lambda^{k+m}(1 - \frac{\lambda}{q\mu}) \prod_{j=1}^{m} e^{-\lambda(x_j - x_{j-1})} e^{-\lambda x_0} \prod_{i=1}^{k} \bar{U}(y_i)$$

$$= \lambda^{k+m}(1 - \frac{\lambda}{q\mu})e^{-\lambda x_m} \prod_{i=1}^{k} \bar{U}(y_i) \tag{5.4}$$

satisfies equations (5.2) and (5.3). The form of the conjectured density implies that the joint distribution of $(x_1(t),\ldots,x_m(t))$ is independent of the state of the system in equilibrium, since the density is of the form

$$\pi_{k,m}(x_1,\ldots,x_m;y_1,\ldots,y_k) = \pi_k(y_1,\ldots,y_k)\pi_m(x_1,\ldots,x_m)$$

where $\pi_k(y_1,\ldots,y_k) = \lambda^k(1 - \frac{\lambda}{q\mu}) \prod_{i=1}^{k} \bar{U}(y_i)$ and $\pi_m(x_1,\ldots,x_m) = \lambda^m e^{-\lambda x_m}$.

By direct substitution we can verify that (5.4) solves (5.2) and (5.3); eg.

$$[\frac{1}{k} \sum_{i=1}^{k} \frac{\partial}{\partial y_i} + \sum_{j=1}^{m} \frac{\partial}{\partial x_j} + \lambda]q_{k,m}(x_1,\ldots,x_m;y_1,\ldots,y_k)$$

$$= [\frac{1}{k} \sum_{i=1}^{k} \frac{\partial}{\partial y_i} + \sum_{j=1}^{m} \frac{\partial}{\partial x_j} + \lambda]\lambda^{k+m}(1 - \frac{\lambda}{q\mu})e^{-\lambda x_m}$$

$$= \lambda^{k+m}(1 - \frac{\lambda}{q\mu})(-\lambda e^{-\lambda x_m}) + \lambda^{k+m+1}(1 - \frac{\lambda}{q\mu})e^{-\lambda x_m} = 0,$$

$$q_{k+1,m}(x_1,\ldots,x_m;y_1,\ldots,y_k,0)$$

$$= \lambda^{k+m+1}(1 - \frac{\lambda}{q\mu})e^{-\lambda x_m} = \lambda q_{k,m}(x_1,\ldots,x_m;y_1,\ldots,y_k)$$

$$q_{k,m}(0,x_1,\ldots,x_{m-1};y_1,\ldots,y_k) = \lambda^{k+m}(1 - \frac{\lambda}{q\mu})e^0 = \lambda^{k+m}(1 - \frac{\lambda}{q\mu})$$

$$= \int_{y=0}^{\infty} \int_{x_m=x_{m-1}}^{\infty} \lambda^{k+m+1}(1 - \frac{\lambda}{q\mu})e^{-\lambda x_m} dx_m dU(y)$$

$$= \int_0^{\infty} \int_{x_{m-1}}^{\infty} q_{k+1,m}(x_1,\ldots,x_m;y_1,\ldots,y_k,y)dx_m dU(y).$$

Hence the departure process from the M/GI/∞-IBF queue is a Poisson process in equilibrium.

5.4 Input and Output Processes and Imbedded Queue Lengths

The output process consists of all times at which units complete a service requirement. The departure process is then obtained by deleting those points in the output process that are assigned a value of 0 by the feedback switch. Since the departure process is always a Poisson process and since the queue length process is insensitive to the form of the service time distribution, given its mean, one might suspect that the output process possesses some insensitivity properties as well. For example, the equilibrium distribution of an arbitrary interoutput interval may be the same for any two processor sharing queues that have the same mean service requirement. In this section we give a counterexample to this conjecture and point out certain order relationships between interoutput intervals from comparable processor sharing queues.

To study the output process, we will modify our definition of the state. Define $y_i(t)$ to be the accumulated service time of unit i at time t on its current pass through the server. Let $\underset{\sim}{Y}(t)$ be the ordered vector of these accumulated service times at time t. Then $(\mathcal{N}, \mathcal{Y}) =$

$\{N(t), \underset{\sim}{Y}(t),\ t \in R^{+}\}$, where $N(t)$ is the queue length at time t, is a Markov process whose reverse process is a vector of queue length and remaining service times on each unit's current pass through the system. The instantaneous failure rate of a unit that has accumulated y units of service is now $\nu(y) = \frac{f(y)}{\bar{F}(y)}$. The rates of the forward and reverse process are given at the jump points (arrivals, feedbacks and departures) as follows (at non-jump points, the vector are either increasing or decreasing linearly at the same rate):

Arrivals	(F)	$(n,\underset{\sim}{y}) \longrightarrow (n+1,(0,\underset{\sim}{y}))$	at rate λ
	(R)	$(n+1,(0,\underset{\sim}{y})) \longrightarrow (n,\underset{\sim}{y})$	at rate $q/n+1$
Departures	(F)	$(n,\underset{\sim}{y}) \longrightarrow (n-1,e_i(\underset{\sim}{y}))$	at rate $q\nu(y_i)/n$
	(R)	$(n-1,e_i(\underset{\sim}{y})) \longrightarrow (n,\underset{\sim}{y})$	at rate $\lambda f(y_i)$
Feedbacks	(F)	$(n,\underset{\sim}{y}) \longrightarrow (n,(0,e_i(\underset{\sim}{y})))$	at rate $p\nu(y_i)/n$
	(R)	$(n,(0,e_i(\underset{\sim}{y}))) \longrightarrow (n,\underset{\sim}{y})$	at rate $pf(y_i)/n$.

The rates for arrivals and departures are similar to those developed in Section 5.2. In the forward process, a feedback requires a service completion and a value of 0 for the feedback switch; hence the rate is $p(\nu(y_i)/n)$. In the reverse process the impending service completion must feedback with remaining service requirement y_i; the rate is therefore $p(f(y_i))1/n$.

Now we can obtain the stationary density of $(\mathcal{N}, \underset{\sim}{\mathcal{Y}})$.

Theorem 5.5. The stationary density of the vector $(\mathcal{N}, \underset{\sim}{\mathcal{Y}})$ exists provided $\lambda/q\mu < 1$ and is given by

$$\pi(\emptyset) = (1 - \lambda/q\mu)$$
$$\pi_k(\underline{y}) = (\frac{\lambda}{q})^k (1 - \frac{\lambda}{q\mu}) k! \prod_{i=1}^{k} \bar{F}(y_i). \tag{5.5}$$

Proof: The detailed balance equations for this process are

$$\pi_k(\underline{y})\lambda = \pi_{k+1}(0,\underline{y})q/k\ 1$$

$$\pi_k(\underline{y})q\mu(y_i)/k = \pi_{k-1}(e_i(\underline{y}))\lambda f(y_i) \quad k = 1,2,\ldots$$

$$\pi_k(\underline{y})p\mu(y_i)/k = \pi_{k+1}(0,e_i(\underline{y}))pf(y_i)/k$$

$$\pi(\emptyset)\lambda = \pi_1(y)q\nu(y)$$

and the normalizing condition is

$$1 = \pi(\emptyset) + \sum_{k=1}^{\infty} \int_0^\infty \int_0^\infty \cdots \int_0^\infty \pi_k(\underline{y})d\underline{y}.$$

The conjectured density satisfies the detailed balance equations; e.g. for feedbacks

$$\pi_k(\underline{y})p\nu(y_i)/k = (\frac{\lambda}{q})^k (1 - \frac{\lambda}{q\mu}) k! \prod_{n=1}^{k} \bar{F}(y_n) \frac{p}{k} \frac{f(y_i)}{\bar{F}(y_i)}$$

$$= (\frac{\lambda}{q})^k (1 - \frac{\lambda}{q\mu})(k-1)!pf(y_i) \prod_{\substack{n=1 \\ n\neq i}}^{k} \bar{F}(y_n)$$

$$\pi_k(0,e_i(\underline{y}))pf(y_i)/k = (\frac{\lambda}{q})^k (1 - \frac{\lambda}{q\mu}) k! \prod_{\substack{n=1 \\ n\neq i}}^{k} \bar{F}(y_n) \cdot \bar{F}(0) \frac{pf(y_i)}{k}$$

$$= (\frac{\lambda}{q})^k (1 - \frac{\lambda}{q\mu})(k-1)!pf(y_i) \prod_{\substack{n=1 \\ n\neq i}}^{k} \bar{F}(y_n)$$

since $\bar{F}(0) = 1$.

By similar manipulations, we can show that the other equations are also solved by $\{\pi(\emptyset),\pi_k(\underline{y})\}$ and this is a probability density provided $\lambda/q\mu < 1$.

From Theorem 5.5, the marginal queue length distribution is given by

$$\pi_0 = \pi(\emptyset) = (1 - \lambda/q\mu)$$

$$\pi_k = \int_{y_1<\ldots<y_k} \pi_k(\underline{y})d\underline{y} = (1 - \lambda/q\mu)(\lambda/q\mu)^k \quad k = 1,2,\ldots$$

Notice that π_k agrees with π_k given in Corollary 5.2. Given that there are $k > 0$ units in the system, their accumulated service times are independent and identically distributed with distribution

$$\Pr(y_1 < y|N(0)=k) = \frac{\Pr(y_i < y, N(0)=k)}{\Pr(N(0)=k)}$$

$$= \frac{\int_0^y [\int_0^\infty \cdots \int_0^\infty (\frac{\lambda}{q})^k (1 - \frac{\lambda}{q\mu}) \prod_{\substack{j=1 \\ j\neq i}}^{k} \bar{F}(y_j)dy_j]\bar{F}(t)dt}{(\frac{\lambda}{q\mu})^k(1 - \frac{\lambda}{q\mu})}$$

$$= \frac{(\frac{\lambda}{q})^k(1 - \frac{\lambda}{q\mu})(\frac{1}{\mu})^{k-1} \int_0^y \bar{F}(t)dt}{(\frac{\lambda}{q\mu})^k(1 - \frac{\lambda}{q\mu})}$$

$$= \mu \int_0^y \bar{F}(t)dt.$$

Independence follows since

$$\pi_k \prod_{i=1}^{k} \Pr(y_i < x_i) = (1 - \frac{\lambda}{q\mu})(\frac{\lambda}{q\mu})^k \prod_{i=1}^{k} \mu \int_0^{x_i} \bar{F}(y_i)dy_i$$

$$= (1 - \frac{\lambda}{q\mu})(\frac{\lambda}{q})^k \int_0^{x_1} \cdots \int_0^{x_k} \prod_{i=1}^{k} \bar{F}(y_i)dy_i$$

$$= \Pr(N(0)=k, y_1 < x_1, \ldots, y_k < x_k).$$

Note that since we looked at accumulated work on each unit's current pass through the system, the actual service time distribution $F(\cdot)$ appears in the stationary density, not the distribution of the

total service time the unit would experience before it left the system ($U(\cdot)$) as in Theorem 5.1. Defining the state as we have here allows us to look at the times at which an output occurs. With the previous definition of a state, the system did not change state at an output time unless the output was a departure. With the feedback structure subsumed in the distribution function $U(\cdot)$, we could only make statements about the departure process, which were sufficient in that section.

The stationary density (5.5) has an interesting structure. Recall that (5.1) was the stationary density of an M/GI/1/PS queue (without feedback) with arrival parameter λ and service requirement distribution $U(\cdot)$. Now (5.5) could be interpreted as the stationary density of an M/GI/1/PS queue (without feedback) with arrival parameter λ/q and service requirement distribution $F(\cdot)$. For the queue with feedback, the service distribution is indeed $F(\cdot)$, but the arrival parameter is λ, not λ/q. To simulate (i.e. produce a sample path of) an interoutput time, then, we could let the system run from $(-\infty,0)$ with arrival rate λ/q, and at 0 "slow down" the arrival process to rate λ. Then the time until the first service completion after time zero will be the forward recurrence time of the stationary interoutput time process. We will use this idea to compare output processes in processor sharing queues with different service requirement distributions but with fixed mean.

First, let us look at the mean time between outputs. Let Σ_F be a processor sharing queue with service requirement distribution $F(\cdot)$ whose mean is $1/\mu$. Let D^o be a stationary interoutput time. From Section 4.4 for the exponential case ($F(t) = 1 - e^{-\mu t}$), D^o had mean q/λ

(Theorem 4.11). The Lemma below proves that the mean interoutput time is q/λ for any F with mean $1/\mu$.

Lemma 5.6. $E(D^o) = q/\lambda$ for Σ_F.

Proof: We compute the stationary output rate in Σ_F. The result will follow by inverting the output rate.

Suppose the system is stationary at time zero and define the function

$$I(\varepsilon) = \begin{cases} 0 & \text{if there is no output in } (0,\varepsilon] \\ 1 & \text{if there is at least one output in } (0,\varepsilon]. \end{cases}$$

Then $E(I(\varepsilon)) = \Pr(\text{at least one output in } (0,\varepsilon])$. The output rate is

$$r = \lim_{\varepsilon \to 0} \frac{E(I(\varepsilon))}{\varepsilon} .$$

Now

$$\lim_{\varepsilon \to 0} \frac{E(I(\varepsilon))}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left[\sum_{n=1}^{\infty} \Pr(N^a(0,\varepsilon]=0)\Pr(N(0)=n)\Pr(\psi(0,\varepsilon]=1 \,|\, N(0)=n) \right]$$

where $\psi(\cdot)$ is the counting measure for the number of service completions in $(0,\cdot]$. Hence

$$r = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left[\sum_{n=1}^{\infty} (1 - \lambda\varepsilon + o(\varepsilon))(1 - \frac{\lambda}{q\mu})(\frac{\lambda}{q\mu})^n \Pr(\psi(0,\varepsilon]=1 \,|\, N(0)=n) \right].$$

But since at time zero, the remaining service times of the n units in service are independent, identically distributed random variables (Corollary 5.2), each decreasing at rate 1/n, we can write

$$\Pr(\psi(0,\varepsilon]=1 \,|\, N(0)=n) = \sum_{i=1}^{n} \Pr(\psi_i(0,\varepsilon]=1) \prod_{\substack{j=1 \\ j \neq i}}^{n} \Pr(\psi_j(0,\varepsilon]=0),$$

where $\psi_i(\cdot)$ counts the number of service completions of unit i in $(0,\cdot]$. Now from Theorem 1.2.12 of Franken, König, Arndt and Schmidt (1981),

$$\sum_{i=1}^{n} \Pr(\psi_i(0,\varepsilon]=1) \prod_{\substack{j=1 \\ j\neq i}}^{n} \Pr(\psi_j(0,\varepsilon]=0)$$

$$= n((\mu(\tfrac{\varepsilon}{n}) + o(\tfrac{\varepsilon}{n}))(1 - \mu(\tfrac{\varepsilon}{n}) + o(\tfrac{\varepsilon}{n}))^{n-1}$$

$$= \mu\varepsilon + o(\varepsilon).$$

Hence

$$\lim_{\varepsilon\to 0} \frac{E(I(\varepsilon))}{\varepsilon} = \lim_{\varepsilon\to 0} \frac{1}{\varepsilon} [\sum_{n=1}^{\infty} (1 - \lambda\varepsilon + o(\varepsilon))(1 - \tfrac{\lambda}{q\mu})(\tfrac{\lambda}{q\mu})^n(\mu\varepsilon + o(\varepsilon))]$$

$$= \mu \lim_{\varepsilon\to 0} (1 - \lambda\varepsilon + o(\varepsilon)) \sum_{n=1}^{\infty} (1 - \tfrac{\lambda}{q\mu})(\tfrac{\lambda}{q\mu})^n$$

$$= \mu(\tfrac{\lambda}{q\mu}) = \frac{\lambda}{q}.$$

5.5 A Comparison of Interoutput Times in Special Systems

Lemma 5.6 says that the mean interoutput time is insensitive to the service requirement distribution, given the mean service requirement. The distribution of an interoutput time, however, is not insensitive, as we will now show by counterexample.

As before, with the system stationary at time zero, let V be the time until the first output and $G(t) = \Pr(V \leq t)$ in the feedback queue. Consider a processor sharing queue without feedback, at which units arrive according to a Poisson process with rate λ with general service time requirements, mean $1/\mu$, but which has initial density (5.5). Clearly, this queue is not stationary, since the (unique) stationary density is obtained by solving the detailed balance equations (as we did in Theorem 5.1) and is given by

$$\pi(\emptyset) = (1 - \frac{\lambda}{\mu})$$

$$\pi_k(\underline{y}) = \lambda^k(1 - \frac{\lambda}{\mu})k! \prod_{i=1}^{k} \bar{F}(y_i).$$

We will take the origin to be time zero. Let V^* be the time until the first departure after time zero from this queue (since this queue has no feedback the output and departure processes are the same process). Let $G^*(t) = Pr(V^* < t)$. Then we have two queues, one a stationary queue with feedback and the other a non-stationary queue without feedback. Both have the same initial distribution and both have a Poisson arrival process with rate λ. Then the time until the first service completion has the same distribution in both queues; i.e.

$$G(x) = G^*(x) \quad \text{for all } x.$$

The first service completion in the feedback queue is the first output; the first service completion in the queue without feedback is the first departure. Thus we can analyze the time until the <u>first</u> output from the stationary queue with feedback by analyzing the time until the <u>first</u> departure from a nonstationary queue wihout feedback. The analysis does not extend to the entire output process, since after the first event the queues are no longer equivalent. But clearly, if the first interval after the zero in the output process of the M/GI/1/PS-IBF queue is not insensitive to the service requirement distribution, neither is the entire output process.

To compare interoutput times in systems with different distributions $F(\cdot)$, we first consider the special case where one of the systems has constant service requirements and the other exponential service requirements. We will later discuss generalizations and make a conjecture about comparing any two service requirement distributions.

Let Σ_1 and Σ_2 denote two single server processor sharing queues with Poisson arrivals, rate λ. Σ_i has service requirement distribution function $F_i(\cdot)$ $(i = 1,2)$ where

$$F_1(x) = \begin{cases} 0, & x < 1 \\ 1, & x \geq 1 \end{cases} \qquad F_2(x) = 1 - e^{-x}, \quad x \geq 0.$$

Assume that both systems are stationary at time zero, and let the state of Σ_i be given by $\{N^{\Sigma_i}(t), \underset{\sim}{Y}^{\Sigma_i}(t),\ t \in \quad^{+}\}$, where $N^{\Sigma_i}(t)$ is the queue length in Σ_i at time t and $\underset{\sim}{Y}^{\Sigma_i}(t)$ the vector of remaining service requirements in Σ_i at time t $(i = 1,2)$. Since both systems have the same arrival process, let $\{T_1^a, T_2^a, \ldots\}$ be the epochs of the arrival process; we will compare Σ_1 and Σ_2 on the same arrival sequence. Now define for $i = 1,2$

$$X_0^{\Sigma_i} = \begin{cases} \min(y_1(t), y_2(t), \ldots, y_n(t)) & \text{if } N^{\Sigma_i}(0) = n > 0 \\ 0 & \text{if } N^{\Sigma_i}(0) = 0 \end{cases}$$

$$X_k^{\Sigma_i} = \min(y_1(T_k^a), y_2(T_k^a), \ldots, y_n(T_k^a)) \quad \text{for } N^{\Sigma_i}(T_k^a) = n > 0,\ k = 1,2,\ldots$$

$$X^{\Sigma_i}(t) = \begin{cases} X_0^{\Sigma_i} - t/N^{\Sigma_i}(0) & 0 < t < T_1^a \\ X_k^{\Sigma_i} - t/N^{\Sigma_i}(T_k^a) & T_k^a < t < T_{k+1}^a \quad k = 1,2,\ldots \end{cases}$$

$$V^{\Sigma_i} = \begin{cases} \inf\limits_{t>0} \{t:\ X^{\Sigma_i}(t) = 0\} & \text{if } N^{\Sigma_i}(0) > 0 \\ \inf\limits_{t>T_1^a} \{t:\ X^{\Sigma_i}(t) = 0\} & \text{if } N^{\Sigma_i}(0) = 0\ . \end{cases}$$

Thus, the process $\{X^{\Sigma_i}(t),\ t > 0\}$ charts the smallest remaining service requirement over time. $\{X^{\Sigma_i}(t),\ t > 0\}$ is strictly decreasing between the times that it hits the x-axis. V^{Σ_i} is the time that $X^{\Sigma_i}(t)$ hits

zero; i.e. it is the forward recurrence time until the first departure.

We need the following Lemma for general systems Σ_1 and Σ_2.

Lemma 5.7. If the service requirement distributions F_1 and F_2 are such that $E(S^{\Sigma_1}) = E(S^{\Sigma_2}) = m$ and $S^{\Sigma_1} \leq_c S^{\Sigma_2}$, then $X_0^{\Sigma_1} \leq_d X_0^{\Sigma_2}$.

Proof: $\Pr(X_0^{\Sigma_i} > x | N_0^{\Sigma_i} = n) = \begin{cases} 1 & n = 0 \\ [\frac{1}{m} \int_x^\infty \bar{F}_i(t)dt]^n & n > 0 \end{cases}$.

Since $S^{\Sigma_1} \leq_c S^{\Sigma_2}$,

$$\int_x^\infty \bar{F}_1(t)dt \leq \int_x^\infty \bar{F}_2(t)dt \text{ for all } x$$

and hence

$$[\frac{1}{m} \int_x^\infty \bar{F}_1(t)dt]^n \leq [\frac{1}{m} \int_x^\infty \bar{F}_2(t)dt]^n \text{ for each } n.$$

Now since $N_0^{\Sigma_1} =_d N_0^{\Sigma_2}$, we have

$\Pr(X_0^{\Sigma_1} > x) \leq \Pr(X_0^{\Sigma_2} > x)$, or $X_0^{\Sigma_1} \leq_d X_0^{\Sigma_2}$.

Now for Σ_1 and Σ_2 we choose a particular realization for N_0 (recall $N_0^{\Sigma_1} =_d N_0^{\Sigma_2}$), and a realization of an arrival sequence $\{T_1^{a'}, T_2^{a'}, \ldots\}$ from a Poisson process with parameter λ/q. Consider the evolution of $\{X^{\Sigma_i}(t), t > 0\}$. Figure 5.1 shows a particular realization of this process over time for Σ_1 and Σ_2. In this realization, Σ_2 has an initial smallest remaining service requirement that is larger than in Σ_1, but because arrivals occur with smaller service requirements, the first output occurs first in Σ_2. The crucial feature in this system is that Σ_2 can make jumps downward when an

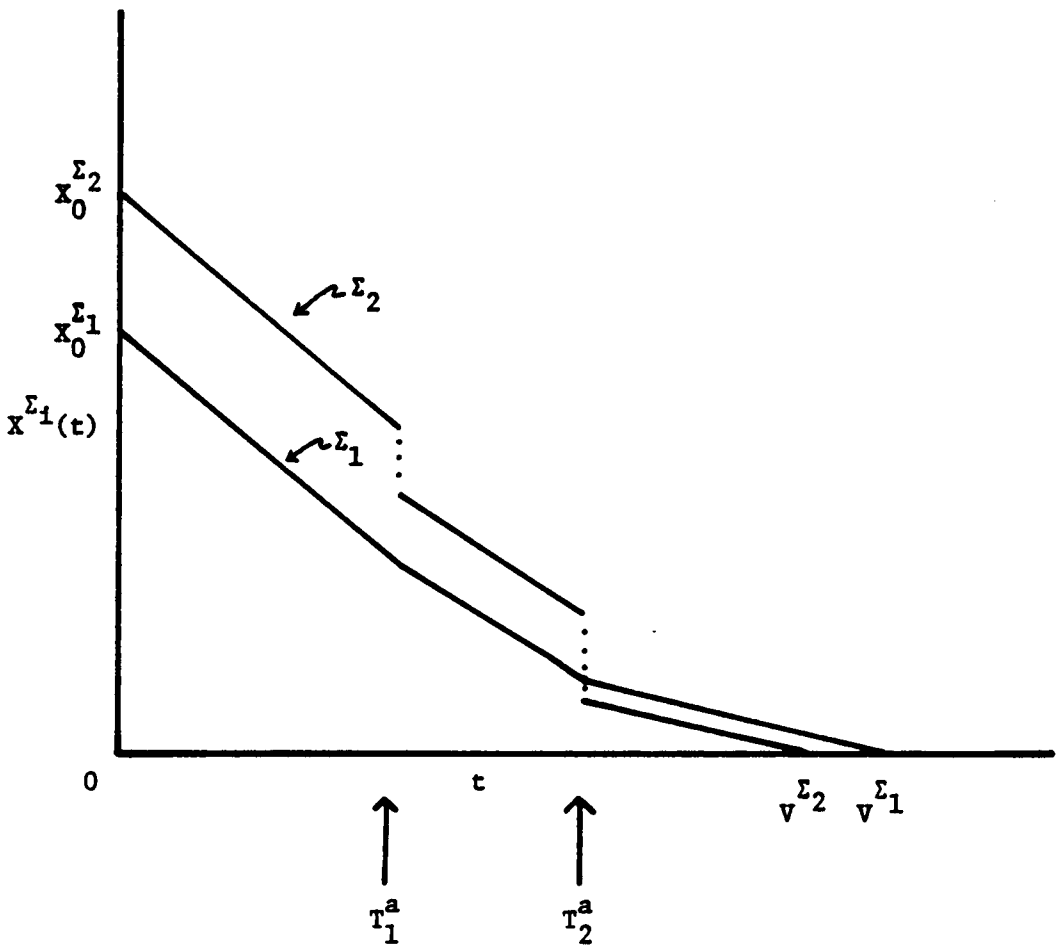

Figure 5.1 Evolution of $X^{\Sigma_i}(t)$ with Poisson arrival process, rate $\frac{\lambda}{q}$

arrival occurs with a smaller service requirement than the smallest one currently in the system. Σ_1 cannot make jumps downward, because all new arrivals have service requirement 1.

Now define two sets as follows (α refers to the arrival parameter of the Poisson arrival process):

$$A^{\alpha} = \{\omega: \ V^{\Sigma_1}(\omega) > V^{\Sigma_2}(\omega)\}$$

$$B^{\alpha} = \{\omega: \ V^{\Sigma_1}(\omega) < V^{\Sigma_2}(\omega)\}.$$

Theorem 5.8. When the arrival parameter is λ/q, $\Pr(A^{\lambda/q}) = \Pr(B^{\lambda/q})$.

Proof: We can write

$$\Pr(A^{\lambda/q}) = \int_0^{\infty} \Pr(V^{\Sigma_1} > y | V^{\Sigma_2} = y) d\Pr(V^{\Sigma_2} < y)$$

$$\Pr(B^{\lambda/q}) = \int_0^{\infty} \Pr(V^{\Sigma_2} > x | V^{\Sigma_1} = x) d\Pr(V^{\Sigma_1} < x).$$

Now when the arrivals occur according to a Poisson process with rate λ/q, V^{Σ_1} and V^{Σ_2} have the same marginal distribution, since the departure process is a Poisson process irrespective of the service time distribution. Clearly they are independent random variables. Hence

$$\Pr(A^{\lambda/q}) = \int_0^{\infty} \Pr(V^{\Sigma_1} > y) d\Pr(V^{\Sigma_2} < y)$$

$$= \int_0^{\infty} \Pr(V^{\Sigma_2} > x) d\Pr(V^{\Sigma_1} < x) = \Pr(B^{\lambda/q}).$$

With the arrival rate α, define three other sets as follows:

$$R_1^{\alpha} = \{\omega: \ X_0^{\Sigma_1} > X_0^{\Sigma_2}\}$$

$$R_2^{\alpha} = \{\omega: \ X_0^{\Sigma_1} < X_0^{\Sigma_2}, V_1^{\Sigma_1} < V_1^{\Sigma_2}\}$$

$$R_3^{\alpha} = \{\omega: \ X_0^{\Sigma 1} < X_0^{\Sigma 2}, V_1^{\Sigma 1} > V_1^{\Sigma 2}\}.$$

Then the sets A^{α} and B^{α} above can be written as

$$A^{\alpha} = R_1^{\alpha} \cup R_3^{\alpha}$$

$$B^{\alpha} = R_2^{\alpha}.$$

This is so because as long as the initial smallest remaining service time in Σ_1 is greater than in Σ_2, the path of $X^{\Sigma 1}(t)$ will always be above that of $X^{\Sigma 2}(t)$. If $X_0^{\Sigma 1} < X_0^{\Sigma 2}$, however, the path of $X^{\Sigma 2}(t)$ may or may not jump below the path of $X^{\Sigma 1}(t)$.

Theorem 5.9. When the arrival parameter is λ, $Pr(A^{\lambda}) < Pr(B^{\lambda})$ and $V^{\Sigma 1} <_d V^{\Sigma 2}$.

Proof: See Figure 5.1. Starting with a sample path from the process $\{X^{\Sigma i}(t), t > 0\}$ with arrival rate λ/q, produce a sample path from the process with arrival rate λ after time zero by deleting an arrival with probability $1 - q$. Call a sample path in the λ/q case ω and a sample path in the λ case ω'. Then

$$\omega \in R_1^{\lambda/q} \Longrightarrow \omega' \in R_1^{\lambda}$$

$$\omega \in R_3^{\lambda/q} \Longrightarrow \omega' \in R_3^{\lambda}$$

$$\omega \in R_2^{\lambda/q} \not\Longrightarrow \omega' \in R_2^{\lambda}.$$

Some sample paths ω in $R_2^{\lambda/q}$ will have ω' paths in R_3^{λ}, because deleting an arrival may cause the Σ_2 system to produce a departure after the Σ_1 output. Figure 5.2 illustrates this possibility. Hence

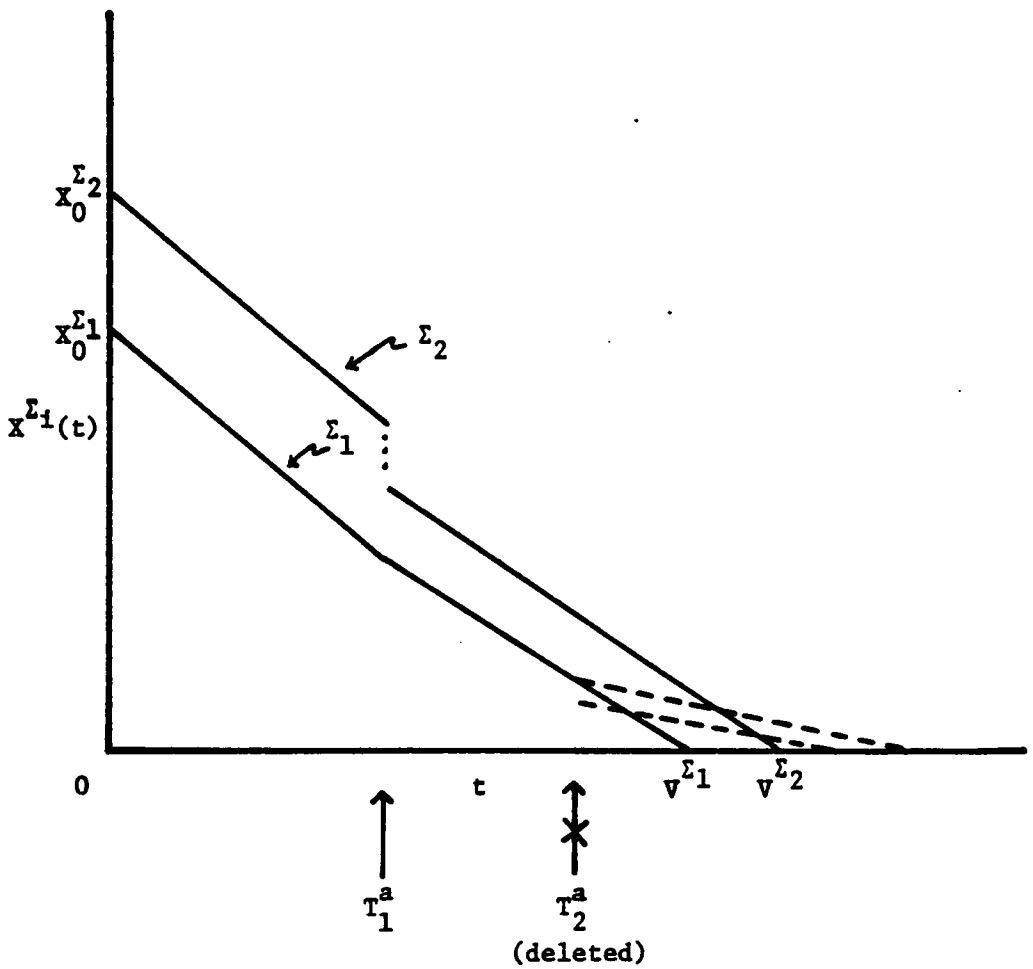

Figure 5.2 Evolution of $x^{\Sigma_i}(t)$ with Poisson arrival process, rate λ

$$R_1^{\lambda/q} = R_1^{\lambda}, \quad R_2^{\lambda/q} \supset R_2^{\lambda}, \quad R_3^{\lambda/q} \subset R_3^{\lambda}$$

and

$$A^{\lambda/q} \supset A^{\lambda} \quad B^{\lambda/q} \subset B^{\lambda}$$

and therefore $Pr(A^{\lambda}) < Pr(B^{\lambda})$ and thus by Theorem 5.8 $V^{\Sigma_1} <_d V^{\Sigma_2}$.

Theorem 5.10. $D^{\Sigma_1} <_c D^{\Sigma_2}$.

Proof: $E(D^{\Sigma_1}) = E(D^{\Sigma_2})$ and $V^{\Sigma_1} <_d V^{\Sigma_2}$, hence by Theorem 2.1 the result follows.

Notice that nowhere in Theorem 5.9 or Theorem 5.10 have we used the fact that Σ_2 has exponential service requirements. That Σ_1 has constant service times is crucial to keep one of the sample paths jump free. Thus our results will generalize to any Σ_2 with GI service times compared against the given Σ_1 system. It is also not crucial that $E(S^{\Sigma_1}) = 1$, but that $E(S^{\Sigma_1}) = E(S^{\Sigma_2})$. Therefore we state without proof the following result.

Theorem 5.11. For two systems Σ_1 and Σ_2, where Σ_1 has constant service requirements and Σ_2 has general service requirements with $E(S^{\Sigma_1}) = E(S^{\Sigma_2})$ and both have Poisson arrivals at rate λ, $D^{\Sigma_1} <_c D^{\Sigma_2}$.

The key idea is that any service requirement distribution has some probability that an arrival may become the next imminent departure. Heuristically what is happening here is that in the stationary case

(i.e. if the arrival parameter were λ/q after time zero as well as before, jumps downward in the general service requirement system occur at a fast enough rate to compensate for the stochastically greater initial smallest remaining service time. When we consider the output process from the feedback queue, however, we are no longer dealing with a stationary system, but one in which arrivals occur at a slower rate. For the output process, there are not enough downward jumps to compensate for the greater initial remaining service requirement, so the interoutput time is convexly slower than for constant service times.

5.6 Summary

In this chapter, we analyzed the processor sharing queue with general independent service requirements. We obtained a Markovian state space by appending the queue length with a vector of accumulated service times of those units in service. We showed that the detailed balance equations hold when the reverse process is defined as in Section 5.2, and we then solved for the stationary queue length distribution. We showed that the distribution is insensitive to the service time distribution given its mean, and that in equilibrium, the accumulated service times for each unit are i.i.d. random variables.

We analyzed the output process by considering the Markov process of queue length and accumulated service times on each unit's current pass through the system. Thus the time until the first output after zero has the same distribution as the time until the first departure after zero from a particular non-stationary processor sharing queue. We compared this distribution for the special case where one system has

general service requirements and the other constant service requirements, and showed that the interoutput time is convexly longer in the constant service requirement system.

From this discussion, we conjecture that the following result holds: if, for two processor sharing queues with Poisson (λ) input and service times S_1 and S_2 (distribution functions F_1 and F_2), where $E(S_1) = E(S_2)$ and $S_1 \leq_c S_2$, then $D^o \leq_c D^o$. We have been unable to prove this result for general service distributions F_1 and F_2. Intuitively, it appears that for two service time distributions with equal means but different variances, the interoutput times will be more variable when the service requirement is more variable. Actually, one would think that this would be true about the departure process as well; the Poisson nature of the departure process is quite surprising.

CHAPTER 6

THE INFINITE SERVER QUEUE WITH FEEDBACK

6.1 Introduction

In this chapter we analyze another symmetric queueing discipline, the infinite server queue. As indicated in Section 2.5, these systems can be modelled as processor sharing queues, where the function $\phi(\cdot)$ takes the form $\phi(n) = n$ for all n. However, they have a simpler structure than other processor sharing queues because units have no influence on each other in the course of receiving service. There are no delays due to units competing for the service resource, and thus the time spent by a unit in the system is only its service time.

Our goal here is to study the various queue length and traffic processes in this system. In the analysis we will proceed along the lines of Chapters 4 and 5, but where we were able to obtain only qualitative results for flow processes in Chapter 5, we will see that in the infinite server queue we can obtain the distribution of the time until the first output for certain service requirements. Thus in some cases we can measure exactly how different the intervals in the output processes are for different service times. This is more than we could do for the processor sharing queue because the actual distributions were difficult to compute.

In this chapter, we also give some insight into the ideas of quasi reversibility (from Chapter 2) and how it applies in a network context. We show in Section 6.4 that the output process from the M/M/∞-IBF queue is a sequence of dependent intervals, and thereby show that an important property of a quasi-reversible node in isolation is lost when

the node is part of a feedback network. (See the discussion in Section 3.4.)

Throughout the chapter, we will draw comparisons with the results of previous chapters. We will point out results for the present system that were not obtained in Chapters 4 and 5. Thus while the infinite server queue shares common properties with the single server processor sharing queue, its simpler structure makes it a system worth studying in its own right.

As in Chapter 5, we consider a system at which units arrive according to a Poisson stream with rate λ. Service requirements are assigned independently according to distribution function $F(\cdot)$. There are infinitely many servers, so each unit's service requirement decreases linearly at rate 1 until it reaches zero. Upon completion of a service requirement, a unit departs with probability q and feeds back with probability p ($p + q = 1$).

6.2 Queue Length Results

To study the time stationary queue length process, we again use the construction in Section 4.2 to convert the queue with feedback into an equivalent queue without feedback that has the same queue length properties. As in Section 5.2 we define the state process to be a vector valued process $(\mathcal{N}, \underset{\sim}{\mathcal{Y}}) = \{N(t), \underset{\sim}{Y}(t),\ t \in R^+\}$, where $N(t)$ is the queue length at time t and $\underset{\sim}{Y}(t)$ is the ordered vector of accumulated service times of each unit in the system at time t. The M/GI/∞ queue that we consider has service requirement distribution $U(\cdot)$, where $U(\cdot)$ is defined as in Section 4.2. We define the reverse process $(\mathcal{N}^r, \underset{\sim}{\mathcal{Y}}^r) = \{N^r(t), \underset{\sim}{Y}^r(t),\ t \in R^+\}$ where $N^r(t)$ is the queue

length at time t and $\underset{\sim}{Y}^r(t)$ the vector of remaining service times at time t. Then the rates at which events occur in the forward and reverse processes are given below.

Arrivals (F) $(n,\underset{\sim}{y}) \longrightarrow (n+1,(0,\underset{\sim}{y}))$ at rate λ

(R) $(n+1,(0,\underset{\sim}{y})) \longrightarrow (n,\underset{\sim}{y})$ at rate 1

Departures (F) $(n,\underset{\sim}{y}) \longrightarrow (n-1,e_i(\underset{\sim}{y}))$ at rate $\mu(y_i)$

(R) $(n-1,e_i(\underset{\sim}{y})) \longrightarrow (n,\underset{\sim}{y})$ at rate $\lambda u(y_i)$

where $\underset{\sim}{y}$ and $e_i(\underset{\sim}{y})$ and $\mu(\cdot)$ are defined as in Section 5.4. These rates are obtained by noting that both real time and accumulated service time are on the same scale in the infinite server queue. In the processor sharing queue, one unit of real time was "worth" only 1/n units of accumulated service time. We can now use our reversibility ideas (see Section 2.3) to determine the stationary density $\{\pi(0),\pi_k(\underset{\sim}{y})\}$.

Theorem 6.1. The unique stationary density $\{\pi(0),\pi_k(\underset{\sim}{y})\}$ exists provided $\lambda,\mu < \infty$ and is given by

$$\begin{aligned} \pi(\emptyset) &= e^{-\lambda/q\mu} \\ \pi_k(\underset{\sim}{y}) &= \lambda^k e^{-\lambda/q\mu} \prod_{i=1}^{k} \bar{U}(y_i) \end{aligned} \tag{6.1}$$

for $\lambda,\mu < \infty$.

Proof: The conditions for stationarity for the M/GI/∞ queue are $\lambda E(S^r) < \infty$ (Gross and Harris (1974)); thus in our case we require $\lambda/q\mu < \infty$, and since $0 < q < 1$, the queue will have a stationary distribution iff $\lambda,\mu < \infty$. To verify that (6.1) is the (unique) stationary probability distribution, we need to show that (6.1)

satisfies the detailed balance equations, since a solution to the detailed balance equations is a solution to the global balance equations (Section 2.3). Using the rates above, the equations are

$$\pi_k(\underset{\sim}{y})\lambda = \pi_{k+1}(0,\underset{\sim}{y})$$

$$\pi_k(\underset{\sim}{y})\mu(y_i) = \pi_{k-1}(e_i(\underset{\sim}{y}))\lambda u(y_i)$$

$$\pi(\phi)\lambda = \pi_1(y)\mu(y)$$

and the normalizing condition is

$$1 = \pi(\phi) + \sum_{k=1}^{\infty} \int_0^{\infty} \cdots \int_0^{\infty} \pi_k(\underset{\sim}{y})d\underset{\sim}{y}.$$

Clearly (6.1) is a solution; e.g.

$$\pi_k(\underset{\sim}{y})\mu(y_i) = \lambda^k e^{-\lambda/q\mu} \prod_{j=1}^{k} \bar{U}(y_j) \cdot \frac{u(y_i)}{\bar{U}(y_i)}$$

$$= \lambda^{k-1} e^{-\lambda/q\mu} \prod_{\substack{j=1 \\ j\neq i}}^{k} \bar{U}(y_j) \cdot \lambda u(y_i) = \pi_{k-1}(e_i(\underset{\sim}{y}))\lambda u(y_i).$$

Since with the $\pi(\phi)$ in the Theorem

$$\pi(\phi) + \sum_{k=1}^{\infty} \int_{y_1<y_2<\cdots<y_k} \cdots \int \pi_k(\underset{\sim}{y})d\underset{\sim}{y}$$

$$= e^{-\lambda/q\mu} + \sum_{k=1}^{\infty} \frac{1}{k!} \int_{y_1<y_2<\cdots<y_k} \cdots \int \pi_k(\underset{\sim}{y})d\underset{\sim}{y}$$

$$= e^{-\lambda/q\mu} + e^{-\lambda/q\mu} \sum_{k=1}^{\infty} \frac{(\lambda/q\mu)^k}{k!} = e^{-\lambda/q\mu}\left(\sum_{k=0}^{\infty} \frac{(\lambda/q\mu)^k}{k!}\right)$$

$$= 1,$$

$\{\pi(\phi), \pi_k(\underset{\sim}{y})\}$ is the unique stationary probability density for $(\mathcal{N}, \underset{\sim}{\mathcal{Y}})$.

Notice that the marginal queue length distribution is Poisson with parameter $\lambda/q\mu$, the traffic intensity. This follows since

$$\pi_k = \lambda^k e^{-\lambda/q\mu} \int_{y_1<y_2<\cdots<y_k} \prod_{i=1}^{k} \bar{U}(y_i)dy_2dy_2\cdots dy_k$$

$$= \frac{\lambda^k e^{-\lambda/q\mu}}{k!} \prod_{i=1}^{k} E(S') = \frac{(\lambda/q\mu)^k e^{-\lambda/q\mu}}{k!} .$$

Also, as we found for the processor sharing queue, given that there are $k > 0$ units in the system, the accumulated service times of each unit in equilibrium are independent, identically distributed random variables with distribution

$$\Pr(y_i < y | N(0) = k) = \frac{\Pr(y_i < y, N(0) = k)}{\Pr(N(0) = k)}$$

$$= \frac{\int_{y_1<y_2<\cdots<y_{i-1}<0<y_{i+1}<\cdots<y_k}^{y} \pi_k(\underset{\sim}{y})dy_1dy_2\cdots dy_k}{\pi_k}$$

$$= \frac{\int_0^y \int_0^\infty \cdots \int_0^\infty (e^{-\lambda/q\mu})\lambda^k \frac{1}{k!} \bar{U}(t) \prod_{\substack{j=1 \\ j\neq i}}^{k} \bar{U}(y_j)dy_jdt}{\frac{(e^{-\lambda/q\mu})(\frac{\lambda}{q\mu})^k}{k!}}$$

$$= \frac{e^{-\lambda/q\mu}\lambda^k(\frac{1}{q\mu})^{k-1} \frac{1}{k!} \int_0^y \bar{U}(t)dt}{\frac{(e^{-\lambda/q\mu})(\frac{\lambda}{q\mu})^k}{k!}}$$

$$= q\mu \int_0^y \bar{U}(t)dt.$$

Independence is apparent from the product form of (6.1).

6.3 Departure Process

We have shown in the previous section that $(\mathcal{N}, \underset{\sim}{\mathcal{Y}})$ is a reversible process, with $(\mathcal{N}^r, \underset{\sim}{\mathcal{Y}}^r)$ defined as above since $\pi_k(\underset{\sim}{y})$ satisfies the detailed balance equation. From this fact, a proof of

the theorem below follows as does that of Theorem 4.4.

Theorem 6.2. In equilibrium the departure process from the M/GI/∞-IBF queue is a Poisson process with rate λ.

Kendall (1964) gives an intuitive argument for the fact that the departure process from the M/GI/∞ queue (without feedback) is a Poisson process. If the incoming units are sorted according to their service requirements, the arrival process can be viewed as a family of independent Poisson inputs. Since units flow independently through the system, the departure times for each unit is a fixed shift of the input time, and hence the departure process is just the independent superposition of the individually shifted inputs, and thus again Poisson.

The independence of the time spent by units in the system is crucial in this intuitive argument. As we saw in Section 5.3, the departure process from the processor sharing queue is also a Poisson process in equilibrium, but clearly the times spent by units in the system are not independent. One wonders if a similar intuitive explanation can be constructed for the processor sharing queue.

In addition, the independence of units in the infinite server queue leads to many simple results, and one might suppose that if any symmetric queue would possess insensitive interoutput intervals, it would be this queue. In the following section we show that even in the M/GI/∞-IBF queue, the internal flow are not insensitive to changes in the service requirement distribution when the queue is part of a

feedback network.

6.4 Output Process

We begin this section with some structural results that show that the output process from the infinite server queue with feedback is not a renewal process. Walrand (1982) extended the "loop criteria" (Section 3.2) to networks of symmetric queues with bounded service rates; by a similar argument we show that for a certain class of queues with unbounded service rates (namely, infinite server queues with exponential service requirements), the loop criteria are necessary for the output process to be renewal.

Let us decompose the M/M/∞-IBF queue into a system of parallel queues in which we can define a flow process corresponding to the output process in the original (feedback) queue. Suppose that upon arrival to the queue, the points in a Poisson (λ) stream are independently assigned a mark from the set $\{1,2,3,\ldots\}$. The mark takes value i with probability qp^{i-1} and represents the number of service requirements to be assigned to the unit (i.e., the number of times the unit feeds back plus one). All units with mark i are assigned to "service bank" i. Service bank i operates as i $\cdot$/M/∞ queues in tandem. Pictorially we represent the constructed system in Figure 6.1.

Since the assignment of marks geometrically thins the original Poisson arrival stream, the arrival process to each service bank is a Poisson process with the appropriate parameter (eg., λqp^{k-1} for bank k) and is independent of the arrival process to any other service bank. We shall refer to any of the k tandem queues in service bank k as a

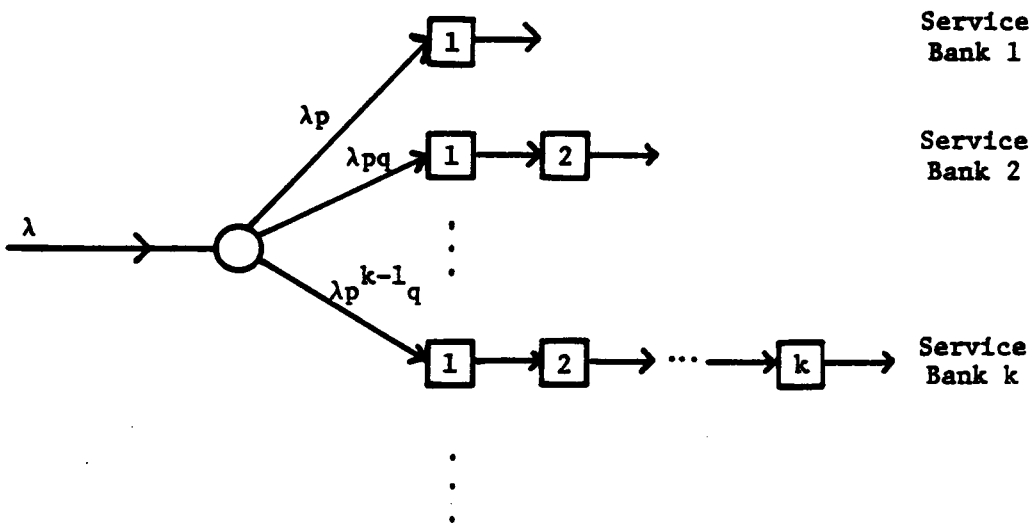

Figure 6.1 The decomposed infinite server queue

stage. The completion of any stage in the tandem sequence corresponds to an output in the original queue.

Define

$T'_n(i,k)$ = time of n^{th} stage i completion from service bank k,

$i = 1,2,\ldots,k,\ n = 1,2,3,\ldots,\ k = 1,2,3,\ldots$

$\mathcal{T}'_k = \bigcup_{i=1}^{k} \{T'_n(i,k),\ n = 1,2,3,\ldots\},\ k = 1,2,3,\ldots$

Then $\mathcal{T}'_k$ is the output process from service bank k in the constructed system, with the understanding that whenever we take the union of point processes, the points are reordered so that $T_1 < T_2 <\ldots$

Since units are assigned to service banks independently and service times are assigned independently, $\mathcal{T}'_k$ is independent of $\mathcal{T}'_j$ and $\mathcal{T}'_k \cap \mathcal{T}'_j = \emptyset$ for $j \neq k$. Moreover

$$\mathcal{T}^o = \bigcup_{k=1}^{\infty} \mathcal{T}'_k,$$

that is, the output process in the original queue is the superposition of the independent output processes in the new system.

Lemma 6.3. $\mathcal{T}'_k$ is a sequence of dependent random variables for $k > 1$.

Proof: Let $k = 2$. Assume the system is in equilibrium at time 0. The system consists of two $\cdot/M/\infty$ queues in tandem (Figure 6.2) with a Poisson arrival process (rate λqp) to the first stage. Define

T_t = time of first output (from either queue) after time t,

$C^{oi}(t)$ = number of outputs in (0,t) from stage i,

$C^o(t)$ = number of outputs in (0,t),

$N_i(t)$ = number of units in stage i at time t (i = 1,2),

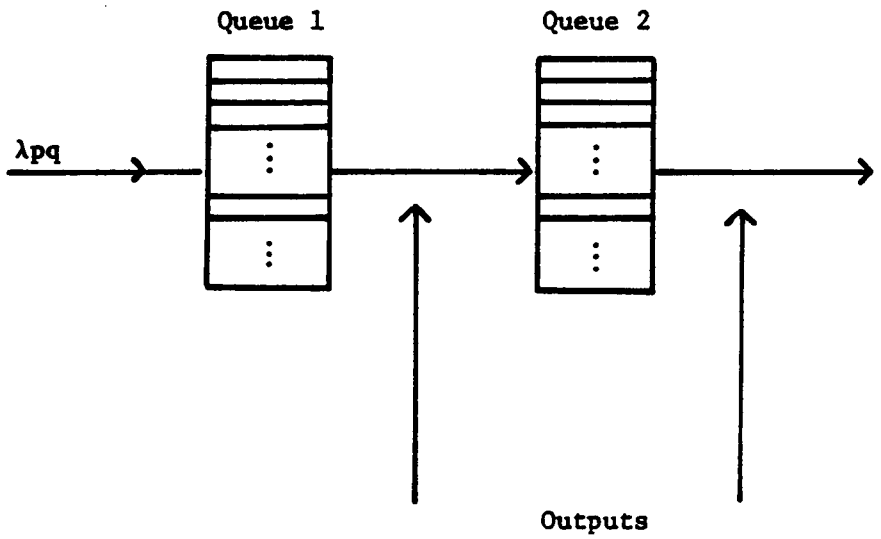

Figure 6.2 Service bank 2

$$N(t) = N_1(t) + N_2(t).$$

Because of the one-to-one relationship between the interval process and the counting process (Section 2.4) it suffices to show that $\Pr(T_t > \tau) \neq \Pr(T_t > \tau | C^o(t))$ for $\tau > t$.

Now $\Pr(T_t > \tau) = \sum_{n=0}^{\infty} \Pr(T_t > \tau | N(t) = n)\Pr(N(t) = n)$

$$= \sum_{n=0}^{\infty} \sum_{j=0}^{n} \Pr(T_t > \tau | N(t) = n)\Pr(N_1(t) = j, N_2(t) = n-j)$$

$$= \sum_{n=0}^{\infty} \sum_{j=0}^{n} e^{-n\mu(\tau-t)} \pi_j \pi_{n-j}.$$

The last equality follows because the service times of units are independent exponentially distributed random variables, and because in equilibrium the queue lengths at queue 1 and queue 2 are independent and each has distribution $\{\pi_k\}$.

If T_t is to be independent of $C^o(t)$ it must be independent of $C^{o1}(t)$. We will show that $\Pr(T_t > \tau | C^{o1}(t) = \ell) \neq \Pr(T_t > \tau)$.

$$\Pr(T_t > \tau | C^{o1}(t) = \ell) = \sum_{n=0}^{\infty} \Pr(T_t > \tau | N(t) = n, C^{o1}(t) = \ell)\Pr(N(t) = n | C^{o1}(t) = \ell)$$

$$= \sum_{n=0}^{\infty} \Pr(T_t > t | N(t) = n)\Pr(N(t) = n | C^{o1}(t) = \ell),$$

since the remaining service times of units in queue at time t are not affected by the output process prior to t (Kelly (1979)).

$$\sum_{n=0}^{\infty} \Pr(T_t > \tau | N(t) = n)\Pr(N(t) = n | C^{o1}(t) = \ell)$$

$$= \sum_{n=0}^{\infty} \sum_{j=0}^{n} e^{-n\mu(\tau-t)} \Pr(N_1(t) = j, N_2(t) = n-j | C^{o1}(t) = \ell)$$

$$= \sum_{n=0}^{\infty} \sum_{j=0}^{n} e^{-n\mu(\tau-t)} \pi(j) \Pr(N_2(t) = n-j | C^{o1}(t) = \ell)$$

$$\neq \sum_{n=0}^{\infty} \sum_{j=0}^{n} e^{-n\mu(\tau-t)} \pi(j)\pi(n-j).$$

The gist of the proof is that outputs from queue 1 in (0,t) change the expected number in queue 2 and hence the chance of an output in the future. This argument extends to any k > 2 service bank to show that the output process from each service bank k (k > 1) is a sequence of dependent intervals.

Theorem 6.4. In equilibrium $\mathcal{T}^o$ is a sequence of dependent random variables.

Proof: Consider an arrival to the M/M/∞-IBF system. With positive probability, the arrival will enter some service bank k > 1. Since $\mathcal{T}_k'^o$ is a sequence of dependent random variables, and since $\mathcal{T}_k'^o \subset \mathcal{T}^o$, the result follows.

Thus even in the M/M/∞-IBF queue, the output process is not a renewal process. To completely characterize the process, then, we need to describe all finite dimensional interval distributions. As we did in the processor sharing queue, we will work with single intervals, as the process itself is complex and not easily tractable analytically.

Using the methods of Section 5.4, we can analyze the distribution of a single interval in the output process. Again, we define a Markov process $(\mathcal{N}, \underset{\sim}{\mathcal{Y}}) = \{N(t), \underset{\sim}{Y}(t),\ t \in \mathbb{R}^+\}$ where now

$$\underset{\sim}{Y}(t) = (y_1(t), y_2(t), \ldots, y_n(t))$$

is the vector of accumulated service times on each unit's current pass in the server. With the similarly defined reverse process $(\mathcal{N}^r, \underset{\sim}{\mathcal{Y}}^r)$ (when $\underset{\sim}{Y}(t)$ is a vector of remaining service times), we can give the

transition rates at jump points for these processes (with $\nu(y) = \frac{f(y)}{\bar{F}(y)}$):

Arrivals (F) $(n,\underset{\sim}{y}) \longrightarrow (n+1,(0,\underset{\sim}{y}))$ at rate λ

(R) $(n+1,(0,\underset{\sim}{y})) \longrightarrow (n,\underset{\sim}{y})$ at rate q

Departures (F) $(n,\underset{\sim}{y}) \longrightarrow (n-1,e_i(\underset{\sim}{y}))$ at rate $q\nu(y_i)$

(R) $(n-1,e_i(\underset{\sim}{y})) \longrightarrow (n,\underset{\sim}{y})$ at rate $\lambda f(y_i)$

Feedbacks (F) $(n,\underset{\sim}{y}) \longrightarrow (n,(0,e_i(\underset{\sim}{y})))$ at rate $p\nu(y_i)$

(R) $(n,(0,e_i(\underset{\sim}{y}))) \longrightarrow (n,\underset{\sim}{y})$ at rate $pf(y_i)$.

These rates look very similar to those in Section 5.4. Since in the infinite server queue each unit is being served at rate 1, the rates for service completions are not divided by n as before.

Theorem 6.5. The stationary density $\{\pi(\emptyset),\pi_k(\underset{\sim}{y})\}$ is given by

$$\begin{aligned} \pi(\emptyset) &= e^{-\lambda/q\mu} \\ \pi_k(\underset{\sim}{y}) &= \left(\frac{\lambda}{q}\right)^k e^{-\lambda/q\mu} \prod_{i=1}^{k} \bar{F}(y_i) \end{aligned} \tag{6.2}$$

for $0 < \lambda,\mu < \infty$, $0 < q \leq 1$.

Proof: The detailed balance equations for this system are

$$\pi_k(\underset{\sim}{y})\lambda = \pi_{k+1}(0,\underset{\sim}{y})q$$

$$\pi_k(\underset{\sim}{y})q\nu(y_i) = \pi_{k-1}(e_i(\underset{\sim}{y}))\lambda f(y_i)$$

$$\pi_k(\underset{\sim}{y})p\nu(y_i) = \pi_k(0,e_i(\underset{\sim}{y}))pf(y_i)$$

$$\pi(\emptyset)\lambda = \pi_1(y)q\nu(y).$$

And the normalizing equation is

$$1 = \pi(\emptyset) + \sum_{k=1}^{\infty} \int_0^{\infty} \cdots \int_0^{\infty} \pi_k(\underline{y})d\underline{y}.$$

The hypothesized density solves these equations; eg.

$$\pi_k(\underline{y})p\mu(y_i) = (\frac{\lambda}{q})^k e^{-\lambda/q\mu} \prod_{j=1}^{k} \bar{F}(y_i)p \frac{f(y_i)}{\bar{F}(y_i)} = (\frac{\lambda}{q})^k e^{-\lambda/q} \prod_{\substack{j=1 \\ j\neq i}}^{k} \bar{F}(y_i)pf(y_i)$$

$$= \pi_{k-1}(0,e_i(\underline{y}))pf(y_i).$$

$\{\pi(\emptyset),\pi_k(\underline{y})\}$ is a proper probability density provided $0 < \lambda,\mu < \infty$.

From Theorem 6.5, the marginal queue length distribution is given by:

$$\pi_0 = \pi(\emptyset) = e^{-\lambda/q\mu}$$

$$\pi_k = \int_{y_1<\ldots<y_k} \pi_k(\underline{y})d\underline{y} = \frac{(e^{-\lambda/q\mu})(\lambda/q\mu)^k}{k!} \quad k = 1,2,\ldots.$$

Given that there are $k > 0$ units in the system, their accumulated service times are independent and identically distributed with distribution

$$\Pr(y_i < y|N(0)=k) = \frac{\Pr(y_i < y, N(0)=k)}{\Pr(N(0)=k)}$$

$$= \frac{\int_0^y [\int_0^{\infty} \cdots \int_0^{\infty} \frac{1}{k!} (\frac{\lambda}{q})^k (e^{-\lambda/q\mu}) \prod_{\substack{j=1 \\ j\neq i}}^{k} \bar{F}(y_j)dy_j]\bar{F}(y_i)dy_i}{\frac{(\frac{\lambda}{q\mu})^k (e^{-\lambda/q\mu})}{k!}}$$

$$= \frac{\frac{1}{k!} (\frac{\lambda}{q})^k (e^{-\lambda/q\mu})(\frac{1}{\mu})^{k-1} \int_0^y \bar{F}(t)dt}{\frac{(\frac{\lambda}{q\mu})^k (e^{-\lambda/q\mu})}{k!}}$$

$$= \mu \int_0^y \bar{F}(t)dt.$$

Independence follows since

$$\pi_k \prod_{i=1}^{k} \Pr(y_i < x_i) = (e^{-\lambda/q\mu})(\frac{\lambda}{q\mu})^k \prod_{i=1}^{k} \mu \int_0^{x_i} \bar{F}(y_i)dy_i$$

$$= (e^{-\lambda/q\mu})(\frac{\lambda}{q})^k \int_0^{x_1} \cdots \int_0^{x_k} \prod_{i=1}^{k} \bar{F}(y_i)dy_i$$

$$= \Pr(N(0) = k, y_1 < x_1, \ldots, y_k < x_k).$$

Again, as in Section 5.4 for the processor sharing queue, (6.2) is the stationary density of an M/GI/∞ queue without feedback whose arrival parameter is λ/q. We will proceed as we did for the single server processor sharing queue; that is, we will compute the time until the first departure after zero from a non-stationary M/GI/∞ queue that begins with density (6.2) at time zero and has arrival parameter λ. We will be able to compute the distribution of this random variable explicitly in certain cases, which we did not do in Chapter 5. From this analysis, we can give some ordering properties of the interoutput intervals.

6.5 Exponential Service Times

Let $F(t) = 1 - e^{-\mu t}$, so that service requirements are exponentially distributed with parameter μ. Assume that the system is stationary at time zero and define

V = time of the first output after 0,

$N^-(t)$ = number of service completions in $(0,t)$ from arrivals before 0,

$N^+(t)$ = number of service completions in $(0,t)$ from arrivals in $(0,t)$.

Now $\Pr(V > t) = \Pr(N^-(t)=0, N^+(t)=0) = \Pr(N^-(t)=0)\Pr(N^+(t)=0)$, since

the service completion times of units in service are independent. We can think of the system as two independent queues, one of which began at $-\infty$ with a Poisson $(\frac{\lambda}{q})$ arrival process that shuts off at 0 and one that starts with an empty queue at 0 and where arrivals occur according to a Poisson process with rate λ. Thus

$$\Pr(N^-(t)=0) = e^{-\frac{\lambda}{q\mu}} + \sum_{n=1}^{\infty} e^{-n\mu t} \frac{(\frac{\lambda}{q\mu})^n e^{-\frac{\lambda}{q\mu}}}{n!} = e^{-\frac{\lambda}{q\mu} - \frac{\lambda}{q\mu} e^{-\mu t}}.$$

From Mirasol (1963)

$$\Pr(N^+(t)=0) = e^{-\lambda t - \frac{\lambda}{\mu} e^{-\mu t} + \frac{\lambda}{\mu}}.$$

Thus

$$\Pr(V > t) = (e^{-\frac{\lambda}{q\mu} - \frac{\lambda}{q\mu} e^{-\mu t}})(e^{-\lambda t - \frac{\lambda}{\mu} e^{-\mu t} + \frac{\lambda}{\mu}})$$

$$= e^{-\lambda t - \frac{\lambda}{\mu}(\frac{1}{q} - 1) + \frac{\lambda}{\mu} e^{-\mu t}(\frac{1}{q} - 1)}$$

$$= e^{-\lambda t - \frac{p\lambda}{q\mu}(1 - e^{-\mu t})}.$$

Note that if $q = 1$, the time until the first output is exponential with parameter λ as expected.

6.6 Deterministic Service Times

Let Σ_1 and Σ_2 denote two infinite server queueing systems. Arrivals occur at each queue according to a Poisson process with rate λ. Service requirements in Σ_1 are constant with mean 1; service requirements in Σ_2 are equally likely to take values $\frac{1}{2}$ and $\frac{3}{2}$. Hence

$$F_1(x) = \begin{cases} 0 & x < 1 \\ 1 & x > 1 \end{cases} \qquad F_2(x) = \begin{cases} 0 & x < \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} < x < \frac{3}{2} \\ 1 & x > \frac{3}{2} \end{cases}$$

Since $E(S^{\Sigma_1}) = E(S^{\Sigma_2}) = 1$, and

$$\int_x^\infty \bar{F}(t)dt = \begin{cases} 1 - x & 0 < x < 1 \\ 0 & x > 1 \end{cases} \qquad \int_x^\infty \bar{F}_2(t)dt = \begin{cases} 1 - x & 0 < x < \frac{1}{2} \\ \frac{3}{4} - \frac{1}{2}x & \frac{1}{2} < x < \frac{3}{2} \\ 0 & x > \frac{3}{2} \end{cases}$$

it follows where that $S^{\Sigma_1} <_c S^{\Sigma_2}$.

We can compute the distribution of an interoutput interval by considering a Poisson arrival process with rate λ/q on $t \in (-\infty,0)$ and rate λ on $t \in [0,\infty)$. For Σ_1, $A(\cdot)$ is the number of arrivals in $(\cdot)$ in Σ_1 and $N(0)$ the number of units in the system at time 0. For Σ_2, we can consider the arrival process to be the superposition of two independent arrival process, each with rate $\lambda/2q$ on $t \in (-\infty,0)$ and $\lambda/2$ on $t \in [0,\infty)$. One arrival stream brings units with service requirement $\frac{1}{2}$ and the other brings units with service requirement $\frac{3}{2}$. Let $A_1(\cdot)$ be the number of arrivals in $(\cdot)$ with service requirement $\frac{1}{2}$ and $A_2(\cdot)$ the number of arrivals in $(\cdot)$ with service requirement $\frac{3}{2}$. Similarly define $N_1(0)$ as the number of units with service requirement $\frac{1}{2}$ in the system at 0, and $N_2(0)$ as the number of units with service requirement $\frac{3}{2}$ in the system at 0.

Now for Σ_1 we have

$$\Pr(V^{\Sigma_1} > t) = \begin{cases} \Pr(A(-1,t-1) = 0) & 0 < t < 1 \\ \Pr(N(0) = 0, A(0,t-1) = 0) & t > 1 \end{cases}$$

$$= \begin{cases} e^{\frac{-\lambda}{q} t} & 0 < t < 1 \\ e^{\frac{-\lambda}{q}} e^{-\lambda(t-1)} = e^{-\lambda(t + \frac{p}{q})} & t > 1 \end{cases}$$

and for Σ_2

$$\Pr(V^{\Sigma_2} > t) = \begin{cases} \Pr(A_1(-\tfrac{1}{2}, t - \tfrac{1}{2}) = 0,\ A_2(-\tfrac{3}{2}, t - \tfrac{3}{2}) = 0) & 0 < t < \tfrac{1}{2} \\ \Pr(N_1(0) = 0, A_1(0, t - \tfrac{1}{2}) = 0, A_2(-\tfrac{3}{2}, t - \tfrac{3}{2}) = 0) & \tfrac{1}{2} < t < \tfrac{3}{2} \\ \Pr(N_1(0) = 0, A_1(0, t - \tfrac{1}{2}) = 0, N_2(0) = 0, A_2(0, t - \tfrac{3}{2}) = 0) & t > \tfrac{3}{2} \end{cases}$$

$$= \begin{cases} e^{-\frac{\lambda}{2q} t} e^{-\frac{\lambda}{2q} t} = e^{-\frac{\lambda}{q} t} & 0 < t < \tfrac{1}{2} \\ e^{-\frac{\lambda}{2q}} e^{-\frac{\lambda}{2}(t - \frac{1}{2})} e^{-\frac{\lambda}{2q}(t - \frac{3}{2})} = e^{-\lambda(t(\frac{q+1}{2q}) + \frac{p}{4q})} & \tfrac{1}{2} < t < \tfrac{3}{2} \\ e^{-\frac{\lambda}{q}} e^{-\lambda(t-1)} = e^{-\lambda(t + \frac{p}{q})} & t > \tfrac{3}{2} \end{cases}$$

It then follows that $\Pr(V^{\Sigma_1} > t) > \Pr(V^{\Sigma_2} > t)$, and since $E(D^{o\Sigma_1}) = E(D^{o\Sigma_2})$, by Theorem 2.1, $D^{o\Sigma_1} <_c D^{o\Sigma_2}$. Figure 6.3 shows a graph of the complementary distribution functions of V^{Σ_1} and V^{Σ_2}.

6.7 Summary

In this chapter we have studied the infinite server queue with different service time distributions. We have shown that the stationary queue length processes at an arbitrary time, at an arrival time and at a departure time are identically distributed. Since the queue length distribution at an arbitrary (stationary) point in time is

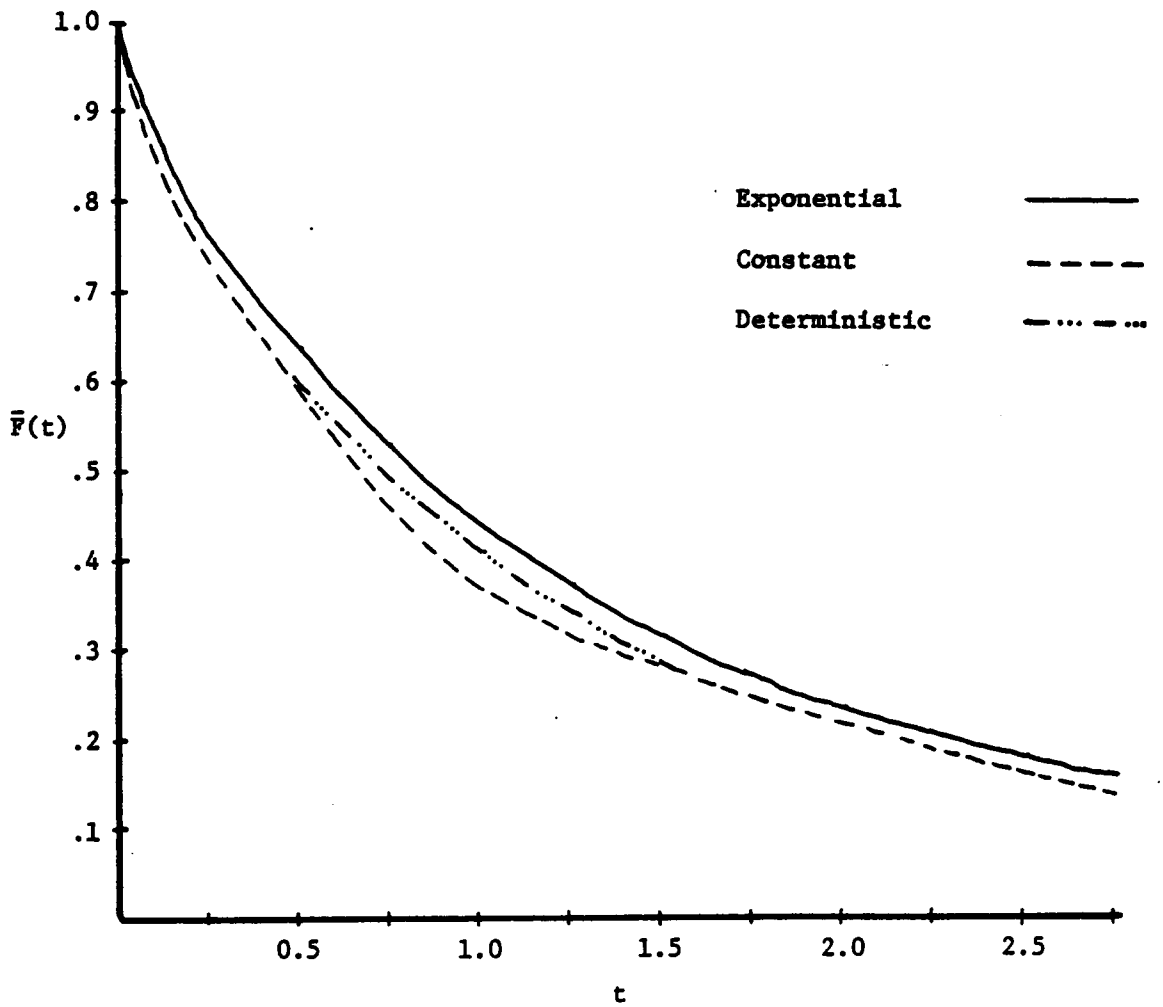

Figure 6.3 Comparison of $\bar{F}(t)$ for exponential, constant, and deterministic service times

insensitive to the service requirement distribution, given its mean, so are queue length distributions imbedded at arrivals and departures. For stationary infinite server queues with IBF and general service times, the queue length and accumulated service times of each unit in service are independent random variables, whether the state process keeps track of the total accumulated service time of a unit in the system or accumulated service itme on each unit's current pass through the server. The accumulated service times are identically distributed under either definition of the state process as well.

The departure process from the M/GI/∞-IBF queue is Poisson with the same parameter as the arrival process irrespective of the service time distribution. Thus the departure process in equilibrium is insensitive to the service time distribution, but we have shown that, in general, the output process does not inherit this insensitivity when the queue is part of this simple feedback network.

For the examples we have given, convexly smaller service requirements resulted in convexly smaller interoutput times. We believe this property to be true in general, although we have not proved it. In the concluding chapter of this paper, we will make some conjectures on order properties of output intervals in general networks of infinite server queues.

CHAPTER 7

CONCLUSIONS AND EXTENSIONS

7.1 Summary

In this paper, we have examined the effect of service discipline on certain properties of a simple queueing network (a queue with IBF). We considered the single server processor sharing case in Chapters 4 and 5 with both exponential and general service requirement distributions. For these systems, we gave conditions under which the queue length process had a stationary probability distribution. We studied the stationary queue length distribution in detail, in the time stationary and the various customer stationary (arrival, departure, input and output) points. We found that the queue length distributions at all these points are equal for the general (independent) service time case. As a consequence, the insensitivity of the arbitrary time stationary queue length distribution is shared by the various imbedded stationary queue length distributions.

We also examined the internal and external flow processes in this system. Because the instantaneous Bernoulli feedback can be thought of as modifying the service requirement distribution, we studied the departure process as if it came from an M/GI/1 queue without feedback. Hence for the queue with feedback, the departure process is a Poisson process with parameter λ, irrespective of the service time distribution (i.e. the departure process is insensitive to the service time distribution). The internal flows, however, are not insensitive to the service time distribution; the distribution of a stationary interoutput interval is different for different service time distributions.

Although the mean stationary interoutput time is insensitive to the service time distribution, given its mean (Lemma 5.6), we showed that the distribution of an interval is not insensitive. We discussed a case where, beginning with convexly ordered service requirements, the output intervals are convexly ordered (Section 5.5). This case leaned heavily on one of the systems having constant service times, so that the input order is almost surely preserved in the output.

Our analysis hinged on comparing the times until the first event in a certain non-stationary process for different service requirement distributions. To obtain the interoutput distribution explicitly proved difficult in the processor sharing queue. We discussed a similar analysis for the infinite server queue in Chapter 6. We again found that the imbedded and arbitrary time stationary queue lengths were equal in distribution and hence both were insensitive. For this queue we could compute the distribution of the time until the first event for special cases, which we did in Sections 6.5 and 6.6. These examples again indicate that the convexity of the service requirement distributions implies the convexity of the interoutput intervals, and hence these processes are not insensitive.

7.2 Discussion

Our analysis of the interoutput times in the processor sharing queue pointed to the importance of two different features of a given service requirement distribution. The first of these is the smallest remaining service time when the system is examined at an arbitrary point in time. In equilibrium the number in the system is insensitive to the distribution of service times (with a given mean); the

distribution of the smallest remaining service time is not. In fact, Lemma 5.7 states that if service requirements are convexly ordered, the smallest remaining service time is ordered in distribution. The second important feature in determining when the next event will occur is how likely an arrival after time zero is to become the unit that will depart next. That is, how likely is it for an entering unit to have a service requirement smaller than the smallest one currently in service? In the case of constant service times this can almost surely never happen. One would think that with equal means, the service requirement distribution with the larger variance would have a greater probability of both long and short service requirements. Between times of arrivals, the graph of the smallest remaining service requirement of units in the system over time decreases linearly (Figure 5.1). If at an arrival time, the arriving unit has a service requirement smaller than the smallest one currently in service, the graph will make a jump downward. Therefore, if incoming units have a greater probability of short service requirements, the graph seems more likely to decrease by jumps. It seems reasonable, then, to conjecture that a service requirement distribution with a larger variance is more likely to have downward jumps than one with a smaller variance.

Now we know that in a stationary system, the initial smallest remaining service time and the number of downward jumps made in the smallest remaining service requirement over time just balance each other to always produce an exponential time until the first departure occurs. But for the output process, the arrivals are only occuring at

rate λ, not rate λ/q. When we delete each point in the λ-stream with probability $1-q$, then, we are (possibly) removing points at which a downward jump may occur. It seems likely that a system that makes more downward jumps would be delayed more by deleting (possible) jump points than would a system that has fewer downward jumps to begin with. In other words, if a system begins with a stochastically large initial smallest remaining service time and relies on downward jumps to produce a Poisson departure process, deleting some of those jumps should have a different effect on the first departure time than in a system that has a smaller smallest remaining service time initially, and makes fewer jumps downward before the smallest remaining service time hits zero. Thus intuitively it would appear that the variability of a service requirement is important in determining when the first output will occur.

The same principle applies in the infinite server queue. In both these queues, an entering unit may finish a service requirement before any or all of the units present at its arrival. The possibility that a unit that arrives later than another unit can leave before the other unit (overtaking) seems to be the common thread between these queues. Although the rate of service changes when a unit arrives or departs in the processor sharing queue (unlike the infinite server queue), the stationary state densities in both cases have the property that accumulated service requirements on units in the queue in equilibrium are independent and identically distributed random variables, unlike in the classical first come, first served discipline. The overtaking possibility appears to be the important feature that not only produces

a Poisson departure process in both these systems, but affects the output process as well.

7.3 Open Problems

There are many directions in which to extend the analyses of this paper. First and foremost would be to come up with a general ordering principle between the service requirements and interoutput times for any service requirement distributions. The discussion above on the effects of incoming units on the path of the minimum remaining service time would appear to be one method to attack this problem. We believe that in comparing two service time distributions, both of which can produce downward jumps in the path, the one more likely to produce jumps would have convexly larger interoutput intervals. This would be a useful result in characterizing the distribution of an interoutput interval. It appears to be a very difficult problem to compute these distributions explicitly.

If the ordering conjecture holds for any M/GI/1/PS-IBF queue, it would be important to try to extend the principle to flow processes in networks of processor sharing queues. The work done on queueing networks in the past indicates that it is often precisely the feedback loops that produce complicated flow processes. One might begin with a simple two node network and examine the intervals between outputs to see if the ordering principle holds there too. Perhaps a simple delayed feedback network would provide a useful starting point.

Throughout this work we have dealt only with single intervals in the internal flow processes. Since we know that in general these interval processes are not renewal processes, there is much work to be

done on characterizing the structure of the dependencies in these processes. One might ask whether or not some functional of the flow (eg., the correlation between consecutive intervals) has the insensitivity property, or is ordered in some way with service requirements.

Finally, we wish to say a word about simulation. We can show that interoutput distributions in processor sharing queues with the same mean service requirement are different, but since we cannot easily compute these distributions, we cannot answer the question, "how different?" It would be useful to develop some statistically sound simulation studies that would help us to determine, for example, the mean time until the first output in comparable systems. We need to know what magnitude of difference we are dealing with. If the differences are small, we may be able to develop some good bounds using systems that we can work with analytically.

BIBLIOGRAPHY

Baskett, F. and F. Palacios (1972), "Processor sharing in a central server queueing model of multiprogramming with applications", Proceedings of the Sixth Annual Princeton Conference on Information Sciences and Systems, Princeton University, Princeton, New Jersey, 598-603.

Beutler, F. J. and B. Melamed (1978), "Decomposition and customer streams of feedback queueing networks in equilibrium", Oper. Res. 26, 1059-1072.

Brumelle, S. L. (1978), "A generalization of Erlang's loss system to state dependent arrival and service rates", Math. Oper. Res. 3, 10-16.

Burman, D. Y. (1980), "Insensitivity in queueing systems", Adv. in Appl. Probab. 13, 846-859.

Burke, P. J. (1976), "Proof of a conjecture on the interarrival-time distribution in an M/M/1 queue with feedback", IEEE Trans. Comm., May, 575-576.

Chandramohan, J. and R. L. Disney (1982), Private communication.

Çinlar, E. (1975), Introduction to Stochastic Processes. Prentice Hall, New Jersey.

Cohen, J. W. (1979), "The multiple phase service network with generalized processor sharing", Acta Inform. 12, 245-284.

D'Avignon, G. R. (1974), "Single server queueing systems with feedback", Ph.D. Dissertation, The University of Michigan, Ann Arbor, Michigan.

D'Avignon, G. R. and R. L. Disney (1976), "Single-server queues with state-dependent feedback", INFOR-Canad. J. Oper. Res. Inform. Process., (14), 71-85.

D'Avignon, G. R. and R. L. Disney (1977), "Queues with instantaneous feedback", Management Sci. 24, 168-180.

Disney, R. L. and P. C. Kiessler (1987), Traffic Processes in Queueing Networks: A Markov Renewal Approach. Johns Hopkins University Press, Maryland (in press).

Disney, R. L. and D. König (1985), "Queueing networks: a survey of their random processes", SIAM Rev. 27, 335-403.

Disney, R. L., D. König and V. Schmidt (1985), "Stationary queue-length and waiting-time distributions in single-server feedback queues", Adv. in Appl. Prob. 16, 437-446.

Disney, R. L., D. C. McNickle and B. Simon (1980), "The M/G/1 queue with instantaneous Bernoulli feedback", Naval Res. Logist. Quart. 27, 635-644.

Foley, R. D. (1982), "The non-homogeneous M/G/∞ queue", Opsearch 19, 40-48.

Franken, P., D. König, U. Arndt and V. Schmidt (1981), Queues and Point Processes. Akademie-Verlag, Berlin.

Gross, D. and C. M. Harris (1974), Fundamentals of Queueing Theory. Wiley, New York.

Hunter, J. J. (1983), "Filtering of Markov renewal queues, I: feedback queues", Adv. in Appl. Prob. 15, 349-375.

Hunter, J. J. (1984), "Filtering of Markov renewal queues, III: semi-Markov processes embedded in feedback queues", Adv. in Appl. Probab. 16, 422-436.

Hunter, J. J. (1985), "Filtering of Markov renewal queues, VI: flow proceses in feedback queues", Adv. in Appl. Probab., 17, 386-407.

Jackson, J. R. (1957), "Networks of waiting lines", Oper. Res. 5, 518-521.

Jackson, J. R. (1963), "Jobshop-like queueing systems", Management Sci. 10, 131-142.

Jansen, U. and D. König (1980), "Insensitivity and steady state probabilities in product form for queueing networks", Elektron. Informationsverarb. Kybernet. 16, 385-397.

Kelly, F. P. (1976), "Networks of queues", Adv. in Appl. Probab. 8, 416-432.

Kelly, F. P. (1979) Reversibility and Stochastic Networks. Wiley, New York.

Kiessler, P. C. (1983), "Reversibility and flows in queueing networks", Ph.D. Dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Kitayev, M. Yu. and S. F. Yashkov (1979), "Analysis of a single-channel queueing system with the discipline of uniform sharing of a device", Izv. Akad. Nauk SSSR Tehn. Kibernet. 17(6), 42-49. (English translation in Engrg. Cybernetics).

Kleinrock, L. (1976), Queueing Systems, Vol. 2, Wiley Interscience, New York.

König, D., K. Matthes and K. Nawrotzki (1967), "Generalization of the Erlang and Engset formulae (a method in queueing theory)", Akademie-Verlag, Berlin (In German).

Melamed, B. (1979), "Characterizations of Poisson traffic streams in Jackson queueing networks", Adv. in Appl. Probab. 11, 422-438.

Mirasol, N. M. (1963), "The output of an M/G/∞ queueing system is Poisson", Oper. Res. 11, 282-284.

Montazer-Haghighi, A. (1977), "Many server queueing systems with feedback", Proceedings of the Eighth National Mathematics Conference, Arya-Mehr University of Technology, Tehran, Iran, 228-249.

Muntz, R. R. (1972), "Poisson departure processes and queueing networks", Proceedings of the Seventh Annual Princeton Conference on Information Sciences and Systems, Princeton University, Princeton, New Jersey, 435-440.

O'Brien, G. L. (1975), "The comparison method for stochastic processes", Ann. Probab. 3, 80-88.

Ross, S. M. (1983), Stochastic Processes. Wiley, New York.

Schassberger, R. (1977), "Insensitivity of steady-state distributions of generalized semi-Markov processes. Part I", Ann. Probab. 5, 87-99.

Schassberger, R. (1978a), "Insensitivity of steady-state distributions of generalized semi-Markov processes. Part II", Ann. Probab. 6, 85-93.

Schassberger, R. (1978b), "Insensitivity of steady-state distributions of generalized semi-Markov processes with speeds", Adv. in Appl. Probab. 10, 836-851.

Schassberger, R. (1978c), "The insensitivity of stationary probabilities in networks of queues", Adv. in Appl. Probab. 10, 906-912.

Simon, B. and R. L. Disney (1984), "Markov renewal processes and renewal processes: some conditions for equivalence", New Zealand Oper. Res. 12, 19-29.

Sonderman, D. (1980), "Comparing semi-Markov processes", Math. Oper. Res. 5, 110-119.

Stoyan, D. (1983), Comparison Methods for Queues and Other Stochastic Models. Wiley, New York.

Takács, L. (1960), Stochastic Processes. Methuen, London.

Takács, L. (1963), "A single server queue with feedback", Bell System Tech. J. 42, 505-519.

Walrand, J. (1982), "Poisson flows in single class open networks of quasireversible queues", Stochastic Process. Appl. 13, 293-303.

Walrand, J. and P. Varaiya (1980), "Interconnections of Markov chains and quasireversible queueing networks", Stochastic Process. Appl. 10, 209-219.

Walrand, J. and P. Varaiya (1981), "Flows in queueing networks: a martingale approach", Math. Oper. Res. 6, 387-404.

Whitt, W. (1980), "The effect of variability in the GI/G/s queue", J. Appl. Probab. 17, 1062-1071.

Wolff, R. (1982), "Poisson arrivals see time averages," Oper. Res. 30, 223-231.

Wyszewianski, R. J. and R. L. Disney (1974), "Feedback queues in the modelling of computer systems: a survey", Technical Report 74-1, Dept. of Ind. and Oper. Eng., University of Mich., Ann Arbor.

Yashkov, S. F. (1980), "Properties of invariance of probabilistic models of adaptive scheduling in shared use systems", Avtomat. i Vychisl. Tehnika 14(6), 56-62. (English translation in Automat. Control Comput. Sci.).

Yashkov, S. F. (1981a), "Some results of analyzing a probabilistic model of remote processing systems", Avtomat. i Vychisl. Tehnika 15(4), 3-11. (English translation in Automat. Control Comput. Sci.).

Yashkov, S. F. (1981b), "On the ergodicity of systems with variable speed of servicing", Izv. Akad. Nauk SSSR Tehn. Kibernet. 19(3), 74-80. (English translation in Engrg. Cybernetics).

The vita has been removed from the scanned document