TRAFFIC PROCESSES AND SOJOURN TIMES IN FINITE MARKOVIAN QUEUES

by

JOHN A. BARNES

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:

RALPH L. DISNEY (CO-CHAIRMAN)

JEFFREY D. TEW (CO-CHAIRMAN)

JOEL A. NACHLAS

PETER C. KIESSLER

MARTIN V. DAY

ROLAND B. MINTON

December 1988
Blacksburg, Virginia

TRAFFIC PROCESSES AND SOJOURN TIMES IN FINITE MARKOVIAN QUEUES

by

JOHN A. BARNES

Committee Co-Chairmen: **Ralph L. Disney**

Jeffrey D. Tew

(ABSTRACT)

This paper gives results on various traffic processes and on the sojourn time distribution for a class of models which operate as Markov processes on finite state spaces. The arrival and the service time processes are assumed to be independent renewal processes with interval distributions of phase-type. The queue capacity is finite. A general class of queue disciplines are considered. The primary models studied are from the $M/E_k/\phi/L$ class. The input, output, departure and overflow processes are analyzed. Furthermore, the sojourn time distribution is determined. Markov renewal theory provides the main analytical tools. It is shown that this work unifies many previously known results and offers some new results. Various extensions, including a balking model, are studied.

ACKNOWLEDGEMENTS

I would first and foremost like to express my appreciation to Professor Ralph Disney who has been a remarkable mentor. He has been an inspiration to me in so many ways, especially by his professionalism and by his dedication to students and co-workers that transcends any academic affiliations.

I would like give thanks to all the members of my committee who have been so helpful and accommodating— to Professors Jeffrey Tew and Joel Nachlas for joining my committee when I really needed them, to Professor Martin Day for helping me stay in touch with my mathematical roots, to Professor Roland Minton for remaining on my committee even when he had excuses for leaving and for his careful reading that found many typing errors that had eluded me, and to Professor Peter Kiessler for his helpful comments and for his willingness to make those trips from Richmond.

I would like to thank Professors Yashaswini Mittal and Robert Foley who as teachers and as former members of my committee were instrumental in my development as a graduate student.

I feel priviledged to have been able to share company with Professors Jeffrey Hunter, Donald McNickle and Masakiyo Miyazawa when they visited Blacksburg. Each of these men have taken an active interest in my research. Their comments and suggestions have been invaluable.

I would like to thank Professor Marcos Magalhaes who as a fellow graduate student worked with me so closely during the research phase. In addition, I wish to thank him along with Professors Eddy Patuwo, Georgia-Ann Klutke and Martin Wortman for sharing with me the agonies and the rewards of being graduate

students.

I offer special appreciation to my wife who not only put up with me during the whole ordeal but learned TEX and typed the manuscript with its many revisions.

I owe more than a debt of apology to my children, and , who have suffered so much neglect from their father.

Finally, I want to give a special thanks to my parents who instilled in me the importance of a quality education.

TABLE OF CONTENTS

APPLICATIONS

EXTENSIONS AND GENERALIZATIONS

Chapter 1

Background and Literature Review

1.1 Introduction

The objective of this dissertation is to present a study of some of the important features of a general class of queueing models which operate in continuous time as Markov processes on finite state spaces. Various stochastic processes relating to the models will be studied. In particular, the stationary Markov state process will be studied along with various discrete time processes that are useful in studying various customer traffic flows within the system. The latter is studied using Markov renewal theory. In addition to the aforementioned stochastic processes, the distribution of the sojourn and waiting times will be developed.

The main class of models that are studied is the $M/E_k/\phi/L$ class operating under a queue discipline taken from a general class. These models include as special cases many of the models that have been studied in isolation. Thus, many of these results are unified under one theory. Moreover, new insights into the nature of these special cases are presented. In the final chapter, various extensions to the $M/E_k/\phi/L$ class are considered.

In Chapter 1, the relevant terminology of queueing theory is defined. Furthermore, the models are defined in detail and the most relevant literature is reviewed. In Chapter 2, the general theory is developed for the $M/E_k/\phi/L$ class of models. Results are presented for some of the important stochastic processes and for the sojourn time distributions. Chapter 3 relates the theory of Chapter 2 to special cases. Some of the results appear to be new. Finally, Chapter 4 contains extensions, open problems, and concluding remarks.

1.2 General Terminology and Definition of Models

A queueing network consists of a collection of nodes that are interconnected in some way. Each node consists of a server which supplies random service times to customers who arrive to the node from outside the network or from other parts of the network. The server may in fact be several servers working in parallel. If a customer arrives to a node and is unable to receive the immediate attention of the server, then the customer may be allowed to join a queue. The maximum number of customers allowed at the server, waiting and being served, is called the queue capacity. The rule which governs the way in which queueing customers receive service is called the queue discipline. For example, if customers are served in the order that they arrive, then the discipline is called first-come-first-served (FCFS). If each customer receives an equal share of the server's attention then the discipline is called processor sharing (PS - a termed coined because of its computer applications.)

After a customer completes its service requirement at a server then the customer leaves the server and is said to output from the server. At this point a decision must be made. The customer either leaves the network or is routed to another server. It may even feedback to the same server it just left. When a customer arrives to a server it may not even stay for service. Another decision may be necessary. For example, if the arriving customer finds the queue at capacity, it may be forced to bypass the server and depart immediately without service. In this case, the customer is said to overflow. As another example, an arriving customer finding n customers in the queue may decide with probability $b(n)$ to bypass the server and depart immediately. This type of behavior is called balking. A server, along with its queue and rules of operation, is called a node of the queueing network.

It is evident that the terminology of queueing theory derives from applications.

For example, when reading the above descriptions of terms, one might get an image of students running around campus on the first day of school, registering for classes, buying textbooks, eating lunch, setting up bank accounts, etc., standing in queues at each step along the way. Of course, in applications, the customers may not be people at all but such things as telephone calls, computer jobs, parts moving on a conveyor or traffic in transportation or communication systems. In addition to the many applications, queueing systems can be studied at a more abstract level. Queueing systems can be used as convenient devices for studying the interaction of various stochastic processes. For example, the flow of customers arriving to a node can be described by an arrival process. The customers flowing into the server form the input process. Those that bypass the server form the overflow process or balking process. Those that leave the server after being served form the output process and those that leave the node form the departure process. The size of the queue when viewed at arbitrary times or as seen at certain discrete times are examples of queue length processes. Most of these processes will be defined in more detail in Section 1.3 and Section 1.4. The balking process will be discussed in Chapter 4. It should be noted that the queue length will be defined to include the customers which are receiving service as well as those which are waiting. The nature of these stochastic processes will be made more precise in later sections. See Figure 1.1 for a diagram of a generic node of a queueing network.

Kendall [1953] originated and many later queueing theorists have embellished a notational shorthand for describing queueing models. The notation takes the form

$$A/S/n/L - QD$$

where

(i) A is replaced by some symbol that refers to the arrival process to the system.

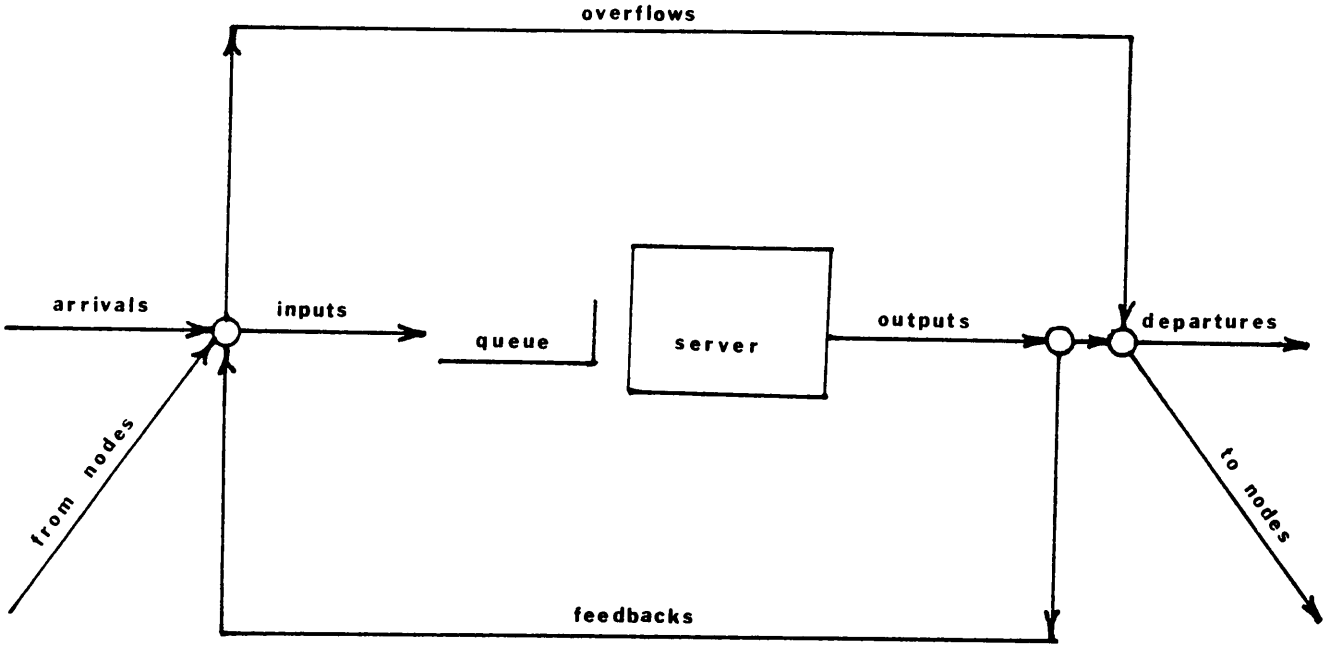

Figure 1.1 Node of a network

For example, if A is replaced by M (for Markovian), then the arrivals occur according to a Poisson process. If A is replaced by GI (for general independent), then the arrivals occur according to a renewal process. Note that M is a special case of GI. If G is used, then the arrival process is general. The general case is usually with no independence assumption, but stationarity of the process is usually retained.

(ii) S is replaced by some symbol that defines the service time process. The same symbols, as in (i), may be used in place of the S descriptions. For example, if S is replaced by M, then the service times are i.i.d., independent and identically distributed, with the exponential distribution.

(iii) n refers to the number of parallel servers for the queue. However, this descriptor will be generalized in this paper.

(iv) L refers to the capacity of the system. However, this sometimes refers to the waiting capacity only. This fourth descriptor is usually omitted when the queue capacity is infinite.

(v) QD is sometimes replaced with some symbol which gives the queue discipline. This descriptor is not always used, especially if the queue discipline is clear from context.

Using the Kendall style notation, the family of models described below are of the M/GI/ϕ/L type. The following assumptions define this family of models which is the primary subject of this dissertation; however, more general models will be discussed in the final chapter.

1. The model consists of a single node.
2. The arrival process is Poisson. That is, if $\{T_n^a\}$ are the arrival epochs then the interarrival intervals, $\{T_{n+1}^a - T_n^a\}$ are independent and identically distributed (a renewal process) with the exponential distribution (parameter, λ). Not only

does this assumption make the model more tractable but makes physical sense (at least approximately) in many real-world applications.

3. The queue capacity may be finite or not. If $L < \infty$ when a customer arrives to find a queue length of L then that customer overflows.

4. Each arriving customer requires an amount of service independent of and identically distributed with every other customer, past and future; i.e., the sequence of service requirements forms a renewal process and so the GI descriptor replaces S. For now we will assume that the service requirements are Erlang-k distributed. That is, $f(t)$, the service time density, is given by

$$f(t) = \frac{k\mu(k\mu t)^{k-1}e^{-k\mu t}}{(k-1)!}$$

In this case, the S descriptor is replaced by E_k.

5. The arrival process and the sequence of service requirements are independent of each other.

6. When there are N customers at the node, the server supplies service at a rate, $\phi(N)$. For example, if we wish to model a node with m actual servers then we define $\phi(N) = N$ if $N \leq m$ and $\phi(N) = m$ for $N > m$. Of course, the ϕ could be used to model a single actual server that is capable of adjusting the speed at which it supplies its service. Note that the function ϕ generalizes the concept embodied in the third descriptor in the Kendall notation.

7. Using the method of Kelly [1979], we will allow a large range of queue disciplines. Suppose that customers in the queue occupy numbered positions $1, 2, \ldots, N \leq L$. When a customer arrives to a queue of size $N < L$, it enters position l with probability $\delta(l, N+1)$. Customers previously occupying positions $l, l+1, \ldots, N$ move to positions $l+1, l+2, \ldots, N+1$, respectively. If $N = L$, then the arriving customer overflows.

8. We further suppose that the attention of the server is possibly shared by the customers. When there are N customers at the node, then $\gamma(l, N)$ is the proportion of the total service effort supplied to the customer in position l.
9. When a customer departs from position l, then the customers previously in positions $l+1, l+2, \ldots, N$ move to positions $l, l+1, \ldots, N-1$, respectively. For example, to model the FCFS discipline, let

$$\delta(N+1, N+1) = 1 \text{ for } N = 0, 1, \ldots, L-1$$

$$\delta(l, N+1) = 0 \text{ for } l < N+1$$

and

$$\gamma(1, N) = 1 \text{ for } N = 1, \ldots, L$$

$$\gamma(l, N) = 0 \text{ for } l > 1.$$

To model the processor sharing discipline we define,

$$\gamma(l, N) = \frac{1}{N} \qquad 0 < N \leq L; \qquad 1 \leq l \leq N.$$

The main aspects of the model to be investigated are the stationary state and traffic processes and the sojourn times. Because of the generality of the model and the number of parameters that can be evaluated, there are many papers in the literature that are at least tangentially related to our research. Indeed, one of the main contributions of this dissertation is to unify and extend many of these results. Some of the results in the existing literature may be more general in some aspects of the model but quite specific in other aspects. For example, the researcher may allow a service time distribution more general than the phase type or an arrival process more general than the Poisson process, but there may only be a single server or the queue discipline may be just FCFS.

Because of the enormity of the relevant literature only the most directly relevant work is reviewed here. When appropriate, more extensive review and survey papers will be noted.

1.3 The State Process

The state process is a stochastic process that gives the state of the queueing system as it evolves in time. In this rather vague description we must clarify what is meant by state and time.

The state of the system at a given time is taken to be a random vector which contains enough information so that the process is Markov. The number of customers in the queue can always be ascertained from this vector. Additionally, the vector may contain components that give information about the expended or remaining service requirements for each of the customers. For our model, the state space is finite or at least countable, which is one of the main reasons for the E_k server. This point will be explained at the beginning of Chapter 3.

The time parameter of the state process may be continuous or discrete. If the process gives the state of the system at arbitrary times then the process is referred to as a continuous time process. We may be more interested in keeping track of the state of the system at certain discrete times such as just before an arrival or just after a departure. This will be the case, for example, when certain traffic processes are studied. Such processes are referred to as embedded processes.

For both the continuous and embedded versions of the state process, we will always assume that the process is stationary or has a stationary distribution. This means that the probability law of the process, as given by the finite dimensional joint distributions, is invariant under translations of the time parameter. The

restrictions on the models under study will always be strict enough that such stationary distributions exist. Since we are dealing with stationary Markov state processes (usually called Markov chains in the embedded case) then the probability law of such a process is completely determined from the stationary distribution and the Markov generator (or the transition matrix in the embedded case). The determination of the stationary distribution (often referred to as the steady state or equilibrium distribution) for queueing models has been a common goal of many researchers in the field. Indeed, of all the possible measures of effectiveness of a queueing system, it seems that the stationary distribution is the most sought after.

On the surface the problem seems straightforward. One must find a probability vector, π which solves a system of linear equations of the form $\pi A=0$ or $\pi P=\pi$, where A is a generator of a Markov process and P is the transition matrix for a Markov chain. However, the problem of computing π from these equations has proven difficult for many models with general parameter settings. Researchers have devised clever and sophisticated tricks to find exact solutions. When exact solutions seemed intractable, researchers have used approximations and simulations in the quest. For some of the relatively simple models, the stationary distributions are found in most queueing textbooks such as Kleinrock [1975] or Gross and Harris [1985]. The birth-death queueing models are notable examples. This type model will be important to us since our model fits into this category when the service times are exponentially distributed.

In the simple birth-death case, the state of the system only consists of the queue length since the expended or remaining service times of the customers are irrelevant because of the forgetfulness property of the exponential distribution. For the finite birth-death queueing system, the stationary distribution for the continuous time

state (queue length) process is given by

$$\pi = B(1, \frac{\alpha_0}{\beta_1}, \frac{\alpha_0\alpha_1}{\beta_1\beta_2}, \ldots, \prod_{i=0}^{L} \frac{\alpha_i}{\beta_{i+1}}) \tag{1.1}$$

where the α_i and β_i are respectively the birth and death rates of the queue. B is a normalizing constant and L is the capacity of the system.

The generator A for the birth-death process has a simple tridiagonal form which makes (1.1) easy to obtain. If the service time distribution of our model is allowed a more general phase type distribution such as the Erlang class, then we lose this simple structure. This is because the holding times in the queue length state are no longer exponential. However, the matrix still has a *block* tridiagonal structure. The steady state distribution in the general setting seems to be unknown. However, certain representative examples will be presented in later chapters.

One important property of stationary distributions concerns the relationship between the continuous time distribution and various embedded distributions. For example, it is well known that for the M/M/1 queue, the stationary distribution embedded just before arrivals and the stationary distribution embedded just after departures are both the same as the continuous time distribution. This is an example of Wolff's [1980] PASTA result. (Poisson arrivals see time averages.) Disney and Kiessler [1987] presented a general formula that relates the continuous time steady state distribution of a Markov queueing system to the discrete time distribution embedded just after a transition from an arbitrary set of transitions (called a traffic set). This formula will be presented and used in later chapters.

Some Markov state processes have the property of reversibility. A process is said to be reversible if the finite dimensional distributions are unchanged when the time parameter is replaced by its negative. Reversible processes are thus stochastically the same if time is run backwards. To help visualize this concept, note that

when time is run backwards the arrivals to a queue become departures and vice versa. It is well known that the stationary queue length process of the M/M/1 queue is reversible, In fact, it is easy to see that any stationary birth-death process is reversible. One useful way to check for reversibility is by verifying the detailed balance equations:

$$\pi(i)A(i,j) = \pi(j)A(j,i)$$

where π is the stationary distribution and A is the generator matrix for the process.

A related property of Markov state processes is the property of dynamic reversibility. A process is dynamically reversible if the process would be reversible when the states are relabelled by a suitable permutation. A process is dynamically reversible if and only if

$$\pi(i) = \pi(i^+) \quad \text{for all } i$$

and

$$\pi(i)A(i,j) = \pi(j^+)A(j^+,i^+)$$

where i^+ is the state i when relabelled.

A stationary Markov process $\{X(t)\}$ may not be reversible, yet, we may still consider the effect of running the process backward in time, $\{X(-t)\}$. This reverse process is another stationary Markov process. A useful way to test two processes $\{X(t)\}$ and $\{X'(t)\}$ to see if they are reverses of each other is to check the following balance conditions:

$$\pi(i)A'(i,j) = \pi(j)A(j,i)$$

where A and A' are generators for $\{X(t)\}$ and $\{X'(t)\}$ respectively. π is the steady state vector for *both* processes. Note that a reversible process is its own reverse. Kelly [1979] is the definitive treatment of the concepts of reversibility and reversing just described.

Of all the stochastic processes associated with a queueing model, the state processes are by far the most studied. In this dissertation, the state process, per se, plays a somewhat ancillary role. These processes will be studied mainly as one component of traffic processes and as an aid in computing sojourn times.

1.4 Traffic Processes

In Markovian queueing models, the flow of various types of customer traffic can be studied by means of Markov renewal processes. (See the Appendix for a brief introduction/review of the definition of Markov renewal processes.) Consider $\{(X_n^r, T_n^r)\}$ and $\{(X_n^{r-}, T_n^r)\}$ where X_n^r and X_n^{r-} are the state processes embedded, respectively, just after and just before state transitions of type r. T_n^r is the time epoch of the *nth* transition of type r. We assume here that the sample paths of the processes are right continuous with left hand limits. Disney and Kiessler [1987] is the definitive work on the topic. The methodology developed in that reference will be used extensively in this dissertation.

The traffic processes of greatest interest in our model are the

(i) arrival processes	$\{(X_n^a, T_n^a)\}$	and	$\{(X_n^{a-}, T_n^a)\}$
(ii) input processes	$\{(X_n^i, T_n^i)\}$	and	$\{(X_n^{i-}, T_n^i)\}$
(iii) departure processes	$\{(X_n^d, T_n^d)\}$	and	$\{(X_n^{d-}, T_n^i)\}$
(iv) output processes	$\{(X_n^o, T_n^o)\}$	and	$\{(X_n^{o-}, T_n^o)\}$
(v) overflow processes	$\{(X_n^v, T_n^v)\}$	and	$\{(X_n^{v-}, T_n^v)\}$

1.5 Departure and Output Processes

One of the most heavily studied traffic processes in queueing systems has been the departure process. Exactly what is meant by the "departure process" varies in

the literature. Often the departure process refers to the point process generated by the departure epochs. This can be studied via the counting process $\{N^d(t)\}$ where $N^d(t)$ is the number of departures which have occurred by time t. Another approach is to study the departure interval process $\{T^d_{n+1} - T^d_n\}$ or just the process $\{T^d_n\}$. The $\{T^d_n\}$ process is just the second component of the bivariate process $\{(X^d_n, T^d_n)\}$. By studying the X and T components jointly, we can gain a richer understanding than either component provides in isolation. Of course, once the joint situation is known then the marginal processes can be determined.

Most of the results on departure processes in the literature give situations where the equilibrium departure process (the point process) is Poisson. One reason this is of interest is that the output of one queue may become the input into a second queue, thus simplifying the analysis of the second queue. The earliest such result is probably that of Burke [1956]. He showed that the departure process from the M/M/s queue (FCFS was assumed) is Poisson with the same parameter as the arrival process. Mirasol [1963] showed that the departure process from an M/GI/∞ queue is likewise Poisson. Of course, the queueing discipline for Mirasol's model is irrelevant since no queueing ever takes place. Mirasol's result actually follows from Doob ([1953], p. 405). In Doob's problem the points of a stationary Poisson process are all shifted to the right by i.i.d. random variables independent of the original process. After such an operation the process is still Poisson with the same parameter. Doob probably had no queueing application in mind when he showed this result. Viewing the departure process as a random translation of the arrival points is an interesting idea; however, it is still an open problem as to how far Doob's result can be generalized.

Disney, Farrell and deMorais [1973] showed that among the M/GI/1/L (FCFS) models, renewal departures are quite special, occurring only in the following cases:

1) Service times are all 0 a.s.

2) L=0 (no queueing capacity)

3) L=1 and the service times are constant (GI=D)

4) L= ∞ and GI=M

Moreover, Poisson departures occur only in case 1, which is rather trivial and uninteresting, and in case 4, which is Burke's result.

If queue disciplines other than FCFS are considered then the M/GI/s/L$\leq$ ∞ model may have a Poisson departure process in certain special cases. Muntz [1972] and Kelly [1979] showed that a queueing node in equilibrium with Poisson arrivals and GI service requirements has a Poisson departure process if the system is quasi-reversible. A node is said to be quasi-reversible if for each time t the evolution of the arrival process after t is independent of the state at t and if the history of the departure process before t is independent of the state at t. Kelly showed further that this condition holds when the queueing discipline is symmetric (i.e. when $\delta(l,n) = \gamma(l,n)$ for all the l,n in the domain of these functions). For example, the queueing discipline is symmetric in the case of last-come-first-served with preemption and resume (LCFS-PR) or the case of processor sharing (PS).

Except for the Disney, et. al., paper previously cited, most of the known conditions for Poisson departures or renewal departures are sufficient conditions. However, Berman and Westcott [1983] gave a necessary condition for the departures from a GI/G/s queue to be a renewal process. (FCFS appears to be assumed.) The condition is that the single interval distribution of the departure process must be the same as the interarrival time distribution. Very few results are available for departure processes which are not even renewal. Disney and deMorais [1973] showed that the outputs from the M/E_k/1/L-FCFS model are Markov renewal and they studied the covariance properties of the process. As a consequence of Disney

and Kiessler [1987], the departure and output processes from the M/GI/ϕ/L type models will always be Markov renewal. How this result impinges on our general model will be explored in later chapters.

For other results concerning departure and output process, see Daley [1975] and Disney and König [1985].

1.6 Arrival and Input Processes

In our model the arrival process is assumed to be Poisson. However, the input process becomes more complex when we allow overflows or balkings. The study of such input processes has received little attention in the literature; although, some researchers have investigated the relationships between the input and output processes. Natvig [1975,1977] investigated this relationship for the birth-death model. In the first paper it is determined that when $s + L > 1$ the input and output processes are identical if and only if $s + L = \infty$, in which case they are both Poisson. In the second paper it is shown that the input and output processes are reverse processes in the sense that

$$P(D_1^i \leq x_1, D_2^i \leq x_2, \ldots, \leq D_n^i \leq x_n) = P(D_1^o \leq x_n, D_2^o \leq x_{n-1}, \ldots, D_n^o \leq x_1)$$

for

$$n = 1, 2, \ldots \quad and \quad 0 \leq x_j \leq \infty$$

where

$$D_j^i = T_j^i - T_{j-i}^i \quad and \quad D_j^o = T_j^o - T_{j-1}^o$$

are respectively, the jth interinput and interoutput intervals. Disney and Kiessler [1987] gave much more general results. They showed that if the state process is a Markov process which is reversible (or dynamically reversible), then the input

and output Markov renewal processes are reverses (or dynamical reverses) of each other. From this it follows that the input and output interval processes are reverses of each other. Since a birth-death process is reversible, then Natvig's results follow as a special case. It will be shown in the next chapter that the results of Disney and Kiessler apply to our general model when certain conditions are met. In particular, the input and output processes are reverses of each other if $E_k = M$ (i.e. k=1) and are dynamic reverses of each other if $\delta \equiv \gamma$.

1.7 Overflow Processes

The overflow process was probably the first traffic process to be studied. Applications to telephone traffic engineering provided the impetus for this research. Palm [1943] considered the following telephone model which in modern notation is known as the GI/M/L/L queueing system. Calls arrive according to a renewal process to an ordered group of L lines. Each call enters the first available line and lasts for an exponentially distributed time interval. If all lines are busy then the call overflows the group. A call is offered to line r only if the first $r-1$ lines are busy, in which case it overflows from this group of $r-1$ lines. It is shown that the time intervals between consecutive overflows from line r form a renewal process with an interval distribution $\phi_r(t)$ which satisfies the following integral recurrence equation:

$$\phi_r(t) = \int_0^t e^{-\mu\lambda}\, d\phi_{r-1}(u) + \int_0^t (1 - e^{-\mu\lambda})\phi_r(t-u)\, d\phi_{r-1}(u)$$

Using Laplace-Stieltjes transforms on this equation yields the following recurrence:

$$\widehat{\phi}_r(s) = \frac{\widehat{\phi}_{r-1}(s+\mu)}{1 - \widehat{\phi}_{r-1}(s) + \widehat{\phi}_{r-1}(s+\mu)}$$

Khinchine [1960] and Riordan [1962] contain good discussions of Palm's problem. Disney and Çinlar [1967] showed that the overflow process is still renewal when

queueing space is added to Palm's model, i.e. the GI/M/N/L, L>N. If the renewal arrival process is made Poisson, then the above models become birth-death models, since we then have the M/M/N/L queue. In this case, the time between overflows from the first r states is simply the first passage time from state r to state $r+1$ where if L is the capacity of the system, we can introduce a dummy state $\Delta \equiv L+1$. It has been shown by several authors (e.g. Karlin and MacGregor [1959], Van Doorn [1984], Branford [1986], and Keilson [1979]) that in this birth-death case,

$$\widehat{\phi}_r(s) = \frac{p_{r-1}(s)}{p_r(s)}$$

where $p_k(s)$ is a polynomial of degree k. This rational transform decomposes so as to imply that the interrenewal distribution is a mixture of r exponential distributions.

We lose the renewal structure of the overflow process when the service time distribution is non-exponential. In later chapters, it is shown that the overflow process from the $M/E_k/\phi/L$ class of models is a Markov renewal process.

Relatively few results are known about overflow processes from non-exponential servers. Halfin [1981] discussed overflows from more general GI/GI/1/L systems, but his analysis extended only to single intervals of the overflow process. Machihara [1987] discussed overflows from the PH/PH/1/L type model (which includes $M/E_k/1/L$ FCFS as a special case). He numerically computed some first and second order moments for the process. In the following chapters the overflow process from the $M/E_k/\phi/L$ model will be determined. Because the service time distributions are in general not exponential and general queue disciplines are allowed, these are new results not previously discussed in the literature.

1.8 Sojourn Times

The sojourn time and the related waiting time are aspects of queueing systems that are heavily studied, going back at least to Lindley [1952]. Perhaps only the queue length has received more attention in the literature. The queue length may be of greatest concern from the point of view of the server since it represents the workload (measured by the number of customers) to be faced. The sojourn time, however, may be of greatest concern to the individual arriving customer. The sojourn time of a customer is influenced by the queue discipline. For example, suppose an arriving customer to a FCFS queue finds N customers already in the system. In this case, the sojourn time will be the total time for these N customers to clear plus the service time of the new customer. However, in a LCFS discipline, the fact that N customers are already present would be of no particular concern to our customer since its sojourn time is just its service time which begins immediately on its arrival.

Even with these very different queue disciplines, the mean waiting time of a customer may be identical in both cases. This is true, for example, in an M/M/1 queue. This result follows from the well known Little's formula

$$E(N) = \lambda E(S)$$

where $E(N)$ is the mean number of customers in the system, $E(S)$ is the mean sojourn time and λ is the mean arrival rate (input rate) of customers. Stidham [1974] gave a proof for this result under very general conditions. The robustness of Little's result indicates that the mean waiting time is of only limited value in understanding customer waiting time. The difference in the waiting of a customer in the FCFS and LCFS queues shows up when the whole waiting time distribution is considered. In fact, one only has to consider the variances to see this difference.

Kingman [1962] showed that for a general G/G/1 queue the waiting time variance is minimized in the FCFS case. Shanthikumar and Sumita [1987] further showed that the maximum variance is attained in the LCFS case. It should be noted that these extrema are taken over the class of work conserving disciplines that do not depend on service times and do not include preemptions. Shanthikumar and Sumita further showed that among the G/M/1 models the results can be extended to all work conserving disciplines that do not depend on service time, even where preemptions are allowed. In this case, it is shown that the waiting time variance is minimized in the FCFS case and maximized in the case of LCFS-PR.

Sojourn times under some other particular queue disciplines are investigated by other researchers. For example, Ott [1984] determined the sojourn time distribution in the M/G/1 queue with processor sharing. The result is in terms of Laplace-Stieltjes transforms and generating functions. Ott extended the results of Coffman, Muntz, and Trotter [1970] who presented the sojourn time distribution for the M/M/1-PS model. Yashkov [1983] and Schassberger [1984] also studied the M/GI/1-PS sojourn time problem. Ramaswani [1984] derived the mean and variance for the sojourn time in the GI/M/1-PS queue. As would be expected, results for the multiserver case are not so extensive. However, several papers are devoted to demonstrating the proposition that the FCFS discipline minimizes waiting in multiserver situations. This proposition is intuitively appealing which may explain the widespread popularity of the little signs which read "Wait here for next available clerk."

Kingman [1970] compared the GI/GI/s queue with the FCFS discipline to the discipline which assigns arriving customers to the servers in a cyclic order, thus creating a separate queueing channel for each server. Kingman asserted, as obvious, the fact that the waiting time mean is larger under the latter discipline. Wolff [1977]

extended Kingman's ideas by showing that the waiting time distribution is smaller in terms of convex ordering than for any other discipline which is independent of service times and arrival times. Stoyan [1983] provided a good discussion and review of the research on bounds and inequalities of the waiting time in the GI/GI/s queue both with FCFS and other allocation disciplines. He included an example that shows that a given customer does not necessarily have a smaller waiting time with a FCFS discipline than with a cyclic allocation of servers. Thus, it is not the case that the FCFS discipline minimizes waiting for every customer in every sample path.

In later chapters the sojourn time distribution will be determined for the class of models studied in this dissertation. This will be accomplished by computing first passage times in a specially constructed process.

Chapter 2

General Results for the $M/E_k/\phi/L$ Models

2.1 Introduction

In this chapter the main theoretical results are presented. The $M/E_k/\phi/L$ family of models as described in Chapter 1 will be the paradigm used to expose the theory. However, the methodology can be adapted to the study of more general models where the service time and interarrival time distributions are more general phase type distributions. These and other extensions will be discussed in Chapter 4. By restricting the model somewhat, the state spaces of the stochastic processes are all finite and thus made more manageable in size and excess notation is avoided. At the same time, the essence of the theory is maintained. Even with a less than completely general model, there is a large number of parameters that must be set in order to define the precise model to be studied. Many of the known results in the literature are unified and extended by the analysis of this chapter. Interesting special cases will be presented in Chapter 3.

The driving force of the theory is the fact that the operation of the model can be described by a Markov process on a finite state space. That is, we can define the state space for the system in such a way that the times between transitions are exponentially distributed with the parameter dependent only on the current state; and, the succession of states forms a Markov chain. The Poisson arrival process easily accommodates this Markovian structure. The complication arises because the service times are not exponentially distributed. However, since the service times are Erlang-k distributed, one can view a service time as a sum of k independent exponentially distributed random variables each with parameter $k\mu$.

In order for a customer to complete service, one can imagine the customer going through k exponential phases of service. Note that these phases do not necessarily correspond to any identifiable physical process of the server. By keeping track of the current service phase of each of the customers in the queue, the Markovian structure can be maintained on a countable state space. This technique is commonly used in the literature and is sometimes referred to as the method of phases. The preceding discussion will be made more mathematically precise in the next section.

Once the Markov process describing the system operation has been developed, then, using a filtering technique, the various traffic processes of interest can be characterized and studied. Moreover, using a variation on the filtering theme, sojourn times distributions can be characterized. This approach to the sojourn time problem appears to be new.

2.2 The Markov State Process

The continuous time state process of the $M/E_k/\phi/L$ model with queue discipline defined by δ and γ (see Section 1.2) must first be described. Let $N \equiv N(t)$ be the queue length at time t. Suppose that customers occupy numbered positions in the queue. Let

$$Y(t) = (s_1(t), s_2(t), \ldots, s_N(t), f(t))$$

where $s_i(t) \in \{1, \ldots, k\}$ is the Erlang phase of service of the customer in position i and where $f(t) \in \{0, 1\}$ is a flip-flop variable which alternates its value whenever an arriving customer instantaneously departs without service (e.g. overflows). Without the $f(t)$ component of the vector, overflow events could not be detected by observing the sample path of the process. Note that $Y(t)$ is a vector with a

random number of components. By convention , let

$$Y(t) = (0, f(t))$$

whenever the system is empty of customers. Let the state space be called E, which is finite.

We assume that the server is always busy when customers are present in the system, i.e.

$$\phi(N) > 0 \text{ if } N > 0.$$

Thus, the system is idle at time t if and only if

$$Y(t) = (0, f(t)).$$

We will make frequent use of the following:

LEMMA 2.1. *Let $X_1, \ldots, X_m$ be independent exponentially distributed random variables with parameters $\sigma_1, \ldots, \sigma_m$, respectively. Then $Z = min(X_1, \ldots, X_m)$ is exponentially distributed with parameter $\sigma = \sum_{i=1}^{N} \sigma_i$. Moreover, $P\{Z = X_i\} = \sigma_i/\sigma$.*

Now we have the following:

THEOREM 2.2. *$\{Y(t)\}$ is a Markov process.*

PROOF: Case 1 Suppose $Y(t) = (0, f(t))$. Then the time until the next transition is the time until an arrival occurs which is exponentially distributed with parameter λ.

Case 2 Suppose $Y(t) = (s_1(t), \ldots, s_N(t), f(t))$. Then the time until the next transition is the time until either an arrival occurs or until a customer completes the current phase of service. This time is the minimum of independent exponentially distributed random variables with parameters

$$\lambda, \gamma(1, N)\phi(N)k\mu, \gamma(2, N)\phi(N)k\mu, \ldots, \gamma(N, N)\phi(N)k\mu \ .$$

So by Lemma 2.1, the time until the next transition is exponentially distributed with parameter

$$\lambda + \phi(N)k\mu \sum_{i=1}^{N} \gamma(i,N) = \lambda + \phi(N)k\mu \ .$$

Furthermore, by the second part of Lemma 2.1, the next state is entered with a probability which depends only on the current state. Therefore, $\{Y(t)\}$ is a Markov process. •

Note that $\{Y(t)\}$ is irreducible and aperiodic. The transitions or jumps in this Markov process occur when there is a change in state. We will assume that the sample paths are right continuous almost surely. The transitions can be categorized into four types: inputs, outputs, phase changes, and overflows. The following definitions will be useful to describe the jumps.

Definition 2.3. Let $\mathbf{y} = (s_1, s_2, \ldots, s_N, f)$ be a state of the system. Let $|\mathbf{y}| = N$ be the number of customers in the system. Then define:

(i) $\varphi_i^+(\mathbf{y})$ is the new state resulting from an arrival to position i in the queue. Note that $|\varphi_i^+(\mathbf{y})| = |\mathbf{y}| + 1$ and $\varphi_i^+(\mathbf{y})$ is defined only if $|\mathbf{y}| < L$.

(ii) $\varphi_i^-(\mathbf{y})$ is the new state resulting from a service completion at position i. Note that $|\varphi_i^-(\mathbf{y})| = |\mathbf{y}| - 1$ and $\varphi_i^-(\mathbf{y})$ is defined only if $s_i = k$ and $|\mathbf{y}| \geq 1$.

(iii) $\varphi_i(\mathbf{y})$ is the new state resulting from a phase change at position i (not resulting in a service completion). Note that $|\varphi_i(\mathbf{y})| = |\mathbf{y}|$ and $\varphi_i(\mathbf{y})$ is defined only if $0 < s_i < k$ and $|\mathbf{y}| \geq 1$.

(iv) $\varphi_f(\mathbf{y})$ is the new state resulting from a change in the flip-flop variable from f to $1 - f$ (e.g. an overflow occurs). Note that $|\varphi_f(\mathbf{y})| = |\mathbf{y}|$ and $\varphi_f(\mathbf{y})$ is defined only if $|\mathbf{y}| = L$.

Using this notation we can now write down the generator for $Y(t)$.

THEOREM 2.4. *The generator, A, for the Markov process* $\{Y(t)\}$ *is given by:*

(i) $A(\mathbf{y}, \varphi_i^+(\mathbf{y})) = \delta(i, N+1)\lambda$ *if* $N < L$;

(ii) $A(\mathbf{y}, \varphi_i^-(\mathbf{y})) = \gamma(i, N)\phi(N)k\mu$ *if* $s_i = k,\ N \geq 1$;

(iii) $A(\mathbf{y}, \varphi_i(\mathbf{y})) = \gamma(i, N)\phi(N)k\mu$ *if* $s_i < k,\ N \geq 1$;

(iv) $A(\mathbf{y}, \varphi_f(\mathbf{y})) = \lambda$ *if* $N = L$;

(v) $A(\mathbf{y}, \mathbf{y}) = -\lambda$ *if* $N = 0$;

(vi) $A(\mathbf{y}, \mathbf{y}) = -(\lambda + \phi(N)k\mu)$ *if* $N > 0$.

PROOF: It is well known (e.g. Çinlar [1975] p. 254) that entries of the generator of a Markov process are given by:

$$A(i,j) = \begin{cases} -\alpha(i) & \text{if} \quad i = j, \\ \alpha(i)P(i,j) & \text{if} \quad i \neq j, \end{cases}$$

where $\alpha(i)$ is the parameter of the exponential holding time in state i and $P(i,j)$ is the transition probability from state i to state j in the underlying Markov chain.

Consider, for example, part (ii) of Theorem 2.4. The exponential parameter of the holding time in state $\mathbf{y}$ is $\lambda + \phi(N)k\mu$ as in the proof of Theorem 2.2. The transition probability is found to be

$$P\Big(\mathbf{y}, \varphi_i^-(\mathbf{y})\Big) \;=\; \gamma(i, N)\phi(N)k\mu/(\lambda + \phi(N)k\mu)$$

since $\phi(N)k\mu/(\lambda+\phi(N)k\mu)$ is the probability that a phase change occurs before an arrival, and, $\gamma(i, N)$ is the probability that given the phase change that it occurs in position i. Therefore,

$$A(\mathbf{y}, \varphi_i^-(\mathbf{y})) \;=\; (\lambda + \phi(N)k\mu)\gamma(i, N)\phi(N)k\mu/(\lambda + \phi(N)k\mu)$$

$$= \gamma(i, N)\phi(N)k\mu\ .$$

Similar proofs could be given for the other parts. •

Now consider the bivariate discrete time stochastic process $\{(Y_n, T_n)\}$ where T_n is the epoch of the *nth* jump of $\{Y(t)\}$ and where $Y_n = Y(T_n+)$. Then $\{(Y_n, T_n)\}$ is a Markov renewal process with the following special structure. Let $\varphi(\mathbf{y})$ be a state reachable from $\mathbf{y}$ in one transition, then the semi-Markov kernel of $\{(Y_n, T_n)\}$ is given by

THEOREM 2.5.

$$\tilde{Q}(\mathbf{y}, \varphi(\mathbf{y}), t) = P\{Y_{n+1} = \varphi(\mathbf{y}), T_{n+1} - T_n \leq t | Y_n = \mathbf{y}\} =$$

$$\begin{cases} P(\mathbf{y}, \varphi(\mathbf{y}))(1 - e^{-\lambda t}) & \text{if } |\mathbf{y}| = 0 \\ P(\mathbf{y}, \varphi(\mathbf{y}))(1 - e^{-(\lambda + \phi(N)k\mu)t}) & \text{if } |\mathbf{y}| = N > 0 \end{cases}$$

where

$$P(\mathbf{y}, \varphi(\mathbf{y})) = A(\mathbf{y}, \varphi(\mathbf{y}))/ - A(\mathbf{y}, \mathbf{y})$$

is the transition probability of the Markov chain, $\{Y_n\}$. *Note that,*

$$P(\mathbf{y}, \varphi(\mathbf{y})) = \lim_{t \to \infty} \tilde{Q}(\mathbf{y}, \varphi(\mathbf{y}), t).$$

PROOF: See Çinlar [1975], Chapter 8. •

The discrete time process $\{Y_n\}$ is referred to as the Markov chain embedded just after all jumps. Because of the way we define $\{Y(t)\}$, every jump causes a change of state; i.e. $Y_{n+1} \neq Y_n$ a.s.. We may find it useful to allow $Y_{n+1} = Y_n$ if an overflow occurs at T_n. That is, when considering the successive states of the jump chain, we may want to drop the flip-flop variable as excess baggage. More precisely we will lump or combine states whenever the states differ only in the flip-flop component.

THEOREM 2.6. *Let* $Y_n = (s_{1,n}, s_{2,n}, \ldots, s_{m,n}, f_n)$ *and*
define $X_n = g(Y_n) = (s_{1,n}, \ldots, s_{m,n})$, *then* $\{(Y_n, T_n)\}$ *is lumpable to* $\{(X_n, T_n)\}$.

PROOF: Apply Lemma 3.2 of Serfoso [1971]. Partition the state space, E, of $\{Y_n\}$:

$$\{[\mathbf{y}] : \mathbf{y} \in E\} \qquad \text{where} \quad [\mathbf{y}] = \{\mathbf{y}, \mathbf{y}'\}$$

where $\mathbf{y}$ differs from $\mathbf{y}'$ only in the flip-flop component. In other words,

$$g(\mathbf{y}) = g(\mathbf{y}').$$

We must show that

$$\tilde{Q}(\mathbf{y}, [\mathbf{z}], t) = \tilde{Q}(\mathbf{y}', [\mathbf{z}], t) \tag{2.1}$$

for each $\mathbf{y}, \mathbf{z} \in E$.

Case 1: $\mathbf{y} \notin [\mathbf{z}]$. (The transition from $\mathbf{y}$ to $[\mathbf{z}]$ is not an overflow.)

$$\tilde{Q}(\mathbf{y}, [\mathbf{z}], t) = \tilde{Q}(\mathbf{y}, \mathbf{z}, t) = \tilde{Q}(\mathbf{y}', \mathbf{z}', t) = \tilde{Q}(\mathbf{y}', [\mathbf{z}'], t) = \tilde{Q}(\mathbf{y}', [\mathbf{z}], t).$$

Case 2; $\mathbf{y} \in [\mathbf{z}]$. (The transition from $\mathbf{y}$ to $[\mathbf{z}]$ is an overflow.)

$$\tilde{Q}(\mathbf{y}, [\mathbf{z}], t) = \tilde{Q}(\mathbf{y}, \mathbf{y}', t) = \tilde{Q}(\mathbf{y}', \mathbf{y}, t) = \tilde{Q}(\mathbf{y}', [\mathbf{z}], t).$$

In either case, (2.1) is satisfied. Therefore, the lumpability of $\{(Y_n, T_n)\}$ to $\{(X_n, T_n)\}$ is proved. •

Let $\{X_n\}$ be the lumped Markov chain of $\{(X_n, T_n)\}$, the corresponding lumped Markov Renewal Process. Let E' be the state space of $\{X(t)\}$ and $\{X_n\}$. Let $Q(t)$ be the semi-Markov kernel of $\{(X_n, T_n)\}$ then

$$Q(\mathbf{x}_1, \mathbf{x}_2, t) = Q(g(\mathbf{y}_1), g(\mathbf{y}_2), t) =$$
$$\begin{cases} \tilde{Q}(\mathbf{y}_1, \mathbf{y}_2, t) = \tilde{Q}(\mathbf{y}_1', \mathbf{y}_2', t) & \text{if } \mathbf{x}_1 \neq \mathbf{x}_2, \\ \tilde{Q}(\mathbf{y}_1, \mathbf{y}_1', t) = \tilde{Q}(\mathbf{y}_1', \mathbf{y}_1, t) & \text{if } \mathbf{x}_1 = \mathbf{x}_2. \end{cases}$$

This result is more easily understood by grouping the states of E by the value of the flop-flop variable. Then $\tilde{Q}(t)$ has the following structure:

$$\tilde{Q}(t) = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} \tilde{Q}_A(t) & \tilde{Q}_B(t) \\ \tilde{Q}_B(t) & \tilde{Q}_A(t) \end{pmatrix} \end{array}$$

and where

$$Q(t) = \tilde{Q}_A(t) + \tilde{Q}_B(t).$$

From the entries of $Q(t)$, it is easy to compute the corresponding entries of $\widehat{Q}_s$, the LS-transform of $Q(t)$.

THEOREM 2.7. *Let* $|\mathbf{x}| = |(s_1, s_2, \ldots, s_N)| = N$.

(i) $\widehat{Q}((0),(1),s) = \dfrac{\lambda}{\lambda + s}$ *(arrival to empty queue);*

(ii) $\widehat{Q}(\mathbf{x}, \varphi_i^+(\mathbf{x}), s) = \dfrac{\delta(i, N+1)\lambda}{\lambda + \phi(N)k\mu + s}$ *(arrival to position i) for* $0 < N < L$;

(iii) $\widehat{Q}(\mathbf{x}, \varphi_i^-(\mathbf{x}), s) = \dfrac{\gamma(i,N)\phi(N)k\mu}{\lambda + \phi(N)k\mu + s}$ *(departure from position i) for* $s_i = k$ *and* $N > 0$;

(iv) $\widehat{Q}(\mathbf{x}, \varphi_i(\mathbf{x}), s) = \dfrac{\gamma(i,N)\phi(N)k\mu}{\lambda + \phi(N)k\mu + s}$ *(phase change at position i) for* $s_i < k$ *and* $N > 0$;

(v) $\widehat{Q}(\mathbf{x}, \mathbf{x}, s) = \dfrac{\lambda}{\lambda + \phi(N)k\mu + s}$ *(arrival that overflows) for* $N > 0$.

PROOF:

(i) $Q((0),(1),t) = P((0),(1))(1 - e^{-\lambda t}) = (1)(1 - e^{-\lambda t})$.
Therefore, $\widehat{Q}((0),(1),s) = \dfrac{\lambda}{\lambda + s}$.

(ii) $Q(\mathbf{x}, \varphi_i^+(\mathbf{x}), t) = P(\mathbf{x}, \varphi_i^+(\mathbf{x}))(1 - e^{-(\lambda + \phi(N)k\mu)t})$

$$\delta(i,N)\frac{\lambda}{\lambda + \phi(N)k\mu}\left(\frac{\lambda + \phi(N)k\mu}{\lambda + \phi(N)k\mu + s}\right)$$

$$\frac{\delta(i,N)\lambda}{\lambda + \phi(N)k\mu + s}. \quad \bullet$$

Similar proofs could be given for the other parts.

The limiting probability $\pi(\mathbf{x}) = \lim_{t\to\infty} P[\{X(t) = \mathbf{x}\}]$ will be of importance to our analysis. Thus the computation of $\pi = (\pi(\mathbf{x}) : \mathbf{x} \in E')$ will be necessary if any quantitative results are to be obtained. Unfortunately, there is little that can be said about π in the general setting of this section. A closed form solution to $\pi A = 0; \pi \mathbf{u} = 1$ would seem to be unknown for this level of generality. However, in Chapter 4, explicit expressions for π will be determined for important special cases.

There is a very simple relationship between the stationary distributions of the Markov processes $\{X(t)\}$ and $\{Y(t)\}$. Suppose the state spaces of $\{X(t)\}$ and $\{Y(t)\}$ are ordered as follows. Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be vectors from the state space E or from E'. Let $\mathbf{x}_1$ precede $\mathbf{x}_2$ if $|\mathbf{x}_1| < |\mathbf{x}_2|$. In the case that $|\mathbf{x}_1| = |\mathbf{x}_2|$, then we will say $\mathbf{x}_1$ precedes $\mathbf{x}_2$ if $\mathbf{x}_1$ precedes $\mathbf{x}_2$ lexicographically. Let π be the stationary vector of $\{X(t)\}$ under the above ordering of states. Let $\widehat{\pi} = [\pi_0, \pi_1]$ be the stationary distribution of $\{Y(t)\}$ under the above ordering where π_f is the subvector corresponding to the states where the flip-flop variable has value f for $f = 0, 1$. Note that if the flip-flop variable is ignored, then the components of π, π_0, and π_1 all correspond to the same states and in the same order. Moreover they are related by the following:

THEOREM 2.8. $\pi_0 = \pi_1 = \frac{1}{2}\pi$

PROOF:

$$\pi A = 0 \text{ and } \pi \mathbf{u} = 1$$

$$\widehat{\pi}\widehat{A} = 0 \text{ and } \widehat{\pi}\mathbf{u} = 1 \tag{2.2}$$

where A and $\widehat{A}$ are respectively the generators for $\{Y(t)\}$ and $\{X(t)\}$ with the states in the above ordering. That these systems have unique solutions follows from the fact that these Markov processes are finite, irreducible and aperiodic. Suppose

$$|\widehat{\pi}| = m,$$

then

$$|\pi| = 2m$$

and

$$|\pi_0| = |\pi_1| = m.$$

Suppose A is partitioned into $m \times m$ blocks according to the value of the flip-flop variable. Then A has the following structure:

$$A = \begin{pmatrix} B & C \\ C & B \end{pmatrix}$$

where

$$B + C = \widehat{A}$$

.

$$(\pi_0 , \ \pi_1) \begin{pmatrix} B & C \\ C & B \end{pmatrix} = (0 , \ 0) \tag{2.3}$$

$$\pi_0 B + \pi_1 C = 0$$

$$\pi_0 C + \pi_1 B = 0$$

$$\pi_0(B + C) + \pi_1(B + C) = 0$$

$$(\pi_0 + \pi_1)(B + C) = 0.$$

Therefore,

$$\pi_0 + \pi_1 = \pi. \tag{2.4}$$

The system (2.3) can be written as

$$(\pi_1 , \ \pi_0) \begin{pmatrix} B & C \\ C & B \end{pmatrix} = (0 , \ 0).$$

But by (2.4) there is a unique solution to the system, (2.3). Therefore,

$$\pi_0 = \pi_1.$$

Combining (2.4) with the preceding equation gives the desired result. •

Even without general expressions for π we can still investigate how π fits into the theory for purposes of obtaining qualitative results for traffic processes and sojourn times. Additionally, the relationships between the stationary vector for

the continuous time process $\{X(t)\}$ and the stationary vectors for the various embedded processes can be explored. These ideas will be discussed in the remainder of Chapter 2.

2.3 Traffic Processes

We primarily will be interested in the traffic processes describing the arrivals, departures, inputs, outputs, and overflows. These processes will be studied by the filtering of Markov renewal processes. (See Çinlar [1969]. Hunter [1983a, 1983b, 1984, 1985] uses this technique extensively.) Figure 2.1 illustrates the various filterings or thinnings of interest. In order to analyze these filtered processes it is useful to consider the notion of traffic sets. Basically, a traffic set is a set of ordered pairs of states of $\{X_n\}$ or $\{Y_n\}$ which represents the one step transitions of interest that are to be filtered from the set of all transitions.

Definition 2.9. A traffic set is a set of ordered pairs of states representing a transition in the state process where the first coordinate represents the state just before the transition and the second coordinate represents the state just after transition.

The following are examples of traffic sets:

(i) $B_i = \{(\mathbf{x},\mathbf{y}) | \mathbf{y} = \varphi_j^+(\mathbf{x}), \quad 1 \le j \le |\mathbf{x}| + 1\}$ (input transitions)

(ii) $B_o = \{(\mathbf{x},\mathbf{y}) | \mathbf{y} = \varphi_j^-(\mathbf{x}), \quad 1 \le j \le |\mathbf{x}|\}$ (output transitions)

(iii) $B_v = \{(\mathbf{x},\mathbf{y}) | \mathbf{y} = \mathbf{x}\}$ (overflow transitions)

(iv) $B_a = B_i \cup B_v$ (arrival transitions)

(v) $B_d = B_o \cup B_v$ (departure transitions).

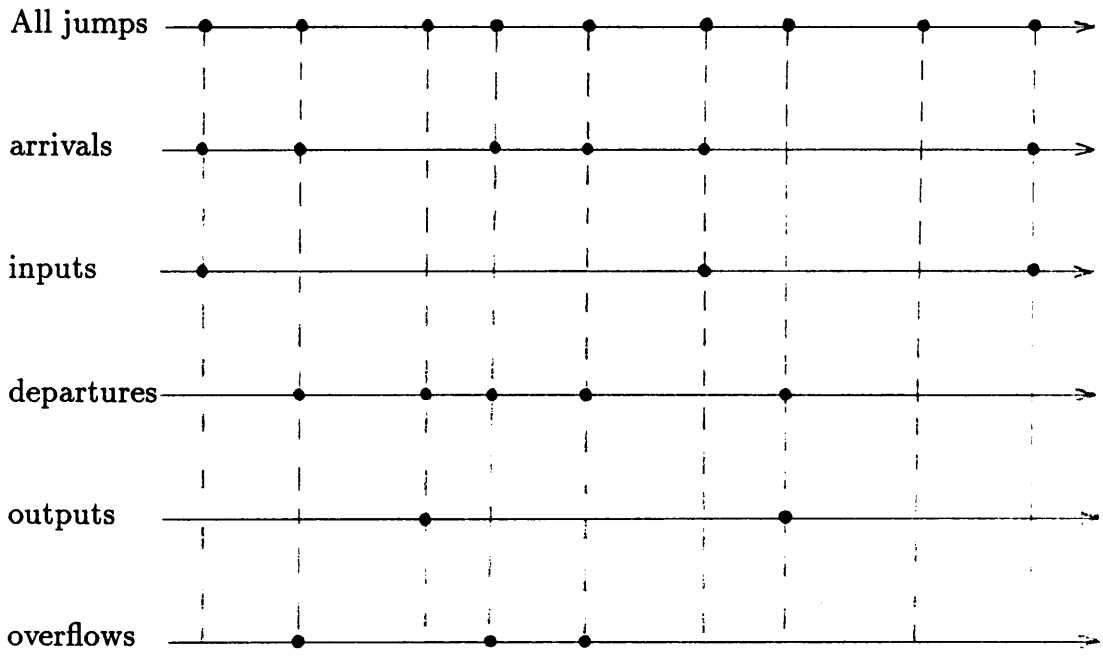

Figure 2.1 Filtering of a sample path

Definition 2.10.

(i) $V_r(\mathbf{x},\mathbf{y},t) = Q(\mathbf{x},\mathbf{y},t) \cdot 1_{B_r}(\mathbf{x},\mathbf{y})$

(ii) $U_r(\mathbf{x},\mathbf{y},t) = Q(\mathbf{x},\mathbf{y},t) \cdot 1_{B_r}(\mathbf{x},\mathbf{y})$ where $r \in \{a,i,d,o,v\}$ and

$$1_{B_r}(\mathbf{x},\mathbf{y}) = \begin{cases} 1 \text{ if } (\mathbf{x},\mathbf{y}) \in B_r \\ 0 \text{ if } (\mathbf{x},\mathbf{y}) \in \bar{B}_r \end{cases}$$

and $\bar{B}_r$ is the complement of B_r.

The matrix U_r has a zero in any entry corresponding to a transition *not* of type r and V_r has a zero in any entry which corresponds to a transition of type r. Note that $U_r(t) + V_r(t) = Q(t)$.

Definition 2.11. Let $\{(X_n^r, T_n^r)\}$ be the stochastic process where T_n^r is the transition epoch of type r and X_n^r be the state of the system just after T_n^r where $r \in \{a,i,d,o,v\}$.

THEOREM 2.12. *$\{(X_n^r, T_n^r)\}$ is a Markov renewal process on E_r' with semi-Markov kernel given by*

$$Q_r(t) = \sum_{j=0}^{\infty} U_r^{(j)}(t) * V_r(t) \tag{2.5}$$

for each $r \in \{a,i,d,o,v\}$, where $E_r' = \{\mathbf{y} : (\mathbf{x},\mathbf{y}) \in B_r,\ A(\mathbf{x},\mathbf{y}) > 0\}$, and where $U_r^{(0)}(t) = I$.

PROOF: See Theorem 3.1 of Disney and Kiessler [1987] for a proof of the result. •

But since this result and other similar results play such a central role in this thesis, an informal description follows which should capture the essence of the theorem.

Suppose that the *nth* transition of type r has just occurred. Then the current state of the system is some $\mathbf{x} \in E_r'$. Suppose the next state of the system is $\mathbf{y}$. This

transition is either of type r or it is not of type r (i.e. $(\mathbf{x},\mathbf{y}) \in B_r$ or $(\mathbf{x},\mathbf{y}) \in \bar{B}_r$). If $(\mathbf{x},\mathbf{y}) \in B_r$ then the $(n+1)st$ type r transition has occurred and

$$P_r\{X^r_{n+1} = \mathbf{y}, T^r_{n+1} - T^r_n \leq t | X^r_n = \mathbf{x}\}$$

is given by the $\mathbf{x},\mathbf{y}$ entry of $V_r(t)$. On the other hand if $(\mathbf{x},\mathbf{y}) \in \bar{B}_r$ then the process regenerates itself but from the new starting state y and this transition is governed by the appropriate entry from $U_r(t)$. Since it is possible that the system goes through any number of transitions from $\bar{B}_r$ before a transition of B_r is attained, then summation (2.5) follows.

For computational purposes we may consider the LS-transform of $Q_r(t)$ given in the following:

THEOREM 2.13. $\widehat{Q}_r(s) = \sum_{j=0}^{\infty} \widehat{U}_r^j(s)\widehat{V}_r(s) = (I - \widehat{U}_r(s))^{-1}\widehat{V}_r(s)$
where $\widehat{U}_r(s)$ and $\widehat{V}_r(s)$ are the LS-transforms of $U_r(t)$ and $V_r(t)$, respectively; and $\widehat{Q}_r(s)$ is the LS-transform of $Q_r(t)$.

PROOF: The summation results from the fact that the *LS*-transform of a convolution is the product of the *LS*-transforms. $\widehat{U}_r(s)$ is non-increasing in s and so $\widehat{U}_r(0) \geq \widehat{U}_r(s)$ for $s > 0$. But $\widehat{U}_r(0)$ is a substochastic matrix and $\|\widehat{U}_r(0)\| < 1$. Hence, $\|\widehat{U}_r(s)\| < 1$ for $s > 0$. Therefore, $\sum_{j=1}^{\infty} \widehat{U}_r^j(s)$ converges to $(I - \widehat{U}_r(s))^{-1}$ for $s > 0$. •

Once the transitional properties of the traffic processes are understood, the embedded stationary distributions can be determined. Let π^r and π^{r-} be respectively, the stationary distributions embedded just after and just before transitions of types r. One could compute $\pi^{r-}P_{r-} = \pi^{r-}$ and $\pi^r P_r = \pi^r$ where

$$P_r = Q_r(\infty) = \widehat{Q}_r(0)$$

and

$$P_r = Q_{r-}(\infty) = \widehat{Q}_{r-}(0).$$

However, in some cases it may be simpler to compute the continuous time distribution, π for $\{X(t)\}$ and find relationships between π and π^r or π^{r-}. One such relationship is the PASTA result which was described in Chapter 1. Since our model has a Poisson arrival process we have $\pi^{a-} = \pi$. A more far-reaching relationship is given by the following:

THEOREM 2.14. *Let π^r be the stationary distribution for $\{X_n^r\}$ then*

$$\pi^r(\mathbf{x}) = \sum_{\mathbf{y}\in E} \pi(\mathbf{y})A(\mathbf{y},\mathbf{x})1_{B_r}(\mathbf{y},\mathbf{x}) / \sum_{\mathbf{z}\in E}\sum_{\mathbf{y}\in E} \pi(\mathbf{y})A(\mathbf{y},\mathbf{z})1_{B_r}(\mathbf{y},\mathbf{z}).$$

PROOF: See Disney and Kiessler [1987], p. 90. •

THEOREM 2.15. *$\{(X_n^{r-}, T_n^r)\}$ is a Markov renewal process on E'' with semi-Markov kernel given by*

$$Q_{r-}(\mathbf{x},\mathbf{y},t) = \left(\sum_{\mathbf{z}\in E} \frac{V_r(\mathbf{x},\mathbf{z},\infty)}{V_r(\mathbf{x},E,\infty)}\right)\left(\int_0^t \left(\sum_{m=0}^{\infty}[U_r(\mathbf{z},\mathbf{y},ds)]^{(m)}\right) V_r(j,E,t-s)\right)$$

where $V_r(\mathbf{x},E,t) = \sum_{\mathbf{z}\in E} V_r(\mathbf{x},\mathbf{z},t)$ and $E_r = \{\mathbf{x} : (\mathbf{x},\mathbf{y}) \in B_r, A(\mathbf{x},\mathbf{y}) > 0\}$.

PROOF: See Theorem 3.4 of Disney and Kiessler [1987]. •

We now have the traffic processes $\{(X_n^r, T_n^r)\}$ and $\{(X_n^{r-}, T_n^r)\}$ completely determined in general. However, computing the entries of the kernels is practically impossible unless either the state space is small or other simplifying assumptions are made. Some important examples will be presented in Chapter 3.

2.4 Sojourn Times

The sojourn time of an arbitrarily tagged customer will be determined by considering the time to absorption in a specially constructed Markov process. This process begins when the customer arrives. The process enters an absorbing state

when the customer departs. If the customer overflows, then the sojourn time is zero. The state space of the process is necessarily larger than the state space for the traffic processes discussed in Section 2.3. Not only must the phase of service of all the customers in the queue be recorded, but, the state vector must contain sufficient information so that the tagged customer's whereabouts is always known.

The process will be defined as follows. Let C be the name of the tagged customer. We define the random process

$$W(t) = (X(t), l(t))$$

where

$$X(t) = (s_1(t), s_2(t), \ldots, s_N(t)),$$

$s_i(t)$ is the phase of service of the customer in position i, and t is the elapsed time since the arrival of C. The process $l(t)$ is the queue position of C at time t.

Let

$$W_0 = (\varphi^+_{l_0}(\mathbf{x}_0), l_0) = W(0+)$$

be the state of the process just after C inputs. Here the queue is in state $\mathbf{x}_0$ just before the input of C. Let Δ be an absorbing state that is reached when C departs.

Define the random variable

$$S_{W_0} = inf\{t : W(t) = \Delta; W(0+) = W_0\},$$

and

$$S = inf\{t : W(t) = \Delta\}.$$

Thus, S_{W_0} is the sojourn time of customer C given that the queue is in state W_0 just after input assuming C does not overflow. Random variable S is the unconditional sojourn time of C. The objective of this section is to characterize the probability distributions of S and S_{W_0}.

It should be noted that $W(t)$ has the same transition epochs as the $X(t)$ process defined in Section 2.1. In fact, the $X(t)$ component of the $W(t)$ process is essentially the same as the $X(t)$ process defined in Section 2.1 except for a shift in the time parameter by the arrival time of C. (Since we restart the time parameter to 0 as soon as customer C enters the queue.) Thus $X(t)$ remains Markov. The $l(t)$ component is determined by the current state of $W(t)$ and the transitions in the $X(t)$ process. Thus it is evident that

THEOREM 2.16. *$\{W(t)\}$ is a Markov process. Moreover if $(\mathbf{x}, l)$ is a generic non-absorbing state with $|\mathbf{x}| = N$, then the generator, $\mathbf{B}$, for the Markov process $\{W(t)\}$ is given by:*

(i) $\mathbf{B}((\mathbf{x}, l), (\varphi_i^+(\mathbf{x}), l)) = \delta(i, N+1)\lambda$ *if* $0 < N < L, i > l$;

(ii) $\mathbf{B}((\mathbf{x}, l), (\varphi_i^+(\mathbf{x}), l+1)) = \delta(i, N+1)\lambda$ *if* $0 < N < L, i \leq l$;

(iii) $\mathbf{B}((\mathbf{x}), l), (\varphi_i^-(\mathbf{x}, l)) = \gamma(i, N)\phi(N)k\mu$ *if* $N > 0$ *and* $s_i = k, i > l$;

(iv) $\mathbf{B}((\mathbf{x}, l), (\varphi_i^-(\mathbf{x}), l-1)) = \gamma(i, N)\phi(N)k\mu$ *if* $N > 0$ *and* $s_i = k, i \leq l$;

(v) $\mathbf{B}((\mathbf{x}, l), (\varphi_i(\mathbf{x}), l)) = \gamma(i, N)\phi(N)k\mu$ *if* $N > 0$ *and* $s_i < k$;

(vi) $\mathbf{B}((\mathbf{x}, l), (\mathbf{x}, l)) = -\phi(N)k\mu$ *if* $N = L$;

(vii) $\mathbf{B}((\mathbf{x}, l), (\mathbf{x}, l)) = -(\lambda + \phi(N)k\mu)$ *if* $0 < N < L$;

(viii) $\mathbf{B}((\mathbf{x}, l), \Delta)) = \gamma(l, N)\phi(N)k\mu$ *if* $N > 0$ *and* $s_l = k$.

Let T_n be the epoch of the *nth* transition of $W(t)$. Let $W_n = W(T_n+)$. Then $\{(W_n, T_n)\}$ is the Markov renewal process embedded at transitions of $\{W(t)\}$. Let $R(t)$ be the semi-Markov kernel for $\{(W_n, T_n)\}$ and $\widehat{R}(s)$ be the L.S. transform of $R(t)$. Let $(\varphi(\mathbf{x}), l')$ be a state reachable from $(\mathbf{x}, l)$ in one transition. Using Theorem 2.5 with the appropriate change in notation, we see that

$$R((\mathbf{x}, l), (\varphi(\mathbf{x}), l'), t)) =$$

$$P(W_{n+1} = (\varphi(\mathbf{x}), l'), T_{n+1} - T_n < t | W_n = (\mathbf{x}, l))$$

$$= P((\mathbf{x}, l), (\varphi(\mathbf{x}), l'))(1 - e^{-(\lambda+\phi(N)k\mu)t}),$$

where

$$P((\mathbf{x}, l), (\varphi(\mathbf{x}), l')) = \frac{B((\mathbf{x}, l), (\varphi(\mathbf{x}), l'))}{-B((\mathbf{x}, l), (\mathbf{x}, l))} .$$

From $R(t)$ it is simple to write down the entries of $\widehat{R}(s)$ in a manner similar to Theorem 2.7.

THEOREM 2.17.

(i) $\widehat{R}((\mathbf{x}, l), (\varphi_i^+(\mathbf{x}), l), s) = \dfrac{\delta(i, N+1)\lambda}{\lambda + \phi(N)k\mu + s}$ *if* $0 < N < L, i > l$;

(ii) $\widehat{R}((\mathbf{x}, l), (\varphi_i^+(\mathbf{x}), l+1), s) = \dfrac{\delta(i, N+1)\lambda}{\lambda + \phi(N)k\mu + s}$ *if* $0 < N < L, i \leq l$;

(iii) $\widehat{R}((\mathbf{x}, l), (\varphi_i^-(\mathbf{x}, l), s) = \dfrac{\gamma(i, N)\phi(N)k\mu}{\lambda + \phi(N)k\mu + s}$ *if* $N > 0$ *and* $s_i = k, i > l$;

(iv) $\widehat{R}((\mathbf{x}, l), (\varphi_i^-(\mathbf{x}), l-1), s) = \dfrac{\gamma(i, N)\phi(N)k\mu}{\lambda + \phi(N)k\mu + s}$ *if* $N > 0$ *and* $s_i = k, i \leq l$;

(v) $\widehat{R}((\mathbf{x}, l), (\varphi_i(\mathbf{x}), l), s) = \dfrac{\gamma(i, N)\phi(N)k\mu}{\lambda + \phi(N)k\mu + s}$ *if* $N > 0$ *and* $s_i < k$.

Once the semi-Markov kernel for $\{(W_n, T_n)\}$ is known , then the sojourn time distribution can be computed using a standard semi-regenerative argument. Let

$$F_{W_0}(t) = P[S_{W_0} \leq t] \tag{2.6}$$

and

$$F(t) = P[S \leq t] \tag{2.7}.$$

THEOREM 2.18.

$$F_{W_0}(t) = R(W_0, \Delta, t) + \sum_{\mathbf{y} \neq \Delta} R(W_0, \mathbf{y}, t) * F_{\mathbf{y}}(t) \tag{2.8}.$$

PROOF: Suppose the initial state is W_0 (just after C inputs assuming C does not overflow). Call the next state $\mathbf{y}$. Either $\mathbf{y} = \Delta$ or $\mathbf{y} \neq \Delta$. If $\mathbf{y} = \Delta$, then

absorption has occurred in which case $F_{W_0}(t) = R(W_0, \Delta, t)$. If on the other hand, $\mathbf{y} \neq \Delta$, then $F_{W_0}(t)$ is the convolution of $R(W_0, \mathbf{y}, t)$ and $F_{\mathbf{y}}(t)$ where $F_{\mathbf{y}}(t)$ is the distribution of the time to absorption from a starting state of $\mathbf{y}$. Summing over all possible states $\mathbf{y} \neq \Delta$ we get (2.8). •

The solution of the Markov renewal equations (2.8) exists and takes a form analogous to the distribution (2.5) of Theorem 2.12. In order to find (2.7) by unconditioning (2.6) we need to know the distribution for the initial state $W_0 = (\varphi_l^+(\mathbf{x}), l)$.

THEOREM 2.19.

$$\nu(\mathbf{x}, l) = P[W_0 = (\varphi_l^+(\mathbf{x}), l)] =$$

$$\pi^{i-}(\mathbf{x})\delta(l, |\mathbf{x}| + 1).$$

PROOF: If $W_0 = (\varphi_l^+(\mathbf{x}), l)$, then the state of the system just before the input of customer C is $\mathbf{x}$. The probability that the system is state $\mathbf{x}$ just before an input is $\pi^{i-}(\mathbf{x})$. Finally, given that C moves into position l with probability $\delta(l, |\mathbf{x}| + 1)$, the result easily follows. •

Now to compute the distribution for the unconditional sojourn time we find:

THEOREM 2.20.

$$F(t) = \sum_{|\mathbf{x}|=L} \pi(\mathbf{x}) + \Big(\sum_{|\mathbf{x}|<L} \pi(\mathbf{x})\Big) \sum_{\mathbf{x},l} \nu(\mathbf{x}, l) F_{W_0}(t)$$

where $W_0 = (\varphi_l^+(\mathbf{x}), l)$.

PROOF: The probability that C overflows is given by $\sum_{|\mathbf{x}|=L} \pi(\mathbf{x})$. Given that C does not overflow (i.e. C inputs), $\sum_{\mathbf{x},l} \nu(\mathbf{x}, l) F_{W_0}(t)$ is the sojourn time distribution. The probability that C does not overflow is given by $\sum_{|\mathbf{x}|<L} \pi(\mathbf{x})$. Thus, the result easily follows. •

In many cases the computation of ν is difficult, if not practically impossible, which causes problems if $F(t)$ is to be found. This may not be a major problem since $F_{W_0}(t)$ may be of primary interest.

In Chapter 3 it will be shown that many known and some new results about sojourn times can be deduced from the results of this section.

Chapter 3

Applications

3.1 Introduction

In this chapter the general theory of Chapter 2 is specialized to obtain known results as well as some results that appear to be new. Thus, the theory unifies many results in queueing theory which have been arrived at in piecemeal fashion. Furthermore, by showing some new results it is seen that our theory extends what is currently known.

In Section 3.2, we study the case of exponentially distributed service times. The model thus operates as a birth-death process. Most of these results are well known. The exponential service time distribution significantly simplifies the model. Not only is the dimension of the state space greatly reduced, but the queue discipline becomes irrelevant in most aspects of the model. The sojourn time is a notable exception.

The sojourn time problem in the case of the exponential server is discussed in Section 3.3. In Section 3.4, we study additional cases where the state process is reversible or dynamically reversible. These properties arise when the queue discipline is symmetric (i.e. $\delta = \gamma$). It will be shown that under this assumption the stationary distribution of the state process has a simple product form solution. Moreover, the queue length distribution is insensitive to the Erlang parameter, k. Some of these results of Section 3.4 were shown in Kelly [1979].

In Section 3.5, we consider the case of queue disciplines without preemption. That is, when a customer begins service, its service remains uninterrupted until completed. With this simplification, the single server model operates as a quasi-birth-death process.

3.2 The M/M/ϕ/L Case

In this case, the service times are independent and identically exponentially distributed with parameter μ. That is, the k parameter of the Erlang-k distribution is taken to be 1. Now, each customer in the system, whether it has received service or not, can be considered to be in its first and only phase of service. Thus, a typical state of the system would be

$$\mathbf{y} = (1, 1, \ldots, 1, f)$$

or

$$\mathbf{x} = (1, 1, \ldots, 1)$$

depending on whether or not we keep track of the overflow flip-flop variable. Note that if $|\mathbf{y}| = |\mathbf{x}| = N \leq L$, then we could just as well call these states

$$\mathbf{y} = (N, f) \quad \text{or} \quad \mathbf{x} = N$$

respectively. Because of this simplification, it is clear that the $\{Y(t)\}$ and $\{X(t)\}$ Markov processes will be unaffected no matter how the queue discipline, as defined by δ and γ, inserts and deletes customers from the queue. Since all customers are in the same phase of service, they are essentially indistinguishable. The $\{X(t)\}$ process is a finite birth-death process; but the $\{Y(t)\}$ is slightly more complicated. The generator for the $\{Y(t)\}$ process is given by

$$
\begin{array}{c}
 \\ (0,0) \\ (1,0) \\ \vdots \\ (L-1,0) \\ (L,0) \\ (0,1) \\ (1,1) \\ \vdots \\ (L-1,1) \\ (L,1)
\end{array}
\begin{array}{c}
\begin{array}{ccccccccccc}
(0,0) & (1,0) & \dots & (L-1,0) & (L,0) & (0,1) & (1,1) & \dots & (L-1,1) & (L,1)
\end{array} \\
\begin{pmatrix}
-\lambda & \lambda & & & & & & & & \\
\mu_1 & -\lambda-\mu_1 & & & & & & & & \\
 & \ddots & \ddots & \ddots & & & & & & \\
 & & & & \lambda & & & & & \\
 & & & \mu_L & -\lambda-\mu_L & & & & & \lambda \\
 & & & & & -\lambda & \lambda & & & \\
 & & & & & \mu_1 & -\lambda-\mu_1 & & & \\
 & & & & & & \ddots & \ddots & \ddots & \\
 & & & & & & & & & \lambda \\
 & & & & \lambda & & & & \mu_L & -\lambda-\mu_L
\end{pmatrix}
\end{array}
$$

where $\mu_n = \phi(n)\mu$. $\{Y(t)\}$ is lumpable to $\{X(t)\}$ which is just the queue length process. The generator, A, of $\{X(t)\}$ is given by

$$
\begin{array}{c}
 \\ 0 \\ 1 \\ 2 \\ \vdots \\ L-1 \\ L
\end{array}
\begin{array}{c}
\begin{array}{ccccccc}
0 & 1 & 2 & & \dots & L-1 & L
\end{array} \\
\begin{pmatrix}
-\lambda & \lambda & & & & & \\
\mu_1 & -(\lambda+\mu_1) & \lambda & & & & \\
 & \mu_2 & -(\lambda+\mu_2) & \ddots & & & \\
 & & \ddots & \ddots & & & \\
 & & & & & -(\lambda+\mu_{L-1}) & \lambda \\
 & & & & & \mu_L & -\mu_L
\end{pmatrix}
\end{array}
$$

THEOREM 3.1. *The stationary distributions for* $\{X(t)\}$ *and* $\{Y(t)\}$ *are given respectively by*

$$\pi = B(1, \frac{\lambda}{\phi(1)\mu}, \frac{\lambda^2}{\phi(1)\phi(2)\mu^2}, \ldots, \frac{\lambda^L}{\prod_i \phi(i)\mu^L})$$

and

$$\widehat{\pi} = \frac{1}{2}(\pi, \pi)$$

where B *is a normalizing constant.*

PROOF: It is easy to verify that $\pi A = 0$. $\widehat{\pi}$ follows from Theorem 2.8. •

THEOREM 3.2. $\{X(t)\}$ *is reversible.*

PROOF: The detailed balance conditions are satisfied; i.e.

$$\pi(i)A(i,j) = \pi(j)A(j,i) \quad \text{for} \quad 0 \leq i \leq L; \quad 0 \leq j \leq L. \qquad \bullet$$

Because $\{X(t)\}$ is a reversible Markov process, then several results concerning the traffic processes in the system follow from the work of Disney and Kiessler [1987]. Let B_r be a traffic set of $\{X(t)\}$. Consider the Markov renewal processes $\{(X_n^r, T_n^r)\}$ and $\{(X_n^{r-}, T_n^r)\}$ which define the traffic processes in our model. Define $B_{\hat{r}} = \{(i,j) : (j,i) \in B_r\}$ to be the reverse traffic set of B_r. Let $\{\widehat{X}(t)\}$ be the reverse Markov process of $\{X(t)\}$ which has generator defined by

$$\widehat{A}(i,j) = \frac{\pi(j)}{\pi(i)} A(j,i).$$

If we embed $\{\widehat{X}(t)\}$ before or after the jumps defined by $B_{\hat{r}}$, then we define Markov renewal processes $\{(\widehat{X}_n^{\hat{r}-}, \widehat{T}_n^{\hat{r}})\}$ and $\{(\widehat{X}_n^{\hat{r}}, \widehat{T}_n^{\hat{r}})\}$, respectively. By Lemma 4.3.5 of Disney and Kiessler [1987], we have the same stationary distribution, π^r, for both processes. Moreover, by Theorem 4.3.6 of the same reference we have that $\{(\widehat{X}_n^{\hat{r}-}, \widehat{T}_n^{\hat{r}})\}$ is the reverse Markov renewal process of $\{(X_n^r, T_n^r)\}$, which is to say that

$$\widehat{\pi}^{\hat{r}}(i)\widehat{Q}_{\hat{r}}(i,j,t) = \pi^{r}(j)Q_{r}(j,i,t)$$

for $i,j \in E$ and $t > 0$

or equivalently

$$\widehat{Q}_{\hat{r}-}(i,j,t) = \frac{\pi^{r}(j)}{\pi^{r}(i)}Q_{r}(j,i,t).$$

Finally, we have by Theorem 4.6.1 of Disney and Kiessler [1987] that $\{(X_n^{\hat{r}-}, T_n^{\hat{r}})\}$ is the reverse Markov renewal process $\{(X_n^{r}, T_n^{r})\}$.

We can now draw the following conclusions concerning our model.

THEOREM 3.3.

(i) $\{(X_n^{a-}, T_n^{a})\}$ *is the reverse MRP of* $\{(X_n^{d}, T_n^{d})\}$ *and*
$\{(X_n^{a}, T_n^{a})\}$ *is the reverse MRP of* $\{(X_n^{d-}, T_n^{d})\}$

(ii) $\{(X_n^{i-}, T_n^{i})\}$ *is the reverse MRP of* $\{(X_n^{o}, T_n^{o})\}$ *and*
$\{(X_n^{i}, T_n^{i})\}$ *is the reverse MRP of* $\{(X_n^{o-}, T_n^{o})\}$

(iii) $\{(X_n^{v}, T_n^{v})\}$ *is the reverse MRP of* $\{(X_n^{v-}, T_n^{v})\}$.

The preceding theorem has some interesting ramifications. By Theorem 2.13.3 of Disney and Kiessler, if two Markov renewal processes are reverses of each other, then their underlying point processes are reverses of each other. For example, $\{T_n^{d}\}$ is the reverse process of $\{T_n^{a}\}$. This means that

$$P(D_m^{d} \le t_0, \ldots, D_{m+k}^{d} \le t_k) = P(D_m^{a} \le t_k, \ldots, D_{m+k}^{a} \le t_0)$$

for all $m, k \in \mathcal{N}$ and $t_0, t_1, \ldots, t_k \in \Re+$

and where

$$D_n^{d} = T_n^{d} - T_{n-1}^{d} \qquad \text{and} \qquad D_n^{a} = T_n^{a} - T_{n-1}^{a} \ .$$

Since $\{T_n^{a}\}$ is assumed to be a Poisson process and hence is renewal, then it follows that $\{T_n^{d}\}$ is the same Poisson process.

THEOREM 3.4.

(i) $\pi^{a-} = \pi^{d} \qquad and \qquad \pi^{a} = \pi^{d-}$

(ii) $\pi^{i-} = \pi^{o} \qquad and \qquad \pi^{i} = \pi^{o-}$

(iii) $\pi^{v} = \pi^{v-}$

Now applying the formula on page 90 of Disney and Kiessler [1987], we can compute these embedded stationary distributions in terms of π, the stationary distribution of $\{X(t)\}$.

THEOREM 3.5.

(i) $\pi^{a-} = \pi^{d} = \pi$

(ii) $\pi^{i-} = \pi^{o} = B^{o}[1, \pi(2)\phi(2)\mu, \pi(3)\phi(3)\mu, \ldots, \pi(L)\phi(L)\mu, 0]$

(iii) $\pi^{v} = \pi^{v-} = [0, 0, \ldots, 0, 1]$

(iv) $\pi^{i} = \pi^{o-} = B^{i}[0, \pi(0), \pi(1), \ldots, \pi(L-1)]$

(v) $\pi^{a} = \pi^{d-} = [0, \pi(0), \pi(1), \ldots, \pi(L-2), \pi(L-1) + \pi(L)]$

Note that (v) must be computed using the generator of $\{Y(t)\}$ instead of $\{X(t)\}$.

Let us now find the semi-Markov kernels for some of the traffic processes of interest as discussed earlier in this section. The semi-Markov kernel for the MRP $\{(X_n, T_n)\}$ (embedded at all jumps) is given by

$$
\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ \vdots \\ L-1 \\ L \end{array}
\begin{array}{c}
\begin{array}{ccccccc} 0 & 1 & 2 & 3 & \dots & L-1 & L \end{array} \\
\left(\begin{array}{ccccccc}
 & a_0(s) & & & & & \\
b_1(s) & & a_1(s) & & & & \\
 & b_2(s) & & a_2(s) & & & \\
 & & \ddots & & \ddots & & \\
 & & & & & & \\
 & & & & & & a_{L-1}(s) \\
 & & & & & b_L(s) & a_L(s)
\end{array}\right)
\end{array}
$$

where

$$
\begin{aligned}
a_0(s) &= \frac{\lambda}{\lambda + s} \\
a_n(s) &= \frac{\lambda}{\lambda + \phi(n)\mu + s}, \qquad 0 < n \leq L \\
b_n(s) &= \frac{\phi(n)\mu}{\lambda + \phi(n)\mu + s}, \qquad 0 < n \leq L
\end{aligned}
$$

Consider $Q_o(t)$, the semi-Markov kernel for the output process $\{(X_n^o, T_n^o)\}$.

$\widehat{U}_o(s)$ is given by

$$
\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ \vdots \\ L-1 \\ L \end{array}
\begin{array}{c}
\begin{array}{ccccccc} 0 & 1 & 2 & 3 & \dots & L-1 & L \end{array} \\
\left(\begin{array}{ccccccc}
 & a_0(s) & & & & & \\
 & & a_1(s) & & & & \\
 & & & a_2(s) & & & \\
 & & & & \ddots & & \\
 & & & & & & \\
 & & & & & & a_{L-1}(s) \\
 & & & & & & a_L(s)
\end{array}\right)
\end{array}
$$

$\widehat{V}_o(s)$ is given by

$$
\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ \vdots \\ L-1 \\ L \end{array}
\begin{array}{c}
\begin{array}{ccccccc} 0 & 1 & 2 & 3 & \dots & L-1 & L \end{array} \\
\begin{pmatrix}
 & & & & & & \\
b_1(s) & & & & & & \\
 & b_2(s) & & & & & \\
 & & b_3(s) & & & & \\
 & & & \ddots & & & \\
 & & & & & & \\
 & & & & & b_L(s) &
\end{pmatrix}
\end{array}
$$

It is easy to show that $\sum_{k=0}^{\infty} \widehat{U}_o^k(s) = \sum_{k=0}^{L-1} \widehat{U}_o^k(s)$ is given by

$$
\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ \vdots \\ L-1 \\ L \end{array}
\begin{array}{c}
\begin{array}{ccccccc} 0 & 1 & 2 & 3 & \dots & L-1 & L \end{array} \\
\begin{pmatrix}
1 & a_0(s) & a_0(s)a_1(s) & a_0(s)a_1(s)a_2(s) & \dots & a_0(s)\dots a_{L-2}(s) & \frac{a_0(s)\dots a_{L-1}(s)}{1-a_L(s)} \\
 & 1 & a_1(s) & a_1(s)a_2(s) & \dots & a_1(s)\dots a_{L-2}(s) & \frac{a_1(s)\dots a_{L-1}(s)}{1-a_L(s)} \\
 & & 1 & a_2(s) & & & \\
 & & & 1 & & & \\
 & & & & \ddots & & \\
 & & & & & & \frac{a_{L-1}(s)}{1-a_L(s)} \\
 & & & & & & \frac{1}{1-a_L(s)}
\end{pmatrix}
\end{array}
$$

And we compute $\widehat{Q}_o(s) = \sum_{k=0}^{\infty} \widehat{U}_o^k(s)\widehat{V}_o(s) =$

$$
\begin{array}{c}
\\
0 \\
1 \\
\vdots \\
L-1 \\
L
\end{array}
\begin{array}{c}
\begin{array}{cccc} 0 & 1 & \cdots & L-1 \end{array} \\
\begin{pmatrix}
a_0(s)b_1(s) & a_0(s)a_1(s)b_2(s) & \cdots & \dfrac{a_0(s)\ldots a_{L-1}(s)b_L(s)}{1-a_L(s)} \\
b_1(s) & a_1(s)b_2(s) & \cdots & \dfrac{a_1(s)\ldots a_{L-1}(s)b_L(s)}{1-a_L(s)} \\
 & & \ddots & \vdots \\
 & & & \dfrac{a_{L-1}(s)b_L(s)}{1-a_L(s)} \\
 & & & \dfrac{b_L(s)}{1-a_L(s)}
\end{pmatrix}
\end{array}
\begin{array}{c}
\\
L \\
0 \\
0 \\
\vdots \\
0 \\
0
\end{array}
$$

Note that we can delete the last row and last column since

$$L \notin E_o' = \{j : (i,j) \in B_o, A(i,j) > 0\} .$$

That is, we would never find the system in state L just after an output.

The entries of $\widehat{Q}_o(s)$ can be explained easily intuitively. For example, consider the entry of row 0 and column 1. Just after the previous output the queue is empty and just after the next output there is one customer. This means that between these consecutive outputs there must have been an arrival to the empty queue, followed by another arrival and then the output from the queue of size 2. The conditional distribution of this inter-output time is then the convolution, $a_0(s)a_1(s)b_2(s)$. For a more complicated example, consider the entry in row 2 and column $L-1$. The previous output left 1 customer and the next output left $L-1$ customers. In the interim there must have been $L-1$ arrivals bringing the queue to the capacity, L. At this point there could be any number of arrivals causing any number of overflows, zero overflows or more. Finally output from the full queue occurs. For example, if there were k overflows, then the conditional distribution

of this inter-output time would be

$$a_1(s)a_2(s)\dots a_{L-1}(s)\Big(a_L(s)\Big)^k b_L(s).$$

Summing over all possible k gives

$$a_1(s)a_2(s)\dots a_{L-1}\Big(\frac{1}{1-a_L(s)}\Big)b_L(s).$$

Using similar computations, we can get the kernel $\widehat{Q}_i(s)$ for the input process, $\{(X_n^i, T_n^i)\}$:

$$\begin{array}{c|c} & \begin{array}{ccccccc} 0 & 1 & 2 & \dots & L-1 & L \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ \vdots \\ L-1 \\ L \end{array} & \left(\begin{array}{ccccccc} 0 & a_0(s) & & & & \\ 0 & b_1(s)a_0(s) & a_1(s) & & & \\ 0 & b_2(s)b_1(s)a_0(s) & b_2(s)a_1(s) & & & \\ & & & & & \\ 0 & b_{L-1}\dots b_1(s)a_0(s) & b_{L-1}(s)\dots b_2(s)a_1(s) & \dots & b_{L-1}(s)a_{L-2}(s) & a_{L-1}(s) \\ 0 & \frac{b_L(s)\dots b_1(s)a_0(s)}{1-a_L(s)} & \frac{b_L(s)\dots b_2(s)a_1(s)}{1-a_L(s)} & \dots & \frac{b_L(s)b_{L-1}(s)a_{L-2}(s)}{1-a_L(s)} & \frac{b_L(s)a_{L-1}(s)}{1-a_L(s)} \end{array}\right) \end{array}$$

And we delete the first row and first column since

$$0 \notin E_i' = \{(i,j) \in B_i, A(i,j) > 0\} .$$

Similar computations could be used to compute the semi-Markov kernel of the departure process, $\{(X_n^d, T_n^d)\}$. These results will not be presented here. However, as noted previously, the underlying point process $\{T_n^d\}$ is the same Poisson process as in the arrival process. Now consider the overflow process $\{(X_n^v, T_n^v)\}$. Note that this is a one state Markov renewal process since the system can only be in state L just after an overflow. Hence the overflow process is actually a renewal process, but not independent of $\{X_n^v\}$.

To compute $\widehat{Q}_v(s)$ first note that $\widehat{U}_v(s)$ is given by

$$\widehat{U}_v(s) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ L-1 \\ L \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & \dots & L-1 & L \end{array} \\ \begin{pmatrix} 0 & a_0(s) & & & & \\ b_1(s) & 0 & a_1(s) & & & \\ & b_2(s) & 0 & \ddots & & \\ & & \ddots & \ddots & & \\ & & & & 0 & a_{L-1}(s) \\ & & & & b_L(s) & 0 \end{pmatrix} \end{array}$$

and $\widehat{V}_v(s)$ is given by

$$\widehat{V}_v(s) = \begin{array}{c} \\ 0 \\ 2 \\ \vdots \\ L \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \dots & & L \end{array} \\ \begin{pmatrix} & & & & & \\ & & & & & \\ & & & & & \\ & & & & & a_L(s) \end{pmatrix} \end{array}$$

To determine $\widehat{Q}_v(s)$ we compute $\sum_{j=0}^{\infty} \widehat{U}_v^j(s)\widehat{V}_v(s) = (I - \widehat{U}_v(s))^{-1}\widehat{V}_v(s)$ and take the entry of this matrix from the Lth row and Lth column. None of the other state, $0, \dots, L-1$, could ever occur just after an overflow, so that all these states could be deleted leaving a 1×1 matrix for the kernel of the overflow process. This single entry of the kernel gives the interoverflow time distribution which completely characterizes the behavior of the renewal overflow process. It can be shown that the interoverflow distribution, $\phi_L(s)$, can be found recursively by a direct computation of $\widehat{Q}_v(s) = (I - \widehat{U}_v(s))^{-1}\widehat{V}_v(s)$:

$$I - \widehat{U}_v(s) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ L-1 \\ L \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \dots & L-1 & L \end{array} \\ \begin{pmatrix} 1 & -a_0(s) & & & & \\ -b_1(s) & 1 & -a_1(s) & & & \\ & -b_2(s) & 1 & \ddots & & \\ & & \ddots & \ddots & & \\ & & & & 1 & -a_{L-1}(s) \\ & & & & -b_L(s) & 1 \end{pmatrix} \end{array}$$

The determinant of $I - \widehat{U}_v(s)$, denoted $D_L(s)$, can be computed by expanding along the last row giving

$$D_{L-1}(s) - b_L(s)a_{L-1}(s)D_{L-2}(s)$$

where D_{L-1} is the determinant of the matrix resulting from deleting the last row and last column. Note that this smaller matrix would be the $I - \widehat{U}_v(s)$ matrix for a system of capacity $L-1$. Now, to complete the computation of the $L \times L$ entry in $(I - \widehat{U}_v(s))^{-1}$ the $L \times L$ entry of the adjoint of $I - \widehat{U}_v(s)$ is computed which is just the cofactor $D_{L-1}(s)$. The $L \times L$ entry of $(I - \widehat{U}_v(s))^{-1}$ is just

$$\frac{D_{L-1}(s)}{D_{l-1}(s) - b_L(s)a_{L-1}(s)D_{L-2}(s)}.$$

Finally, when $(I - \widehat{U}_v(s))^{-1}$ is multiplied by $\widehat{V}_v(s)$, (which only has a nonzero entry, $a_L(s)$, in the $L \times L$ position) we have

$$\phi_L(s) = \frac{a_L(s)D_{L-1}(s)}{D_{L-1}(s) - b_L(s)a_{L-1}(s)D_{L-2}(s)}$$

for $L = 1, 2, \dots$ and where $D_0 = D_{-1} = 1$.

This characterization of the interoverflow distribution is consistent with the known results discussed in Chapter 1.

3.3 The Sojourn Time in M/M/ϕ/L

In this section, the sojourn time in the M/M/ϕ/L queue is analyzed. The queue discipline is arbitrary within the class of disciplines defined by δ and γ. In particular, the sojourn times under the FCFS discipline and under the LCFS discipline with preemption are computed for the single server case with queue capacity $L = 2$.

Consider an arriving customer, C, at an arbitrary input epoch. At this epoch, the sojourn time of C begins. Define

$$W(t) = (N(t), l(t))$$

where $N(t)$ is the total number of customers in the queue at time t and $l(t)$ is the position in the queue of customer C at time t. The time parameter t is taken to be zero at the input epoch of customer C. Note that the $N(t)$ component of $W(t)$ is a scalar with $0 \leq N(t) \leq L$ whereas in the general case of chapter 2, this component was a vector. Thus, the dimension of the state space of $\{W(t)\}$ has been reduced significantly.

Let $N_0 = N(0-)$ be the queue length just before the arrival of C. Let l_0 be the initial queue position of C. Thus

$$W_0 = W(0+) = (N_0 + 1, l_0)$$

gives the state of the process just after C inputs. Recall that Δ is taken to be an absorbing state of $\{W(t)\}$ which is entered when C departs from the system. Therefore, the sojourn time of C, or the time that elapses between the input of C and the output of C, is just the first passage time of the $\{W(t)\}$ process from the state W_0 to the absorbing state Δ; or, in the case that C overflows then we let $W(0+) = \Delta$ and the sojourn time is zero.

The random variables of interest are the conditional and unconditional sojourn times:

$$S_{W_0} = inf\{t : W(t) = \Delta; W_0 = (N_0 + 1, l_0)\}$$

and

$$S = inf\{t : W(t) = \Delta\}$$

respectively.

Recall from Chapter 2 that $\{W(t)\}$ is a Markov process and $\{(W_n, T_n)\}$ is the Markov renewal process embedded just after jumps of $\{W(t)\}$. The semi-Markov kernel parallels the results of Theorem 2.17. In terms of LS transforms, $\widehat{R}(s)$ is given by:

(i) $\widehat{R}\Big((N,l),(N+1,l),s\Big) = \dfrac{\delta(i,N+1)\lambda}{\lambda+\phi(N)k\mu+s}$ if $N < L, \quad i > l$;

(ii) $\widehat{R}\Big((N,l),(N+1,l+1),s\Big) = \dfrac{\delta(i,N+1)\lambda}{\lambda+\phi(N)k\mu+s}$ if $N < L, \quad i \leq l$;

(iii) $\widehat{R}\Big((N,l),(N-1,l),s\Big) = \dfrac{\gamma(i,N)\phi(N)k\mu}{\lambda+\phi(N)k\mu+s}$ if $s_i = k, \quad i > l$;

(iv) $\widehat{R}\Big((N,l),(N-1,l-1),s\Big) = \dfrac{\gamma(i,N)\phi(N)k\mu}{\lambda+\phi(N)k\mu+s}$ if $s_i = k, \quad i \leq l$;

(v) $\widehat{R}\Big((N,l),(N,l),s\Big) = \dfrac{\gamma(i,N)\phi(N)k\mu}{\lambda+\phi(N)k\mu+s}$ if $s_i < k$.

The conditional sojourn time distribution, $F_{W_0}(t)$, satisfies the Markov renewal equation in Theorem 2.18. By Theorem 2.19 the distribution of the initial state, W_0, is given by

$$\nu(N_0, l_0) = P[W_0 = (N_0 + 1, l_0)] =$$

$$\pi^{i-}(N_0)\ \delta(l_0, N_0 + 1)$$

assuming that C is not an overflow. The unconditional sojourn time distribution is thus given by

$$F(t) = \pi(L) + \Big(\sum_{n<L} \pi(n)\Big) \sum_{n,l} \nu(n,l) F_{(n+1,l)}(t).$$

To illustrate the above results, consider the M/M/1/3-FCFS model. The state space of $\{W(t)\}$ is

$$E = \Big\{(1,1),(2,1),(2,2),(3,1),(3,2),(3,3)\Big\} \cup \Big\{\Delta\Big\}.$$

Recall that for the FCFS queue discipline,

$$\delta(N+1, N+1) = 1, \text{ if } N < L;$$

and

$$\delta(l, N+1) = 0 \text{ for } l \neq N+1;$$

Furthermore,

$$\gamma(1, N) = 1;$$

and

$$\gamma(l, N) = 0 \text{ for } l \neq 1 \ .$$

Thus the semi-Markov kernel for $\{(W_n, T_n)\}$ is given by

$$\widehat{R}(s) = \begin{array}{c} \\ (1,1)^* \\ (2,1) \\ (2,2)^* \\ (3,1) \\ (3,2) \\ (3,3)^* \\ \Delta \end{array} \begin{array}{c} \begin{array}{ccccccc} (1,1)^* & (2,1) & (2,2)^* & (3,1) & (3,2) & (3,3)^* & \Delta \end{array} \\ \begin{pmatrix} & a(s) & & & & & b(s) \\ & & & a(s) & & & b(s) \\ b(s) & & & & a(s) & & \\ & & & & & & c(s) \\ & c(s) & & & & & \\ & & c(s) & & & & \\ & & & & & & \end{pmatrix} \end{array}$$

where

$$a(s) + \frac{\lambda}{\lambda+\mu+s}, \qquad b(s) = \frac{\mu}{\lambda+\mu+s}, \qquad c(s) = \frac{\mu}{\mu+s} \ .$$

Note that the states marked with an asterisk are the only ones possible at the start of C's sojourn time, since these states are the states such that C is at the back of the queue. The vector of conditional sojourn times as determined from Theorem 2.18 is given by

$$\begin{matrix}(1,1)\\(2,1)\\(2,2)\\(3,1)\\(3,2)\\(3,3)\end{matrix}\begin{pmatrix} b(s)+a(s)b(s)+a^2(s)c(s) \\ - \\ b(s)\Big[b(s)+a(s)b(s)+a^2(s)c(s)+a(s)c(s)[b(s)+a(s)c(s)]\Big] \\ - \\ - \\ c(s)\Big[b(s)[b(s)+a(s)b(s)+a^2(s)c(s)]+a(s)c(s)[b(s)+a(s)c(s)]\Big] \end{pmatrix}$$

which simplifies to

$$\begin{matrix}(1,1)\\(2,2)\\(3,3)\end{matrix}\begin{pmatrix} \frac{\mu}{\mu+s} \\ (\frac{\mu}{\mu+s})^2 \\ (\frac{\mu}{\mu+s})^3 \end{pmatrix}$$

The suppressed entries in the above vector represent the distribution of the time to absorption from states that cannot possibly initiate a sojourn time. Therefore, we do not bother to compute these.

The previous result is obvious by a more direct argument. For example, the entry $\Big(\mu/(\mu+s)\Big)^2$ is just the LS transform of the convolution of two exponential service time distributions, which is certainly the sojourn time distribution of a customer who enters the second position of a queue operating under a FCFS

discipline. The interesting aspect of the result is the fact that it was arrived at as a special case of a more general theory.

The entries of the unsimplified version of the above vector were determined using a useful technique involving a tree graph. The tree illustrates the various ways that $\{(W_n, T_n)\}$ can reach Δ. Each branch of the tree represents a transition of $\{W_n\}$ and is labelled with the appropriate entry of $\widehat{R}(s)$. Thus to find the distribution of the time to absorption from some state we simply multiply along each path from that starting state to Δ and then sum these products. Figure 3.1 illustrates the transition tree for this example. This technique offers a simple procedure in many cases, especially when computations are done by hand. More importantly, it provides a picture of the semi-regenerative structure inherent in the Markov renewal equation. Often the tree has infinitely long paths, but if a recursive structure is identified then the technique still may prove useful.

To compute the unconditional sojourn time distribution we need the stationary distribution of the starting state, W_0. It is easy to compute

$$\nu = B[1, 0, \rho, 0, 0, \rho^2]$$

where $B = \dfrac{1}{1+\rho+\rho^2}$ and $\rho = \dfrac{\lambda}{\mu}$. That is

$$\nu = \left(\frac{\mu^2}{\mu^2+\lambda\mu+\lambda^2}, 0, \frac{\lambda\mu}{\mu^2+\lambda\mu+\lambda^2}, 0, 0, \frac{\lambda^2}{\mu^2+\lambda\mu+\lambda^2}\right).$$

Thus the unconditional sojourn time distribution given that an overflow did not occur, is given by

$$\left(\frac{\mu}{\mu+s}\right)\left(\frac{\mu^2}{\mu^2+\lambda\mu+\lambda^2}\right) + \left(\frac{\mu}{\mu+s}\right)^2\left(\frac{\lambda\mu}{\mu^2+\lambda\mu+\lambda^2}\right) + \left(\frac{\mu}{\mu+s}\right)^3\left(\frac{\lambda^2}{\mu^2+\lambda\mu+\lambda^2}\right)$$

$$= \left(\frac{\mu^2}{\mu^2+\lambda\mu+\lambda^2}\right)\frac{(\mu+s)^2+\lambda\mu(\mu+s)+\lambda^2\mu}{(\mu+s)^3}.$$

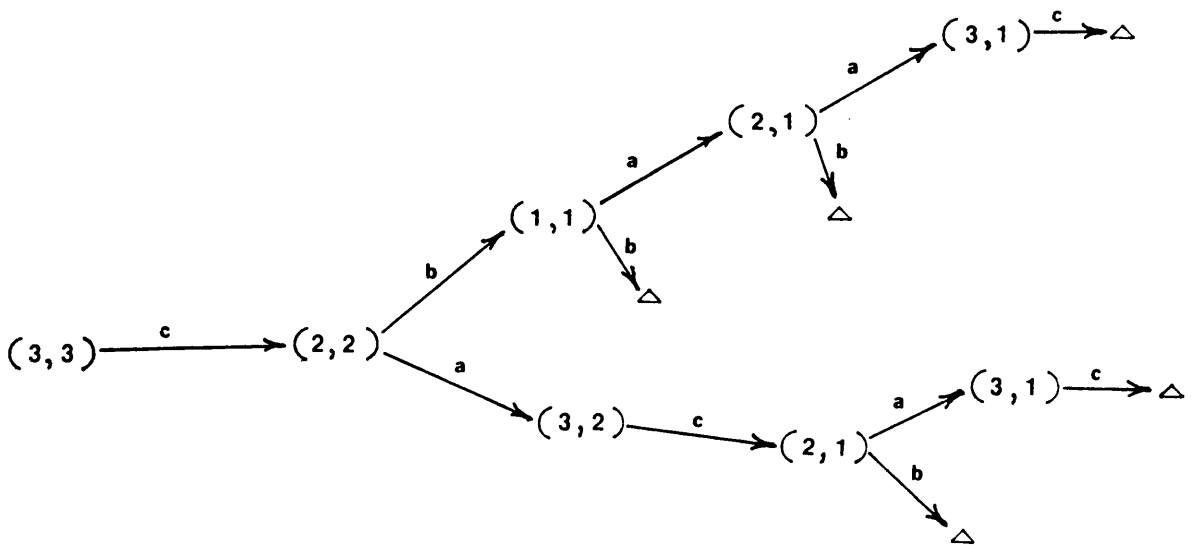

Figure 3.1 Tree graph for M/M/1/3-FCFS sojourn time

That is, the unconditional sojourn time is a linear combination of an exponential, an Erlang-2 and an Erlang-3 distribution. Once this transform is determined we can compute moments. For example, the expected sojourn time is given by

$$E[S] = -\widehat{F}'(0) = \frac{\mu + 2\lambda\mu + 3\lambda^3}{\mu(\mu^2 + \lambda\mu + \lambda^2)}$$

$$= \frac{1}{\mu}\Big(\frac{\mu + 2\lambda\mu + 3\lambda^2}{\mu^2 + \lambda\mu + \lambda^2}\Big) = E[\text{service time}] \times E[N] \ .$$

This last equality also follows from Little's formula.

For another example, consider the M/M/1/3 model operating with the LCFS-PR discipline. If an arriving customer finds the system full then the customer overflows as usual.

The state space of the model is still

$$E = \Big\{(1,1),(2,1),(2,2),(3,1),(3,2),(3,3)\Big\} \cup \Big\{\Delta\Big\},$$

but now we have

$$\delta(1, N+1) = 1 \text{ if } N < L,$$

$$\delta(l, N+1) = 0 \text{ if } N < L, 1 < l \leq N+1,$$

and

$$\gamma(1, N+1) = 1 \text{ if } N < L,$$

$$\gamma(l, N+1) = 0 \text{ if } N < L \text{ and } 1 < l \leq N+1.$$

The semi Markov kernel for $\{(W_n, T_n)\}$ is given by

$$
\begin{array}{c}
 \\
(1,1)^* \\
(2,1)^* \\
(2,2) \\
(3,1)^* \\
(3,2) \\
(3,3) \\
\Delta
\end{array}
\begin{array}{c}
\begin{array}{ccccccc}
(1,1)^* & (2,1)^* & (2,2) & (3,1)^* & (3,2) & (3,3) & \Delta
\end{array} \\
\begin{pmatrix}
 & a(s) & & & & & b(s) \\
 & & & a(s) & & & b(s) \\
b(s) & & & & a(s) & & \\
 & & & & & & c(s) \\
 & c(s) & & & & & \\
 & & c(s) & & & & \\
 & & & & & &
\end{pmatrix}
\end{array}
$$

where again

$$a(s) = \frac{\lambda}{\lambda + \mu + s}, \qquad b(s) = \frac{\mu}{\lambda + \mu + s}, \qquad c(s) = \frac{\mu}{\mu + s}\ .$$

After much algebraic manipulation, the vector of conditional sojourn time distributions is obtained:

$$
\begin{array}{c}
 \\
(1,1) \\
(2,1) \\
(3,1)
\end{array}
\begin{array}{c}
\begin{array}{c}
\end{array} \\
\begin{pmatrix}
\dfrac{b(s)}{1 - a(s)\frac{b(s)}{1-a(s)c(s)}} \\
\dfrac{b(s)}{1 - a(s)c(s)} \\
c(s)
\end{pmatrix}
\end{array}
$$

where the unconditional sojourn time distribution, given that an overflow did not occur, is easily obtained by multiplying the above vector by the distribution for the starting state of customer C,

$$(\frac{\mu^2}{\mu^2 + \lambda\mu + \lambda^2}, \frac{\lambda\mu}{\mu^2 + \lambda\mu + \lambda^2}, \frac{\lambda^2}{\mu^2 + \lambda\mu + \lambda^2})\ .$$

Note that the sojourn time conditional on the starting state (3,1) is just $\mu/(\mu + s)$ which is the distribution of just one service time. This makes sense since new

arrivals do not preempt the customer in service when the queue is at capacity. The entries in the vector of conditional sojourn times exhibit a recursive structure in this case. That this structure continues to a capacity of L can be verified using mathematical induction. The transition tree of Figure 3.2 helps to visualize the recursive structure of the following result:

$$\widehat{F}_{(N+1,1)}(s) = \frac{b(s)}{1 - a(s)\widehat{F}_{(N,1)}(s)}, \qquad N < L.$$

3.4 Symmetric Queue Disciplines

One of the key features of the M/M/ϕ/L class of queues is the property of reversibility of the continuous time state processes, $X(t)$ and $Y(t)$. The reversibility property is independent of the queue discipline in that special case. When the exponential service times are generalized to the E_k server for $k \geq 2$, then the reversibility of the continuous time state process is lost. Moreover, in this more general setting, the continuous time state process will depend on the queue discipline. In this section, a class of queue disciplines is considered which causes the state processes to be dynamically reversible. By salvaging this weaker notion of reversibility, the queue retains some of the simplicity inherent in the exponential server case. For example, the stationary queue length distribution maintains a product form. Moreover, certain relationships among the traffic processes are retained.

Recall that a queue discipline, as defined by the δ and γ functions, is said to be symmetric if $\delta = \gamma$. That is, when an arrival occurs to a queue of length n, then that arriving customer enters position l with probability $\gamma(l, n+1)$ which is also the proportion of the total service effort directed toward the customer in position l when $n+1$ customers are in the queue.

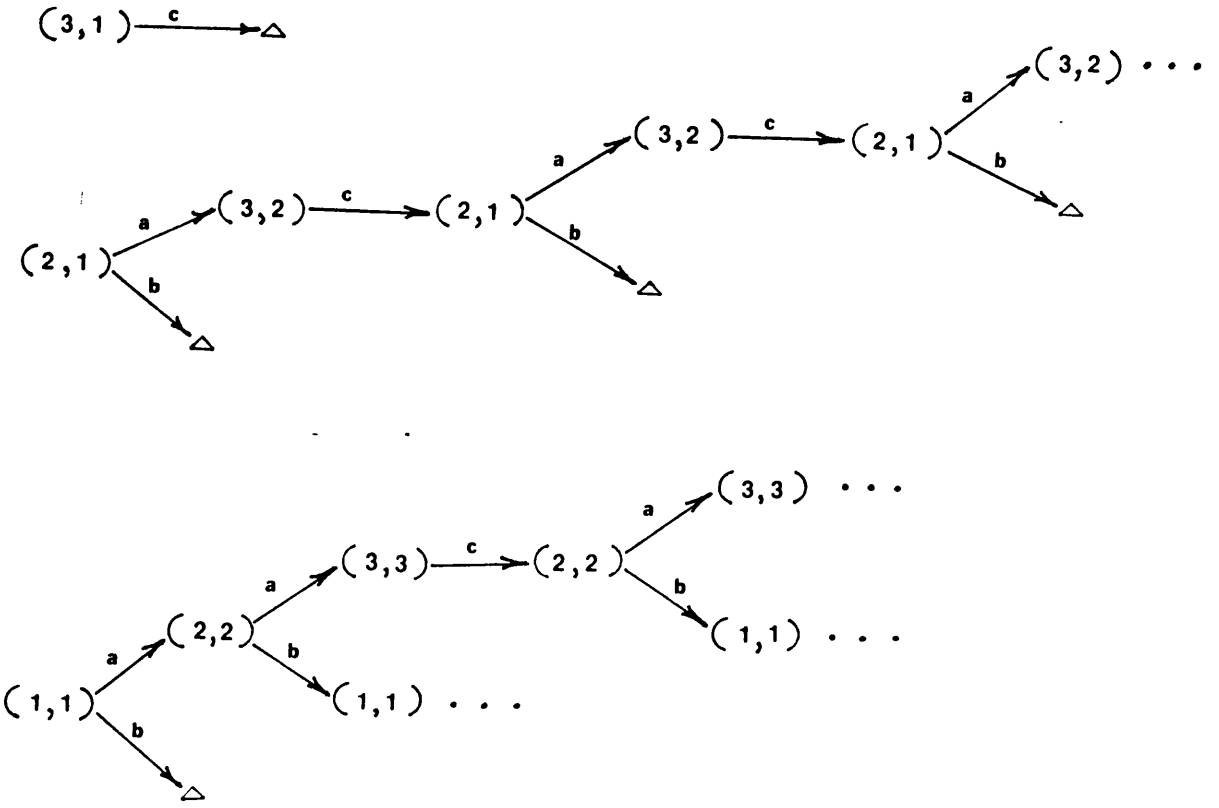

Figure 3.2 Tree graph for M/M/1/3-LCFS-PR sojourn time

Kelly [1979] presented some results concerning the $M/E_k/\phi$ queue with the symmetric queue discipline. In general, Kelly assumed an infinite capacity but his results are readily adaptable to the finite case. Kelly's model has the additional feature of including customer classes, but this feature is omitted here. To be more precise, we put all our customers into the same class. One primary result from Kelly is the product form stationary distribution for the continuous time state process. By suitably adapting the results of Section 3.3 of Kelly [1979], we get the following:

THEOREM 3.6. *The stationary distribution for the continuous time state process, $\{X(t)\}$, for the $M/E_k/\phi/L$ model with symmetric queue discipline is given by*

$$\pi(\mathbf{x}) = B(\frac{\lambda}{k\mu})^n(\prod_{i=1}^{n}\phi(i))^{-1} \quad \textit{if} \quad |\mathbf{x}| = n \tag{3.1}$$

where B is a normalizing constant which allows $\sum_{\mathbf{x}}\pi(\mathbf{x}) = 1$.

Note that in the single server case,

$$\pi(\mathbf{x}) = B(\frac{\lambda}{k\mu})^n \quad \text{if} \quad |\mathbf{x}| = n.$$

It is apparent that each state such that $|\mathbf{x}| = n$ has the same stationary probability. Since there are k^n such states, then the marginal stationary queue length probability is given by

$$P[|\mathbf{x}| = n] = B(\frac{\lambda}{\mu})^n(\prod_{i=1}^{n}\phi(i))^{-1}$$

Note that this queue length probability is independent of k. Thus, it can now be stated that:

THEOREM 3.7. *For symmetric queue disciplines, the stationary queue length distribution for the $M/E_k/\phi/L$ queue is insensitive to the choice of the E_k service time distribution and depends on this distribution only through its mean, μ^{-1}.*

To see that $\{X(t)\}$ is dynamically reversible, consider the following permutation of the states:

$$\mathbf{x} = (s_1, s_2, \ldots, s_n) \rightarrow \mathbf{x}^+ = (k - s_1 + 1, k - s_2 + 1, \ldots, k - s_n + 1) \quad \text{if } |\mathbf{x}| = n > 0$$

$$\mathbf{x} = (0) \rightarrow \mathbf{x}^+ = (0)$$

Note that in this rearrangement of states, the current phase of service of the customer in position l is being replaced by the remaining phases to be traversed (including the current uncompleted phase).

THEOREM 3.8. *For symmetric queue disciplines, the state process, $\{X(t)\}$, for the $M/E_k/\phi/L$ queue is dynamically reversible.*

PROOF: Let $\mathbf{x} = (s_1, s_2, \ldots, s_n)$ be the state just before transition. Let $\mathbf{y}$ be the state just after transition. It must be shown that for all $\mathbf{x}, \mathbf{y}$ that

$$\pi(\mathbf{x}) = \pi(\mathbf{x}') \quad \text{if } |\mathbf{x}| = |\mathbf{x}'| \tag{3.2}$$

and

$$\pi(\mathbf{x})A(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y}^+)A(\mathbf{y}^+, \mathbf{x}^+). \tag{3.3}$$

Equation (3.2) follows directly from (3.1). To prove (3.3) consider the possible cases:

(i) $\mathbf{y} = \varphi_l^+(\mathbf{x}) = (s_1, \ldots, s_{l-1}, 1, s_l, \ldots, s_n)$

$$\pi(\mathbf{x})A(\mathbf{x}, \mathbf{y}) = \pi(s_1, \ldots, s_n)A((s_1, \ldots, s_n), (s_1, \ldots, s_{l-1}, 1, s_l, \ldots s_n))$$

$$= B(\frac{\lambda}{k\mu})^n(\prod_{i=1}^{n} \phi(i))^{-1}\lambda\delta(l, n+1). \tag{3.4}$$

$$\pi(\mathbf{y}^+)A(\mathbf{y}^+, \mathbf{x}^+) = \pi(k - s + 1, \ldots, k - s_{l-1} + 1, k, k - s_l + 1, \ldots, k - s_n + 1) \cdot$$

$$A((k-s_1+1, \ldots, k-s_{l-1}+1, k, k-s_l+1, \ldots, k-s_n+1), (k-s_1+1, \ldots, k-s_n+1))$$

$$= B(\frac{\lambda}{k\mu})^{n+1}(\prod_{i=1}^{n+1}\phi(i))^{-1}k\mu\gamma(l,n+1)\phi(n+1)$$

$$= B(\frac{\lambda}{k\mu})^{n}(\prod_{i=1}^{n}\phi(i))^{-1}\lambda\gamma(l,n+1) \tag{3.5}$$

Using the fact that $\delta = \gamma$, we have equality of expressions (3.4) and (3.5).

(ii) $\mathbf{y} = \varphi_l^-(\mathbf{x}) = (s_1, \ldots, s_{l-1}, s_{l+1}, \ldots, s_n)$

$$\pi(\mathbf{x})A(\mathbf{x},\mathbf{y}) = \pi(s_1, \ldots, s_{l-1}, k, s_{l+1}, \ldots, s_n) \cdot$$

$$A((s_1, \ldots, s_{l-1}, k, s_{l+1}, \ldots, s_n), (s_1, \ldots, s_{l-1}, s_{l+1}, \ldots, s_n))$$

$$= B(\frac{\lambda}{k\mu})^{n}(\prod_{i=1}^{n}\phi(i))^{-1}k\mu\gamma(l,n)\phi(n) \tag{3.6}$$

$$\pi(\mathbf{y}^+)A(\mathbf{y}^+,\mathbf{x}^+) = \pi(k-s+1, \ldots, k-s_{l-1}+1, k-s_{l+1}+1, \ldots, k-s_n+1) \cdot$$

$$A((k-s_1+1, \ldots, k-s_{l-1}+1, k-s_{l+1}+1, \ldots, k-s_n+1),$$

$$(k-s_1+1, \ldots, k-s_{l-1}+1, 1, k-s_{l+1}+1, \ldots, k-s_n+1))$$

$$= B(\frac{\lambda}{k\mu})^{n-1}(\prod_{i=1}^{n}\phi(i))^{-1}\lambda\delta(l,n)\phi(n)$$

$$= B(\frac{\lambda}{k\mu})^{n}(\prod_{i=1}^{n}\phi(i))^{-1}k\mu\delta(l,n)\phi(n) \tag{3.7}$$

Again we get equality of (3.6) and (3.7) by virtue of the fact that $\delta = \gamma$.

(iii) $\mathbf{y} = \varphi_l(\mathbf{x}) = (s_1, \ldots, s_l + 1, \ldots, s_n)$

$$\pi(\mathbf{x})A(\mathbf{x},\mathbf{y}) = B(\frac{\lambda}{k\mu})^{n}(\prod_{i=1}^{n}\phi(i))^{-1}k\mu\gamma(l,n)\phi(n) \tag{3.8}$$

$$\pi(\mathbf{y}^+)A(\mathbf{x}^+,\mathbf{y}^+) = \pi(k-s_1+1, \ldots, k-s_l, \ldots, k-s_n+1) \cdot$$

$$A((k-s_1+1, \ldots, k-s_l, \ldots, k-s_n+1), (k-s_1+1, \ldots, k-s_l+1, \ldots, k-s_n+1))$$

$$= B(\frac{\lambda}{k\mu})^{n}(\prod_{i=1}^{n}\phi(i))^{-1}k\mu\gamma(l,n)\phi(n) \tag{3.9}$$

And (3.8) and (3.9) are equal.

(iv) $\mathbf{x} = (0)$ and $\mathbf{y} = (1)$

$$\pi(\mathbf{x})A(\mathbf{x},\mathbf{y}) = B\lambda\delta(1,1) \tag{3.10}$$

$$\pi(\mathbf{y}^+)A(\mathbf{y}^+,\mathbf{x}^+) = \pi((2))A((2),(0))$$

$$= B(\frac{\lambda}{k\mu})(\phi(1))^{-1}k\mu\gamma(1,1)\phi(1)$$

$$= B\lambda\gamma(1,1)\ . \tag{3.11}$$

(3.10) and (3.11) are equal since $\delta = \gamma$. Thus by Theorem 1.14 of Kelly [1979], we have dynamic reversibility of $\{X(t)\}$. •

To see a relatively simple example of $\{X(t)\}$, consider the $M/E_2/1/2$ case with an arbitrary symmetric queue discipline. The state space is

$$\{(0),(1),(2),(1,1),(1,2),(2,1),(2,2)\}$$

and the generator is given by

$$G = \begin{array}{c} \\ (0) \\ (1) \\ (2) \\ (1,1) \\ (1,2) \\ (2,1) \\ (2,2) \end{array} \begin{array}{c} \begin{array}{ccccccc} (0) & (1) & (2) & (1,1) & (1,2) & (2,1) & (2,2) \end{array} \\ \begin{pmatrix} -\lambda & \lambda & & & & & \\ & -(\lambda+2\mu) & 2\mu & \lambda & & & \\ 2\mu & & -(\lambda+2\mu) & & \gamma_{12}\lambda & \gamma_{22}\lambda & \\ & & & -2\mu & \gamma_{22}2\mu & \gamma_{12}2\mu & \\ & \gamma_{22}2\mu & & & -2\mu & & \gamma_{12}2\mu \\ & \gamma_{12}2\mu & & & & -2\mu & \gamma_{22}2\mu \\ & & 2\mu & & & & -2\mu \end{pmatrix} \end{array}$$

Note that since $\delta = \gamma$, the $\gamma's$ could just as well be replaced by $\delta's$.

The stationary distribution for $\{X(t)\}$ is found to be

$$\pi = B\left[1, \frac{\lambda}{2\mu}, \frac{\lambda}{2\mu}, \left(\frac{\lambda}{2\mu}\right)^2, \left(\frac{\lambda}{2\mu}\right)^2, \left(\frac{\lambda}{2\mu}\right)^2, \left(\frac{\lambda}{2\mu}\right)^2\right] \tag{3.12}$$

The marginal queue length distribution is

$$\pi = B\left[1, \frac{\lambda}{\mu}, \left(\frac{\lambda}{\mu}\right)^2\right] . \tag{3.13}$$

It is interesting to note that (3.12) and (3.13) are the same for any queue discipline within the symmetric class of disciplines. Further note that (3.13) is the same for any E_k distribution and in particular it matches the stationary queue length distribution for the exponential server case as seen in Section 3.2.

Since $\{X(t)\}$ is dynamically reversible, then from results in Disney and Kiessler [1987], we have some interesting relationships among the traffic processes.

THEOREM 3.9.

(i) $\{(X_n^{a-}, T_n^a)\}$ *is the dynamic reverse MRP of* $\{(X_n^d, T_n^d)\}$ *and* $\{(X_n^a, T_n^a)\}$ *is the dynamic reverse MRP of* $\{(X_n^{d-}, T_n^d)\}$.

(ii) $\{(X_n^{i-}, T_n^i)\}$ *is the dynamic reverse MRP of* $\{(X_n^o, T_n^o)\}$ *and* $\{(X_n^i, T_n^i)\}$ *is the dynamic reverse MRP of* $\{(X_n^{o-}, T_n^o)\}$.

(iii) $\{(X_n^v, T_n^v)\}$ *is the dynamic reverse MRP of* $\{(X_n^{v-}, T_n^v)\}$.

THEOREM 3.10.

(i) $\pi^{a-} = \pi^d = \pi$ *and* $\pi^a = \pi^{d-}$

(ii) $\pi^{i-} = \pi^o$ *and* $\pi^i = \pi^{o-}$

(iii) $\pi^v = \pi^{v-}$

The next result follows from Theorem 3.10 and equations (3.1) and (3.2).

THEOREM 3.11. *The embedded distributions in Theorem 3.10 have the same insensitivity property as* π. *Moreover, these embedded distributions have the property that any two states* $\mathbf{x}$ *and* $\mathbf{x}'$, *such that* $|\mathbf{x}| = |\mathbf{x}'|$, *have the same embedded probability.*

For an example of a traffic process in a model with a symmetric queue discipline, consider the output process for the $M/E_2/1/2$ - LCFS-PR. The generator for $\{X(t)\}$ is given by

$$
\begin{array}{c}
\\
(0) \\
(1) \\
(2) \\
(1,1) \\
(1,2) \\
(2,1) \\
(2,2)
\end{array}
\begin{array}{c}
\begin{array}{ccccccc}
(0) & (1) & (2) & (1,1) & (1,2) & (2,1) & (2,2)
\end{array} \\
\begin{pmatrix}
-\lambda & \lambda & 0 & 0 & 0 & 0 & 0 \\
0 & -(\lambda+2\mu) & 2\mu & \lambda & 0 & 0 & 0 \\
2\mu & 0 & -(\lambda+2\mu) & 0 & \lambda & 0 & 0 \\
0 & 0 & 0 & -2\mu & 0 & 2\mu & 0 \\
0 & 0 & 0 & 0 & -2\mu & 0 & 2\mu \\
0 & 2\mu & 0 & 0 & 0 & -2\mu & 0 \\
0 & 0 & 2\mu & 0 & 0 & 0 & -2\mu
\end{pmatrix}
\end{array}
$$

The stationary vector is given by

$$\pi = B\left[1, \frac{\lambda}{2\mu}, \frac{\lambda}{2\mu}, \left(\frac{\lambda}{2\mu}\right)^2, \left(\frac{\lambda}{2\mu}\right)^2, \left(\frac{\lambda}{2\mu}\right)^2, \left(\frac{\lambda}{2\mu}\right)^2\right]$$

which also follows from (3.12). The semi-Markov kernel for the MRP $\{(X_n, T_n)\}$ (embedded at all jumps of $X(t)$) is given by

$$
\begin{array}{c} \\ (0) \\ (1) \\ (2) \\ (1,1) \\ (1,2) \\ (2,1) \\ (2,2) \end{array}
\begin{array}{c}
\begin{array}{ccccccc} (0) & (1) & (2) & (1,1) & (1,2) & (2,1) & (2,2) \end{array} \\
\begin{pmatrix}
0 & c(s) & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & b(s) & a(s) & 0 & 0 & 0 \\
b(s) & 0 & 0 & 0 & a(s) & 0 & 0 \\
0 & 0 & 0 & a(s) & 0 & b(s) & 0 \\
0 & 0 & 0 & 0 & a(s) & 0 & b(s) \\
0 & b(s) & 0 & 0 & 0 & a(s) & 0 \\
0 & 0 & b(s) & 0 & 0 & 0 & a(s)
\end{pmatrix}
\end{array}
$$

Consider the output process $\{(X_n^o, T_n^o)\}$. This process is a MRP on state space $\{(0),(1),(2)\}$ since any of these states are possible after an output. The semi-Markov kernel is given by

$$
Q_o(s) = \begin{array}{c} \\ (0) \\ (1) \\ (2) \end{array}
\begin{array}{c}
\begin{array}{ccc} (0) & (1) & (2) \end{array} \\
\begin{pmatrix}
c(s)b^2(s) & \dfrac{c(s)a(s)b^2(s)}{[1-a(s)]^2} & \dfrac{c(s)a(s)b^3(s)}{[1-a(s)]^2} \\
b^2(s) & \dfrac{a(s)b^2(s)}{[1-a(s)]^2} & \dfrac{a(s)b^2(s)}{[1-a(s)]^2} \\
b(s) & 0 & \dfrac{a(s)b^2(s)}{[1-a(s)]^2}
\end{pmatrix}
\end{array}
$$

The stationary vector is

$$
\pi^o = \left[\frac{\mu}{\lambda+\mu}, \frac{\frac{1}{2}\lambda}{\lambda+\mu}, \frac{\frac{1}{2}\lambda}{\lambda+\mu}\right]
$$

Note that $\pi((1)) = \pi((2))$, which is indicated in Theorem 3.5.

The fact that $\pi^o = \pi^{i-}$ would be useful in computing the sojourn time in the case of symmetric queue disciplines. Other than this fact, there seems to be no other particular simplification of the sojourn time problem in the case of symmetric queue disciplines.

3.5 Queue Disciplines Without Preemptions

There are certain advantages in considering disciplines that do not allow preemptions. The main one is a reduction in the size of the state space. Moreover, if the single server case is considered, then the model operates in continuous time as a quasi-birth-death process. Hence, all the known results concerning this type of Markov process would apply. Recall that in the general setting, the generator of $\{X(t)\}$ and the semi-Markov kernel of $\{(X_n, T_n)\}$ both have a tridiagonal block structure, the size of the diagonal block at the nth level (corresponding to a queue length of n) being $k^n \times k^n, \quad 0 < n \leq L$. When no preemptions are allowed then the largest block is $k^S \times k^S$ where S is the number of servers. Hence, in the single server case, all the blocks are $k \times k$ (except for the boundary case at $n = 0$, where the blocks are $1 \times 1, 1 \times k$, and $k \times 1$). Besides the reduction in dimension, there are historical reasons for considering the case of no preemptions. In earlier days of queueing theory, preemptions were not considered when comparing waiting times for various queue disciplines, e.g. Takacs [1963] and Kingman [1961].

Queue disciplines without preemptions have certain disadvantages compared to the earlier examples in this chapter. For one, the queue disciplines are not symmetric, so, unless the service times are exponentially distributed, there is no reversibility or dynamic reversibility of the state space. Hence, the relationships between traffic processes as discussed in Sections 3.3 and 3.4 no longer hold. More-

over, the stationary distributions no longer exhibit product form. One important observation concerning the case of no preemption is that the state process and the various traffic processes are not affected by the specific queue discipline within this class. Sojourn and waiting times, however, can be affected.

In order to use the δ, γ, and ϕ functions to model queue disciplines with no preemption we use the following construction. Designate one position (i.e. position 1) as the service position. A customer remains in this position until service is complete.

$$\gamma(l,N) = \begin{cases} 1 & \text{if } l = 1 \\ 0 & \text{if } l > 1 \end{cases}, \quad 0 < N \leq L.$$

The way in which arriving customers are inserted in the queue is given by the δ function.

For example, for the FCFS discipline,

$$\delta(l,N+1) = \begin{cases} 1 & \text{if } l = N+1 \\ 0 & \text{if } l < N+1 \end{cases}, 0 \leq N < L.$$

For the LCFS (no preemption)

$$\delta(l,N+1) = \begin{cases} 1 & \text{if } l = 2 \\ 0 & \text{if } l \neq 2 \end{cases}, 1 \leq N < L$$

and

$$\delta(1,1) = 1.$$

In other words, for the LCFS discipline, an arriving customer always enters position 2 if the queue is not yet full and a customer is already in service. If the server is idle then the customer enters directly to position 1 and service begins.

Another classic example is that of random order of service. To model this discipline let

$$\delta(l,N+1) = p_l$$

where

$$0 \leq p_l \leq 1, \text{ for all } l \quad \text{and} \quad \sum_{l=2}^{N+1} p_l = 1.$$

At any time t, the state of the system, for single server queues without preemptions, is of the form

$$X(t) = (s(t), 1, 1, \ldots 1)$$

where only one customer is being served and is in phase $s(t)$ of service. All other customers, whether they arrived before or after the customer in service, have not yet received any service and hence are still in stage 1 of service. This is obviously true since a customer receives continuous and unshared service until that customer's service is completed at which time it departs the system.

Since all customers not currently in service are in stage 1, there is nothing in the state vector to distinguish them. When a new arrival is inserted in the queue the state vector would be the same no matter where the customer is inserted. So, the actual queue discipline within the no preemption class is of no importance in studying the $\{X(t)\}$ process. Thus, it reduces the notation to rename the vector $(s(t), 1, 1, \ldots, 1)$ to $(N(t), s(t))$ where $N(t)$ is the total number of customers in the system.

To illustrate how one might take advantage of the special structure afforded by disallowing preemptions, consider the $M/E_k/1/L$ model with no preemptions.

When the states are ordered lexicographically, then the generator of the Markov process $\{X(t)\} = \{(N(t), s(t))\}$ has the following form:

$$A = \begin{pmatrix} A_{00} & A_{01} & & & \\ A_{10} & A_{11} & A_{12} & & \\ & A_{21} & A_{11} & A_{12} & \\ & & A_{21} & A_{11} & A_{12} \\ & & & A_{21} & A_{22} \end{pmatrix}$$

where

$$A_{00} = (-\lambda) \qquad (1 \times 1)$$

$$A_{01} = (\lambda, 0, \ldots, 0) \qquad (1 \times k)$$

$$A_{10} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ k\mu \end{pmatrix} \qquad (k \times 1)$$

$$A_{21} = \begin{array}{c} \\ (N,1) \\ (N,2) \\ \vdots \\ (N,k) \end{array} \begin{array}{c} \begin{array}{cccc} (N-1,1) & (N-1,2) & \ldots & (N-1,k) \end{array} \\ \begin{pmatrix} & & & \\ & & & \\ & & & \\ k\mu & & & \end{pmatrix} \end{array}$$

$$A_{12} = \begin{array}{c} \\ (N,1) \\ (N,2) \\ \vdots \\ (N,k) \end{array} \begin{array}{c} \begin{array}{cccc} (N+1,1) & (N+1,2) & \ldots & (N+1,k) \end{array} \\ \begin{pmatrix} \lambda & & & \\ & \lambda & & \\ & & \ddots & \\ & & & \lambda \end{pmatrix} \end{array}$$

$$A_{11} = \begin{array}{c} \\ (N,1) \\ (N,2) \\ \vdots \\ (N,k-1) \\ (N,k) \end{array} \begin{array}{c} \begin{array}{cccc} (N,1) & (N,2) & \ldots & (N,k) \end{array} \\ \begin{pmatrix} -(\lambda+k\mu) & k\mu & & \\ & & & \\ & & \ddots \quad \ddots & \\ & & & k\mu \\ & & & -(\lambda+k\mu) \end{pmatrix} \end{array}$$

$$A_{22} = \begin{array}{c} \\ (L,1) \\ \vdots \\ (L,k-1) \\ (L,k) \end{array} \begin{array}{c} \begin{array}{cccc} (L,1) & \cdots & (L,k-1) & (L,k) \end{array} \\ \begin{pmatrix} & & & \\ & & & \\ & & -k\mu & k\mu \\ & & & -k\mu \end{pmatrix} \end{array}$$

Several authors have used the special quasi-birth-death structure as an aid for computing the stationary distribution for the state process. See Neuts [1981] for a discussion. However, the particular processes studied have been on an infinite state space (i.e. $L = \infty$).

In this section, a simple matrix solution to the stationary equations will be outlined. This technique was used by Disney [1972]. The steady state equation $\pi A = 0$ can be rewritten as:

$$\begin{aligned} \pi_0 A_{00} + \pi_1 A_{10} &= 0 \\ \pi_0 A_{01} + \pi_1 A_{11} + \pi_2 A_{21} &= 0 \\ \pi_{i-1} A_{12} + \pi_i A_{11} + \pi_{i+1} A_{21} &= 0 \qquad 2 \le i \le L-1 \\ \pi_{L-1} A_{12} + \pi_L A_{LL} &= 0 \end{aligned} \tag{3.14}$$

where

$$\pi_0 = \pi(0)$$

and

$$\pi_i = (\pi(i,s))_{s=1,2,\ldots,k}$$

Equation (3.14) is a second order difference equation with the other three equations providing boundary conditions. The solution to this boundary value problem can then be approached in a manner analogous to a method found in many elementary

texts of differential equations. Write (3.14) as a system of two first order difference equations

$$\pi_{i-1}A_{12} = -\pi_i A_{11} - \pi_{i+1}A_{21}$$

$$\pi_i = \pi_i$$

Noting that $A_{12}^{-1} = \lambda^{-1}I$ then

$$\pi_{i-1} = -\pi_i\lambda^{-1}A_{11} - \pi_{i+1}\lambda^{-1}A_{21}$$

$$\pi_i = \pi_i$$

or in matrix form

$$(\pi_{i-1}, \pi_i) = (\pi_i, \pi_{i+1}) \begin{pmatrix} -\lambda^{-1}A_{11} & I \\ -\lambda^{-1}A_{21} & 0 \end{pmatrix}$$

Thus the solution of this first order equation has the form

$$\mathbf{z}_i = \mathbf{z}_{i+1}B^{L-i}$$

where $z_i = (\pi_{i-1}, \pi_i)$ and $2 \leq i \leq L-1$. In order to find the π_i vectors from the $\mathbf{z}_i$ vectors and to find the boundary values π_0 and π_L, we use the boundary equations and the fact that $\pi\mathbf{u} = 1$. Details will be omitted.

The main point to be made from the previous discussion is that the quasi-birth-death structure, in the case of no preemption, provides several ready made tools for computing the stationary state distribution.

The traffic processes $\{(X_n^r, T_n^r)\}$ for the M/E$_k$/1/L model are not affected by the actual queue discipline within the class of disciplines without preemption. This fact follows directly from the fact that $\{X(t)\}$ has this same invariance property.

One complicating factor for this class of disciplines is that we lose the dynamic reversibility between the arrival and departure processes, between the input and

output processes and between the overflow process with itself. This can be checked in the following example.

To illustrate some of the above mentioned points, consider the $M/E_2/1/2$ example under, say a FCFS queue discipline. First, consider the continuous time process on the former version of the state space. That is, let $X(t) = (s_1(t), s_2(t))$ where $s_i(t)$ is the phase of the customer in position i. The state space is thus

$$\{(0), (1), (2), (1,1), (1,2)^*, (2,1), (2,2)^*\}$$

where the states marked with an asterisk are not possible and thus could be omitted. Consider the new version of the state space. That is, let $X(t) = (N(t), s(t))$, then the state space is

$$\{(0), (1,1), (1,2), (2,1), (2,2)\}$$

where the order of the states is the same as the corresponding order in the former representation with the impossible states omitted.

Since the $M/E_2/1/2$ FCFS example has a relatively small state space in both versions and for ease in comparing with earlier examples, the former state space will be used. The generator of $X(t) = (s_1(t), s_2(t))$ is given by:

$$\begin{array}{c} \\ (0) \\ (1) \\ (2) \\ (1,1) \\ (1,2)^* \\ (2,1) \\ (2,2)^* \end{array} \begin{array}{c} \begin{array}{ccccccc} (0) & (1) & (2) & (1,1) & (1,2)^* & (2,1) & (2,2)^* \end{array} \\ \begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 & 0 & 0 \\ 0 & -(\lambda+2\mu) & 2\mu & \lambda & 0 & 0 & 0 \\ 2\mu & 0 & -(\lambda+2\mu) & 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & -2\mu & 0 & 2\mu & 0 \\ 0 & 0 & 0 & 0 & -2\mu & 0 & 2\mu \\ 0 & 2\mu & 0 & 0 & 0 & -2\mu & 0 \\ 0 & 0 & 2\mu & 0 & 0 & 0 & -2\mu \end{pmatrix} \end{array}$$

The stationary distribution is given by

$$\pi = B\Big(1, \frac{\lambda(\lambda+2\mu)}{(2\mu)^2}, \frac{\lambda}{2\mu}, \frac{\lambda^2(\lambda+2\mu)}{(2\mu)^2}, 0, \frac{\lambda^2(\lambda+4\mu)}{(2\mu)^3}, 0\Big)$$

The semi-Markov kernel for the MRP $\{(X_n, T_n)\}$ (embedded at all jumps of $X(t)$) is given by

$$\begin{array}{c} \\ (0) \\ (1) \\ (2) \\ (1,1) \\ (1,2)^* \\ (2,1) \\ (2,2)^* \end{array} \begin{array}{c} \begin{array}{ccccccc} (0) & (1) & (2) & (1,1) & (1,2)^* & (2,1) & (2,2)^* \end{array} \\ \begin{pmatrix} 0 & c(s) & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & b(s) & a(s) & 0 & 0 & 0 \\ b(s) & 0 & 0 & 0 & 0 & a(s) & 0 \\ 0 & 0 & 0 & a(s) & 0 & b(s) & 0 \\ 0 & 0 & 0 & 0 & a(s) & 0 & b(s) \\ 0 & b(s) & 0 & 0 & 0 & a(s) & 0 \\ 0 & 0 & b(s) & 0 & 0 & 0 & a(s) \end{pmatrix} \end{array}$$

where

$$c(s) = \frac{\lambda}{\lambda+s} \qquad a(s) = \frac{\lambda}{\lambda+2\mu+s} \qquad b(s) = \frac{2\mu}{\lambda+2\mu+s}$$

where states marked by $*$ can be deleted so the kernel reduces to

$$\begin{array}{c} \\ (0) \\ (1) \\ (2) \\ (1,1) \\ (2,1) \end{array} \begin{array}{c} \begin{array}{ccccc} (0) & (1) & (2) & (1,1) & (2,1) \end{array} \\ \begin{pmatrix} 0 & c(s) & 0 & 0 & 0 \\ 0 & 0 & b(s) & a(s) & 0 \\ b(s) & 0 & 0 & 0 & a(s) \\ 0 & 0 & 0 & a(s) & b(s) \\ 0 & b(s) & 0 & 0 & a(s) \end{pmatrix} \end{array}$$

Consider, for example, the output process, $\{(X_n^o, T_n^o)\}$. The semi-Markov kernel is given by

$$\widehat{Q}_o(s) = \sum_{k=0}^{\infty} \widehat{U}_o^k(s)\widehat{V}_o(s) = (I - \widehat{U}_o(s))^{-1}\widehat{V}_o(s).$$

Just after an output the system is in state (0) or (1). Hence we need to consider only that portion of $\widehat{Q}_o(s)$.

$$I - \widehat{U}_o(s) = \begin{pmatrix} 1 & \frac{-\lambda}{\lambda+s} & 0 & 0 & 0 \\ 0 & 1 & \frac{-2\mu}{\lambda+2\mu+s} & \frac{-\lambda}{\lambda+2\mu+s} & 0 \\ 0 & 0 & 1 & 0 & \frac{-\lambda}{\lambda+2\mu_s} \\ 0 & 0 & 0 & 1-\frac{\lambda}{\lambda+2\mu+s} & \frac{-2\mu}{\lambda+2\mu+s} \\ 0 & 0 & 0 & 0 & 1-\frac{\lambda}{\lambda+2\mu+s} \end{pmatrix}$$

$$\widehat{V}_o(s) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \frac{2\mu}{\lambda+2\mu+s} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{2\mu}{\lambda+2\mu+s} & 0 & 0 & 0 \end{pmatrix}$$

Note that because of the sparseness of $\widehat{V}_o(s)$, only certain entries of $[I-\widehat{U}_o(s)]^{-1}$ need to be computed.

Performing the appropriate computations yields the following kernel for the output process:

$$\widehat{Q}_o(s) = \begin{array}{c} \\ (0) \\ (1) \end{array} \begin{array}{c} \begin{array}{cc} (0) & \qquad\qquad (1) \end{array} \\ \begin{pmatrix} c(s)(b(s))^2 & \frac{c(s)a(s)(b(s))^2}{[1-a(s)]^2} + \frac{c(s)a(s)b(s))^2}{1-a(s)} \\ (b(s))^2 & \frac{a(s)(b(s))^2}{[1-a(s)]^2} + \frac{a(s)(b(s))^2}{1-a(s)} \end{pmatrix} \end{array}$$

Note that the state space of this MRP is $\{(0),(1)\}$ since these are the only states possible just after an output.

The embedded Markov chain has a transition matrix given by

$$\widehat{Q}_o(0) = \begin{pmatrix} b^2 & 1-b^2 \\ b^2 & 1-b^2 \end{pmatrix}$$

where $b = 2\mu/(\lambda + 2\mu)$.

So the steady state vector satisfying

$$\pi^o \widehat{Q}_o(0) = \pi^o; \qquad \pi^o \mathbf{u} = 1$$

is given by

$$\pi^o = [b^2, 1-b^2]$$

Now consider the overflow process $\{(X_n^v, T_n^v)\}$ from the $M/E_2/1/2-FCFS$ queue. We have the kernel,

$$\widehat{Q}_v(s) = \sum_{k=0}^{\infty} \widehat{U}_v^k(s)\widehat{V}_v(s) = (I - \widehat{U}_v(s))^{-1}\widehat{V}_v(s)$$

where

$$I - \widehat{U}_v(s) = \begin{array}{c} \\ (0) \\ (1) \\ (2) \\ (1,1) \\ (2,1) \end{array} \begin{array}{c} \begin{array}{ccccc} (0) & (1) & (2) & (1,1) & (2,1) \end{array} \\ \begin{pmatrix} 1 & -c(s) & 0 & 0 & 0 \\ 0 & 1 & -b(s) & -a(s) & 0 \\ -b(s) & 0 & 1 & 0 & -a(s) \\ 0 & 0 & 0 & 1 & -b(s) \\ 0 & -b(s) & 0 & 0 & 1 \end{pmatrix} \end{array}$$

and

$$\widehat{V}_v(s) = \begin{array}{c} \\ (0) \\ (1) \\ (2) \\ (1,1) \\ (2,1) \end{array} \begin{array}{c} \begin{array}{ccccc} (0) & (1) & (2) & (1,1) & (2,1) \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & a(s) & 0 \\ 0 & 0 & 0 & 0 & a(s) \end{pmatrix} \end{array}$$

$\widehat{Q}_v(s)$ could be computed to give the kernel of the overflow process on the state space $\{(1,1),(2,1)\}$. A recursive procedure has not been developed for computing $\widehat{Q}_v(s)$ in general, as was done for the exponential server. However, the nature of the overflow process is now better understood.

We now turn our attention to the sojourn time for the single server queue and the queue disciplines without preemptions . In order to study the sojourn time for this class of models, the continuous time state process, $\{X(t)\} = \{(N(t), s(t))\}$, will be used. Applying the general results of Section 2.4, we define

$$W(t) = (X(t), l(t)) = (N(t), s(t), l(t))$$

Recall that $l(t)$ is the position in the queue of the tagged customer C. Now C can move around in the queue before his service commences. But, once he reaches position 1, he remains there until his service is complete. This last comment and the fact that there is only one server are the reasons that this example is simpler to analyze than the general sojourn time problem. It is convenient to define

$$\Delta' = \{\mathbf{x} : \mathbf{x} = (n, 1, 1)\}$$

That is, Δ' is the collection of states of $\{W(t)\}$ that C enters when the server is first reached. Thus, the first passage time of $W(t)$ to reach Δ' is actually the *waiting* time of C. Once the distribution for the waiting time is known, then the sojourn time distribution can be found by convolving the waiting time distribution with the service time distribution. In a manner analogous to Section 3.3, define the conditional waiting time by

$$S'_{W_0} = inf\Big\{t : W(t) = \Delta'; W(0+) = W_0\Big\}$$

and the waiting time by

$$S' = inf\Big\{t : W(t) = \Delta'.\Big\}$$

Note that here we define the waiting time only for the customer that does not overflow. The results of Section 2.4 can now be readily adapted to find the waiting time distribution.

For purposes of illustration consider the $M/E_2/1/3$ case with FCFS. The generator for $\{W(t)\}$ can be found by applying Theorem 2.16. To compute, for example, the conditional waiting time distribution S'_{W_0} where $W_0 = (3,2,3)$, we could adapt 2.17 or apply the graphical technique to find $\widehat{F}_{W_0}(s)$, the Laplace-Stieltjes transform of $F_{W_0}(t)$, the conditional waiting time distribution:

$$\Big[\frac{2\mu}{2\mu+s}\Big]\Big[(\frac{2\mu}{2\mu+s})^2(\frac{\lambda}{\lambda+2\mu+s}) + (\frac{2\mu}{2\mu+s})(\frac{\lambda}{\lambda+2\mu+s})(\frac{2\mu}{\lambda+2\mu+s}) + (\frac{2\mu}{\lambda+2\mu+s})^2\Big]$$

which simplifies to

$$\Big[\frac{2\mu}{2\mu+s}\Big]^3.$$

This result is obvious since C must wait for the customer in service, at the time of the arrival of C, to finish the second phase of service plus the next customer in line to receive two phases of service . Therefore, the waiting time of C is the convolution of three exponential distributions of rate 2μ.

Chapter 4

Extensions and Generalizations

4.1 Introduction

One of the essential aspects of the $M/E_k/\phi/L$ class of models is the fact that these models operate in continuous time as Markov processes on finite state spaces. Various extensions to this class of models can be considered without losing this property. The methodologies of this dissertation can be adapted to handle these extensions. However, many of these extensions require that the state space become larger and more complicated. In this chapter some of these extensions will be discussed. Many of the details will be omitted since usually the details are just a rehashing of Chapters 2 and 3 and add little to our understanding of the theory. Many of the questions that arise in this chapter will be left for future research.

4.2 Balking

One interesting extension of the $M/E_k/\phi/L$ class of models provides a generalization of the parameter L. Consider the parameter L as follows. Suppose that there are N customers in the queue just before the arrival of a customer. That customer will join the queue with probability $\beta(N) = 1$ if $N < L$ and with probability $\beta(N) = 0$ if $N = L$. We could just as well consider a more arbitrary function, β. The function β could then be used to model balking, a feature which is common in many real world systems. Hence, a customer who finds N customers in the system upon arrival will join the queue with probability $\beta(N)$ and will depart without service (balk) with probability $1 - \beta(N)$. In order to keep a finite state space, assume that $\beta(N) = 0$ for $N \geq L$. That is, the queue capacity remains finite. This balking system could be referred to using the notation $M/E_k/\phi/\beta_L$. Very

few changes to the theory of Chapter 2 are needed in order to study this extended model.

Let

$$Y(t) = (s_1(t), s_2(t), \ldots, s_N(t), f(t))$$

where $s_i(t) \in \{1, 2 \ldots, k\}$ is the phase of service of the customer in position i at time t. But, $f(t) \in \{0, 1\}$ now changes its value whenever a balk occurs. This flip-flop variable may change from any state rather than just from states corresponding to a full capacity. Most of the results of Chapter 2 go through with little alteration. For example, we get

THEOREM 4.1. *$\{Y(t)\}$ is a Markov process.*

PROOF: The proof is an exact replica of the proof of Theorem 2.12. •

The generator of $\{Y(t)\}$ can be derived by a small change in Theorem 2.4. Let $\mathbf{y} = (s_1, s_2, \ldots, s_N, f)$ be a state of the system. Interpret $\varphi_f(\mathbf{y}) = (s_1, s_2, \ldots, s_N, 1-f)$ as the new state that occurs when there is a balk.

THEOREM 4.2. *The generator, A, for the Markov process $\{Y(t)\}$ is given by:*

(i) $A(\mathbf{y}, \varphi_i^+(\mathbf{y})) = \delta(i, N+1)\beta(N)\lambda \quad \text{if} \quad N < L;$

(ii) $A(\mathbf{y}, \varphi_i^-(\mathbf{y})) = \gamma(i, N)\phi(N)k\mu \quad \text{if} \quad s_i = k, \ N \geq 1;$

(iii) $A(\mathbf{y}, \varphi_i(\mathbf{y})) = \gamma(i, N)\phi(N)k\mu \quad \text{if} \quad s_i < k, \ N \geq 1;$

(iv) $A(\mathbf{y}, \varphi_f(\mathbf{y})) = (1 - \beta(N))\lambda \quad \text{if} \quad N \leq L;$

(v) $A(\mathbf{y}, \mathbf{y}) = -\lambda \quad \text{if} \quad N = 0;$

(vi) $A(\mathbf{y}, \mathbf{y}) = -(\lambda + \phi(N)k\mu) \quad \text{if} \quad N > 0.$

PROOF: Only part (i) and (iv) are different from the corresponding parts of Theorem 2.3. For part (i) there are two subcases to consider: $N = 0$ and $0 < N \leq L$. If $N = 0$, then $\mathbf{y} = (0, f(t))$ and the exponential holding time parameter is λ. In this case $P(\mathbf{y}, \varphi_i^+(\mathbf{y})) = \delta(1, 1)\beta(0) = \beta(0)$ which is the probability that the arrival

enters the system in position 1. If $0 < N \leq L$, then the exponential holding time parameter is $\lambda + \phi(N)k\mu$. Furthermore, $P(\mathbf{y}, \varphi_i^+(\mathbf{y})) = \delta(i, N+1) \cdot \beta(N) \cdot \lambda/(\lambda + \phi(N)k\mu)$ which is the probability that the transition is an arrival and not a phase change (which is $\lambda/(\lambda + \phi(N)k\mu)$ times the probability that given the transition is an arrival that the arriving customer enters the queue (which is $\beta(N)$) times the probability that given the arriving customer joins the queue that the customer enters position i (which is $\delta(i, N+1)$). Therefore, multiplying the transition probability by the holding time parameter we get (i).

Part (iv) can be verified by a similar argument. •

Consider the Markov renewal process $\{(Y_n, T_n)\}$ obtained by embedding $\{Y(t)\}$ at transition epochs. The semi-Markov kernel has the same structure as the one given in Theorem 2.5. Moreover, just as in Chapter 2, $\{(Y_n, T_n)\}$ is lumpable to $\{(X_n, T_n)\}$ which eliminates the flip-flop component from the state space.

The traffic processes filtered from $\{(X_n, T_n)\}$ for the balking model are defined just as in Section 2.3. An interesting new traffic process to be investigated is the balking process. To study this process one needs only to re-interpret the overflow process. Definition 2.9 (iii) defined the overflow traffic set $B_v = \{(\mathbf{x}, \mathbf{y}) | \mathbf{x} = \mathbf{y}\}$. Under the new interpretation B_v becomes the balking traffic set. The traffic process $\{(X_n^v, T_n^v)\}$ now could be called the balking process.

The sojourn time as discussed in Section 2.4 remains valid with little change. Each λ in Theorems 2.16 and 2.17 should be replaced by $\beta(N)\lambda$. Note that a balking customer like the overflow customer has a sojourn time of 0.

The special cases of Chapter 3 are easily adapted to accommodate balking. The exponential server case of section 3.2 is especially interesting. The $M/M/\phi/\beta_L$ model becomes the general finite birth-death model. Unlike the $M/M/\phi/L$ model which had a constant birth rate λ, the extended model has a birth-rate dependent

on the queue length. This birth rate will be denoted by $\lambda_n = \lambda\beta(n)$. The generator for $\{Y(t)\}$ becomes

$$
\begin{array}{c}
\\
(0,0) \\
(1,0) \\
\vdots \\
(L-1,0) \\
(L,0) \\
(0,1) \\
(1,1) \\
\vdots \\
(L-1,1) \\
(L,1)
\end{array}
\begin{array}{c}
\begin{array}{cccccccccc}
(0,0) & (1,0) & \dots & (L-1,0) & (L,0) & (0,1) & (1,1) & \dots & (L-1,1) & (L,1)
\end{array} \\
\begin{pmatrix}
-\lambda_0 & \lambda_0 & & & & & & & & \\
\mu_1 & -\lambda_1-\mu_1 & & & & & & & & \\
 & \ddots & \ddots & \ddots & & & & & & \\
 & & & & \lambda_{L-1} & & & & & \\
 & & & \mu_L & -\lambda_L-\mu_L & & & & & \lambda_L \\
 & & & & & -\lambda_0 & \lambda_0 & & & \\
 & & & & & \mu_1 & -\lambda_1-\mu_1 & & & \\
 & & & & & & \ddots & \ddots & \ddots & \\
 & & & & & & & & & \lambda_{L-1} \\
 & & & & \lambda_L & & & & \mu_L & -\lambda_L-\mu_L
\end{pmatrix}
\end{array}
$$

where $\mu_n = \phi(n)\mu$. The generator for $\{X(t)\}$ becomes

$$
\begin{array}{c}
\\
0 \\
1 \\
2 \\
\vdots \\
L-1 \\
L
\end{array}
\begin{array}{c}
\begin{array}{ccccccc}
0 & 1 & 2 & & \dots & L-1 & L
\end{array} \\
\begin{pmatrix}
-\lambda_0 & \lambda_0 & & & & & \\
\mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & & & & \\
 & \mu_2 & -(\lambda_2+\mu_2) & \ddots & & & \\
 & & \ddots & \ddots & & & \\
 & & & & & -(\lambda_{L-1}+\mu_{L-1}) & \lambda_{L-1} \\
 & & & & & \mu_L & -\mu_L
\end{pmatrix}
\end{array}
$$

The stationary distribution for $\{X(t)\}$ becomes

$$\pi = B\Big(1, \frac{\lambda\beta(0)}{\mu\phi(1)}, \frac{\lambda^2\beta(0)\beta(1)}{\mu^2\phi(1)\phi(2)}, \cdots \frac{\lambda^L \prod_i \beta(i-1)}{\mu^L \prod_i \phi(i)}\Big)$$

where B is the normalizing constant.

The discussion on reversibility in Section 3.3 remains valid for the M/M/ϕ/β_L model. In particular, Theorems 3.2, 3.3, and 3.4 remain true.

The stationary distributions for the various embedded Markov chains can be computed as for Theorem 3.5 using the formula on page 90 of Disney and Kiessler [1987].

THEOREM 4.3.

(i) $\pi^{a-} = \pi^d = \pi$;

(ii) $\pi^a = \pi^o = B^o[1, \pi(2)\phi(2)\mu, \pi(3)\phi(3)\mu, \ldots, \pi(L)\phi(L)\mu]$;

(iii) $\pi^v = \pi^{r-} = B^v[1, \pi(1)(\beta(1)\lambda + \phi(1)\mu), \ldots, \pi(L)(\beta(L)\lambda + \phi(L)\mu)]$;

(iv) $\pi^i = \pi^{o-} = B^i[0, \pi(0)\beta(0)\lambda, \ldots, \pi(L-1)\beta(L-1)\lambda]$;

(v) $\pi^a = \pi^{d-}$

$= B^a[0, \pi(0)\beta(0)\lambda, \ldots, \pi(L-2)\beta(L-2)\lambda, \pi(L-1)\beta(L-1)\lambda + \pi(L)(1-\beta(L))\lambda]$.

The semi-Markov kernels for the traffic processes in Section 3.2 are essentially the same . However, we must redefine the notation as follows:

$$a_0(s) = \frac{\beta(0)\lambda}{\beta(0)\lambda + s};$$

$$a_n(s) = \frac{\beta(n)\lambda}{\beta(n)\lambda + \phi(n)\mu + s} \qquad 0 < n \leq L;$$

$$b_n(s) = \frac{\phi(n)\mu}{\beta(n)\lambda + \phi(n)\mu + s} \qquad 0 < n \leq L.$$

The sojourn time discussion of Section 3.3 remains valid if the λ's in the kernel of $\widehat{R}(s)$ are replaced by $\lambda\beta(N)$.

Consider the $M/E_k/\phi/\beta_L$ model with symmetric queue discipline ($\delta = \gamma$). Kelly [1979] does not discuss this model; however, his results can be extended to include it. For example, our Theorem 3.6 can be updated to get the stationary distribution for the continuous time state process for the $M/E_k/\phi/\beta_L$ model:

$$\pi(\mathbf{x}) = B(\frac{\lambda}{k\mu})^n \prod_{i=1}^{n} \beta(i-1)\Big(\prod_{i=1}^{n} \phi(i)\Big)^{-1} \quad \text{if} \quad |\mathbf{x}| = n$$

where B is the normalizing constant. Again, we get the insensitivity property of Theorem 3.7 since

$$P[|\mathbf{x}| = n] \; = B(\frac{\lambda}{\mu})^n \prod_{i=1}^{n} \beta(i-1)\Big(\prod_{i=1}^{n} \phi(i)\Big)^{-1}$$

is independent of the parameter k.

The dynamic reversibility of $\{X(t)\}$ follows by a proof analogous to Theorem 3.8. Hence, the relationships among the traffic processes given in Theorems 3.3 and 3.4 follow.

4.3 Other Extensions

As noted in the previous section, it is possible to analyze extensions to the basic model of Chapter 3 using the same basic methodology. In the basic model it was assumed that the service time is Erlang-k distributed. More general phase type distributions could be used. For example, the service time distribution could be made a mixture of Erlang distributions. That is, one might envision several parallel channels. Channel i consists of k_i phases. Each phase of the channel i is exponentially distributed with parameter $k_i\mu_i$. When a customer enters the server then the customer chooses one of the channels randomly. When all of the phases in that channel have been completed then the service of the customer is finished. The state space of $\{Y(t)\}$ must be made larger. It is now necessary to include in the

state description the channel and the phase within the channel for each customer in the system.

Neuts [1981] has developed an interesting generalization of the previously described mixture of Erlangs. Suppose that there are m phases each of which lasts for an exponentially distributed amount of time with parameter dependent on the phase. A customer entering service would traverse through the phases randomly according to a Markov chain until some dummy absorbing phase is reached, at which time the customer's service is finished. Such a service time distribution is said to be of phase type and is denoted by PH. Essentially, a phase type distribution is the time to absorption in a finite state Markov process with an absorbing state. The LS-transform of such a distribution is rational just as for the Erlang and mixtures of Erlangs previously described. One interesting feature of the PH class of distributions is that they are dense in the set of all probability distributions on $[0, \infty)$. This fact makes the PH distribution useful in modelling for purposes of analysis and simulation of queueing systems. It seems that the general results of Chapters 2 and 3 can be adapted to accommodate service times which are PH distributed since that the finiteness and Markovianess are maintained. The details will be left for future research.

Until now, it has been assumed that the arrival process is Poisson. The methods of Chapter 2 can be adapted to allow for more general arrival processes. For example, it might be assumed that the interarrival times are independent and identically Erlang distributed. Consider for example an $E_m/E_k/\phi/L$ system. One could think of the time between arrivals as consisting of m independent exponential phases each with parameter $m\lambda$. In this case the interarrival time is Erlang-m distributed with a mean interarrival time of $1/\lambda$. In addition to the customers in the queue at any given time, one could envision another customer in the process

of arriving in some phase of arriving; however, the arrival does not actually occur until this arriving customer completes all m phases. With this interpretation, we need to append an extra component to the state description which marks the progress of the arriving customer. Thus,

$$Y(t) = (s_a(t), s_1(t), s_2(t), \ldots, s_N(t), f(t))$$

describes the state of the system at time t where $s_a(t) \in \{1, \ldots, m\}$ is the phase of the arriving customer and $s_i(t) \in \{1, \ldots, k\}$, $i = 1, \ldots, N$ is the phase of service of the customer in position i and $f(t)$ is the flip-flop variable which changes value when an overflow occurs. The necessary adaptations to Chapter 2 will not be developed herein, but will be left for future research.

It has been seen that additions could be made to the basic model of Chapter 2. The main problem with such extensions is that the state space of the system increases in dimensionality. This makes computation more tedious; however, many of the basic qualitative results are preserved.

The results in this dissertation open up many questions that need to be explored.

4.4 Conclusions

In this section the main results will be reviewed. The specific contributions of this dissertation will be highlighted. Furthermore, some of the open problems will be discussed. Although the primary emphasis of this research has been the $M/E_k/\phi/L$ class, it is evident that the theory can be adapted to include other extensions as long as the system operates as a Markov process on a finite state space.

The three major aspects of these models which have explored are:

1. The continuous time state processes
2. The embedded traffic processes
3. The sojourn and waiting time distributions.

One could argue after considering the literature of the field that these are the three most important aspects of queueing systems. Perhaps the main contribution of this dissertation is the fact that the results are provided for a general class of models which contains as special cases so many of the examples that have been historically studied in isolation.

The key assumptions are the Markovian property and the finiteness. The Markovianess of the state process is accomplished by the method of phases. Since all distributions on $[0, \infty)$ can be approximated by phase type distributions, this assumption is not necessarily a practical limitation. However, it would be of theoretical interest to investigate how the results could be extended to GI distributions, both for the service times as well as the inter-arrival times. Admittedly, the finiteness assumption is a convenience since it removes the need to invoke the theory of infinite dimensional matrices. However, it also happens to be the case that finite capacity queues are in need of study. In fact, historically, the finite capacity case largely has been avoided. Often, the assumption of an infinite capacity makes analysis more tractable. In such cases the results must be considered approximations for any real world system which surely must be of finite capacity.

The continuous time state processes are finite state Markov processes. The generators are presented for the general $M/E_k/\phi/L$ models of Chapter 2 and for the special cases of Chapter 3, as well as some of the extensions in Chapter 4. The stationary distributions are computed only for certain special cases including the birth-death type model and for the case of the symmetric queue disciplines. Most of these results follow from well known results on Markov processes, including Kelly

[1979]. The problem of finding the stationary distribution in the general problem is in need of further research, although, there is a long history of work on this problem.

The traffic processes in our class of models have been characterized using Markov renewal processes. Much of the underlying theory for this approach is due to Disney and Kiessler [1987]. This dissertation represents one of the first major applications of their theory of traffic processes. The results on the overflow process provide an important contribution to the field. In most prior research it has been assumed that the service times are exponentially distributed from which it follows that the overflow process is renewal. This dissertation appears to be the first work to fully characterize the overflow process when the service times are not exponentially distributed. There seems to be no publications that study the balking process.

The sojourn time results may be among the more important contributions of this dissertation. Although others have recognized the sojourn time and waiting time as first passage times, the Markov renewal approach appears to be novel. By using the δ, γ and ϕ functions, the sojourn time distribution was determined in a unified fashion for a large class of queue disciplines, both for single and multiple servers. Further research is needed to see how this unified theory can be exploited for practical purposes, such as for comparing sojourn times for various particular disciplines.

Many of the results could no doubt be extended to networks containing more than one queueing node. Perhaps the customers could be assigned types which could be used to assign priorities of service. The possibilities seem almost limitless. However, it must be remembered that the more elaborate the model the greater the curse of dimensionality. Even though theoretical results are possible for the

more complicated models, there are practical computational limits that may be exceeded.

APPENDIX

The following discussion is intended to be a brief introduction to the Markov renewal process. More extensive treatments can be found elsewhere; e.g. Çinlar [1969].

The structure of the Markov renewal process makes it very useful in many diverse areas where stochastic systems are modelled, including queueing theory. The basic character of such a system is that the system moves from state to state at discrete times and the duration of time between state changes is random with a distribution dependent on the current state and the next state.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. A Markov renewal process is a certain bivariate Markov chain $\{(X_n, T_n); \quad n = 0, 1, \ldots\}$ on $(\Omega, \mathcal{F}, P)$ with state space $E \times \Re^+$. For our purposes, E is countable; but, in general, this restriction can be relaxed. Assume that

$$0 = T_0(\omega) < T_1(\omega) < \ldots \qquad a.s.$$

$$\text{and} \qquad \lim_{n\to\infty} T_n(\omega) = \infty \qquad a.s.$$

From the Markov structure we have

$$P[X_{n+1} = j, T_{n+1} \leq t \mid (X_0, T_0), (X_1, T_1), \ldots, (X_n, T_n)]$$

$$= P[X_{n+1} = j, T_{n+1} \leq t \mid (X_n, T_n)]. \qquad (A1)$$

From (A1) it follows that

$$P[X_{n+1} = j, T_{n+1} - T_n \leq t \mid (X_0, T_0), \ldots, (X_n, T_n)]$$

$$= P[X_{n+1} = j, T_{n+1} - T_n \leq t \mid X_n] \qquad (A2)$$

We further assume that the right hand side of (A2) is independent of n, so that the process is time homogeneous. Thus, define

$$A(i, j, t) = P[X_{n+1} = j, T_{n+1} - T_n \leq t \mid X_n = i]$$

The matrix

$$A(t) = [A(i,j,t)]_{i,j\in E}$$

is called the semi-Markov kernel of the Markov renewal process $\{(X_n, T_n)\}$. It is well known that the stochastic behavior of a Markov chain is completely determined by the initial state distribution and the transition functions. Since $T_0 = 0$ a.s., we simply need the initial distribution ν of X_0 and the semi-Markov kernel $A(t)$ to completely determine the stochastic behavior of a Markov renewal process.

Note that $A(i,j,t)$ is a (possibly) defective probability distribution and

$$\lim_{t\to\infty} P[X_{n+1} = j, T_{n+1} - T_n \le t \mid X_n = i]$$

$$= P[X_{n+1} = j \mid X_n = i] = P(i,j)$$

defines a transition matrix for the Markov chain $\{X_n\}$, called the embedded Markov chain. That is, if the successive states of a Markov renewal process are observed without regard to the time intervals between the state changes, then the succession of states, $\{X_n\}$ is a Markov chain. If the initial state distribution ν of $\{X_n\}$ is the stationary distribution (i.e. $\nu P = \nu$ and $\nu \mathbf{u} = 1$), then $P[X_n = i] = \nu(i)$ is independent of n.

Since

$$P[X_{n+1} = j, T_{n+1} - T_n \le t \mid X_n = i]$$

$$= P[T_{n+1} - T_n \le t \mid X_n = i, X_{n+1} = j]\ P[X_{n+1} = j \mid X_n = i]$$

$$= P[T_{n+1} - T_n \le t \mid X_n = i, X_{n+1} = j]\ P(i,j).$$

then it can be shown that

$$P[T_1 - T_0 \le t_1, \ldots, T_{n+1} - T_n \le t_{n+1} \mid X_0, X_1 \ldots, X_{n+1}]$$

$$= P[T_1 - T_0 \le t \mid X_0, X_1]\ P[T_2 - T_1 \le t_2 \mid X_1, X_2] \ldots P[T_{n+1} - T_n \le T_{n+1} \mid X_n, X_{n+1}].$$

This result means that the time increments $\{T_{n+1} - T_n\}$ are conditionally independent given the Markov chain $\{X_n\}$. As an important special case, if the state space, E, consists of a single state then the Markov renewal process simply reduces to a renewal process. If

$$P[T_{n+1} - T_n \leq t \mid X_n = i, X_{n+1} = j]$$

$$= P[T_{n+1} - T_n \leq t \mid X_n = i] = 1 - e^{-\mu(i)t}$$

where $\mu(i) > 0$, then, in this special case, the Markov renewal process is in fact a Markov process. So, Markov renewal theory contains as special cases two of the most important classes of stochastic processes used in operations research: renewal processes and Markov process.

BIBLIOGRAPHY

Berman, M. and Westcott, M. 1983. On queueing systems with renewal departure processes. *Adv. Appl. Prob.* 15: 657-673.

Burke, P.J. 1956. The output of queueing system. *Oper. Res.* 4:699-704.

Burke, P.J. 1958. The output process of a stationary $M/M/s$ queueing system. *Ann. Math. Statist.* 39: 1144-1152.

Branford, A.J. 1986. On a property of finite-state birth and death processes. *J. Appl. Prob.* 23: 859-866.

Çinlar, E. 1969. Markov renewal theory. *J. Appl. Prob.* 1: 123-187.

Çinlar, E. and Disney, R.L. 1967. Streams of overflows from a finite queue. *Oper. Res.* 15: 131-134.

Coffman, E.G., Muntz, R.R. and Trotter, H. 1970. Waiting time distributions for processor sharing systems. *J. Assoc. Comput. Mach.* 17: 123-130.

Daley, D.J. 1976. Queueing output processes. *Adv. Appl. Prob.* 8: 395-412

Disney, R.L. and deMorais, P.R. 1976. Covariance properties for the departure process of $M/E_k/1/N$ queues. *A.I.I.E. Trans.* 8:169-175.

Disney, R.L.; Farrell, R.L.; and deMorais, P.R. 1973. A characterization of $M/G/1$ queues with renewal departures. *Mgt. Sci.* 20: 1222-1228.

Disney, R.L. and Konig, D. 1985. Queueing Networks: A survey of their random processes. *SIAM Rev.* 27: 335-403.

Disney, R.L. and Keissler, P.C. 1987. *Traffic Processes in Queueing Networks: A Markov Renewal Approach.* Johns Hopkins Univ. Press: Baltimore.

Doob, J.L. 1953. *Stochastic Processes.* John Wiley: NY.

Gross, D.G. and Harris, C.M. 1985. *Fundamentals of Queuing Theory.* John

Wiley and Sons: NY

Halfin, S. 1981. Distribution of the interoverflow time for the $GI/G/1$ loss system. *Math. of O. R.* 6: 563-570.

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

Hunter, J.J. 1983. Filtering of Markov renewal queues, II: birth-death queues. *Adv. Appl. Prob¿* 15: 376-391.

Hunter, J.J. 1984. Filtering of Markov renewal queues, III: semi-Markov processes embedded in feedback queues. *Adv. Appl. Prob.* 16: 422-436.

Hunter, J.J. 1985. Filtering of Markov renewal queues, IV: flow processes in feedback queues. *Adv. Appl. Prob.* 17: 386-407.

Karlin, S. and MacGregor, J.L. 1959. A characterization of the birth and death process. *Proc. Acad. Nat. Sci. U.S.A.* 45: 375-79.

Keilson, J. 1979. *Markov Chain Models: Rarity and Exponentiality.* Springer-Verlag: NY.

Kelly, F.P. 1979. *Reversibility and Stochastic Networks.* John Wiley: Chichester.

Kendall, D.G. 1953. Stochastic processes occurring in the theory of queues and their analysis by the method of imbedded Markov chains. *Ann. Math. Statist.* 24: 338-354.

Khinchine, A.Y. 1960. *Mathematical Models in the Theory of Queueing.* Griffin: London.

Kingman, J.F.C. 1962. The effect of queue disciplines on waiting time variance. *Proc. Comb. Phil. Soc.* 58: 163-164.

Kingman, J.F.C. 1970. Inequalities in the theory of queues. *J. Roy. Stat. Soc. B.* 32: 102-110.

Kleinrock, L. 1975. *Queueing Systems, Vol. 1: Theory.* John Wiley and Sons: NY.

Lindley, D.V. 1952. Theory of queues with a single server. *Proc. Camb. Phil. Soc.* 48: 277-289.

Machihara, F. 1987. First passage times of $PH/PH/1/K$ and $PH/PH/1$ queues. *J. Oper. Res. Soc. of Japan.* 30: 1-25.

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

Muntz, R.R. 1972. Poisson departure processes and queueing networks. *Proc.* 7^{th} *Annual Conf. Info. Sci. and Systems.* Princeton, N.J.: 435-440.

Natvig, B. 1977. On the reversibility of the input and output processes for a general birth-and-death queueing model. *J. Appl. Prob.* 14: 876-883.

Natvig, B. 1975 On the input and output processes for a general birth-and-death queueing model. *Adv. Appl. Prob.* 7: 576-592.

Neuts, M.F. 1981 *Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach.* Johns Hopkins Univ. Press: Baltimore.

Ott, T.J. 1984. The sojourn time distribution in the M/G/1 queue with processor sharing. *J. Appl. Prob.* 21: 360-378.

Palm, C. 1943. Intensity fluctuations in telephone traffic engineering. (in German) *Ericsson Technics* 44: 1-189.

Ramaswami, V. 1984. The sojourn time in the GI/M/1 queue with processor sharing. *J. Appl. Prob.* 21: 437-442.

Riordan, J. 1962. *Stochastic Service Systems.* John Wiley: NY.

Schassberger, R. 1984. A new approach to the M/G/1 processor-sharing queue. *Adv. Appl. Prob.* 16: 202-213.

Serfozo, R.F. 1971. Functions of semi-Markov processes. *Siam J. Appl. Math.*

20: 530-535.

Shanthikumar, J.G. and Sumita, U. 1987. Convex ordering of sojourn times in single-server queues: Extremal properties of FIFO and LIFO disciplines. *Adv. Appl. Prob.* 24: 737-748.

Stoyan, D. 1983. *Comparison Methods for Queues and Other Stochastic Models.* edited by D.J. Daley. John Wiley: NY.

VanDoorn, E.A. 1984. On the overflow process from a finite Markovian queue. *Performance Evaluation* 4: 233-240.

Wolff, R.W. 1977. An upper bound for multi-channel queues. *J.Appl. Prob.* 14: 884-888.

Wolff, R.W. 1982. Poisson arrivals see time averages. *Oper. Res.* 30: 223-231.

Yashkov, S.F. 1983. A derivation of response time distribution for a M/G/1 processor-sharing queue. *Problems of Control and Information.* 12: 133-148.

The vita has been removed from the scanned document