VACATION QUEUES WITH MARKOV SCHEDULES

by

M. A. Wortman

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

Joel A. Nachlas, Co-Chairman

Jeffery D. Tew

I. M. Besieris

Kenneth B. Hannsgen

September, 1988

Blacksburg, Virginia

VACATION QUEUES WITH MARKOV SCHEDULES

by

M. A. Wortman

Adviser: Ralph L. Disney

Industrial Engineering and Operations Research

(ABSTRACT)

Vacation systems represent an important class of queueing models having application in both computer communication systems and integrated manufacturing systems. By specifying an appropriate server scheduling discipline, vacation systems are easily particularized to model many practical situations where the server's effort is divided between primary and secondary customers.

A general stochastic framework that subsumes a wide variety of server scheduling disciplines for the M/GI/1/L vacation system is developed. Here, a class of server scheduling disciplines, called Markov schedules, is introduced. It is shown that the queueing behavior M/GI/1/L vacation systems having Markov schedules is characterized by a queue length / server activity marked point process that is Markov renewal and a joint queue length / server activity process that is semi-regenerative. These processes allow characterization of both the transient and ergodic queueing behavior of vacation systems as seen immediately following customer service completions, immediately following server vacation completions, and at arbitrary times.

The state space of the joint queue length / server activity process can be systematically particularized so as to model most server scheduling disciplines appearing in the literature

and a number of disciplines that do not appear in the literature. The Markov renewal nature of the queue length / server activity marked point process yields important results that offer convenient computational formulae. These computational formulae are employed to investigate the ergodic queue length of several important vacation systems; a number of new results are introduced. In particular, the M/GI/1 vacation with limited batch service is investigated for the first time, and the probability generating functions for queue length as seen immediately following service completions, immediately following vacation completions, and at arbitrary times are developed.

Acknowledgements

I thank Dr. Ioannis Besieris for the advice and support he has offered throughout the course of my doctoral studies. Dr. Kenneth Hannsgen has provided me with many of the mathematical tools essential to my work, and I thank him for his efforts. I thank Dr. Jeffery Tew for the quality of his advice and the insight provided by his criticisms.

Thanks are due to Dr. Joel Nachlas whose professionalism and wisdom have, for some years, provided me with a source of sound guidance. (His skill with darts on the other hand has been for me a source of frustration and humiliation that has tested the bounds of my sportsmanship.) The value Joel's friendship and generosity cannot be measured, and hence, no amount of thanks could be enough.

It is not possible to express to Dr. Ralph Disney the depth of my appreciation; attempts to do so would appear effusive and trivial. In his role as adviser he has unselfishly acted as teacher, tutor, taskmaster, and friend. To study with a true scholar, is an opportunity afforded but few students, and it is the high point of my academic life to have studied with Ralph Disney.

I wish to thank Drs. Jeffery Hunter and John Daigle for their many helpful discussions on queueing and stochastic processes. I thank Drs. Daniel Hodge and William Blackwell for their support during my present stay in Blacksburg.

Finally I thank two special members of my family: my father and my wife . To my father, I offer my thanks for providing a role model that has served me well. To my wife, I offer a heartfelt thanks for her patience and love which are a continuing source of inspiration.

Table of Contents

1. Introduction

Queues attended by a single vacationing server have, in recent years, received much attention in the queueing literature. Such queueing systems are most often referred to as "vacation systems" (or "vacation models"). A vacation system, by its most general description, consists of a single-server queue where customers arrive to the queue according to a stochastic process; customer service times are drawn from general distributions. Under specified conditions the server, upon completion of a customer's service, will abandon further customer services to begin a vacation period of random length. When a vacation period is over, the server, again under specified conditions, either begins a customer service or begins another vacation period.

Vacation systems arise naturally as models for many computer communication systems and production systems. In such systems, it often happens that a server's work is divided between two classes of customers: primary and secondary. From the perspective of primary customers, work performed on secondary customers is equivalent to a vacation by the server. While no attempt is made here to justify the validity or accuracy of vacation models in particular applications, it is helpful to consider a pair of simple examples that illustrate vacation models.

Example 1.1 Routine maintenance in computer communications systems: In addition to transmitting and receiving data, processors in computer communication systems perform a variety of testing and maintenance tasks designed to enhance system reliability. Here, managing and processing data is considered the processor's primary activity, while maintenance is considered a secondary activity. The way in which maintenance is scheduled relative to data management and processing is dependent upon system

requirements. Two typical processor scheduling disciplines are illustrated by the following:

i) Since maintenance activity is most often divided into small tasks, whenever the processor finds that there are no primary jobs in the system to service, it begins work on a maintenance task. Upon completing work on this maintenance task, if primary jobs have entered the system, then the processor resumes working on primary jobs. However, if upon completing a maintenance task the processor finds no primary jobs in the system, the processor immediately begins another maintenance task. Here, data management and processing have priority over maintenance activity; however, maintenance tasks are never preempted. Clearly, when primary jobs are being served, the system behaves as a typical single-queue, single-server system. When primary jobs are absent from the system, the server (processor) takes a vacation (to perform maintenance) and continues to take vacations until upon return from a vacation it finds at least one primary job in the system.

ii) An obvious drawback to the processor scheduling discipline of i) is that heavy traffic in the primary jobs can defer maintenance activity for prolonged periods. A processor scheduling discipline that insures maintenance is performed regularly is given by "limiting to m" the number of primary jobs that may be served before a maintenance task is performed. The resulting queueing model indicates that the server takes a vacation upon becoming idle (with respect to primary jobs) or after serving m consecutive primary jobs, whichever comes first.

Example 1.2 Preventive maintenance in production systems: Consider a machine used to assemble items from regular parts batches that arrive at random times to the machine. When the machine becomes idle, preventive maintenance is performed on the

machine. Parts batches arriving to the machine during preventive maintenance must wait for service. Clearly, the machine can be idle following preventive maintenance, and parts arriving to an idle machine where preventive maintenance is completed are unaffected by the maintenance. As in Example 1.1, maintenance is considered a vacation by the assembly items. Note that there is exactly one vacation following each busy period.

Examples 1.1 and 1.2 serve to illustrate that vacation system operation is largely governed by the server scheduling discipline. Typical analyses of such vacation systems focus upon queue length and waiting time distributions. Note that these two examples provide no information regarding: 1) the nature of the stochastic process that governs arrivals to the system, 2) the order in which arrivals are served, 3) queue capacity, 4) the distribution of customer service times, or 5) the distribution of vacation times. Typically, these five fundamental items of information are required in addition to the the server scheduling discipline for any analysis of system performance.

In the developments that follow, a general class of server scheduling disciplines (Markov schedules) is identified. As will be shown, M/GI/1/L vacation systems operating with Markov schedules have a common, well defined stochastic structure. A formal exposition of this common stochastic structure is the focus of the research presented here. To the author's knowledge, identification of the class of Markov schedules, and development of the common stochastic structure for M/GI/1/L vacation systems having Markov schedules is new.

Loosely described, M/GI/1/L vacation systems with Markov schedules refer to vacation systems having the following operational characteristics: 1) Poisson arrival streams, 2) customer service periods drawn from a general distribution that generally depend upon queue length, 3) server vacation periods, drawn from a general distribution that generally

depend upon queue length, and 4) queue capacities that may be either finite or infinite.

The importance of the M/GI/1/L vacation system with Markov schedules is found in the generality of the model. It is easily shown that most (if not all) of the server scheduling disciplines for M/GI/1/L vacation systems considered in the literature are special cases of Markov schedules. Thus, the stochastic processes and their probability structures that underlie these systems provide a general framework for analyzing a wide variety of vacation systems.

While development of a formal theory for the operation of the M/GI/1/L vacation system with Markov schedules is deferred to Chapter 2, it is appropriate to here review some of the important such systems reported in the literature that are subsumed by our system. Doshi (1986) and Takagi (1987) offer excellent review papers discussing vacation models. Details regarding the analysis of specific systems reviewed here are given in these papers.

The *M/GI/1 vacation system with exhaustive service* is a variation of the classical M/GI/1 queue. Here, the server begins a vacation of random length each time the system becomes empty. If upon returning from vacation the server finds the system empty, it immediately begins another vacation. The server continues to operate in this manner until upon return from vacation it finds at least one customer waiting in the queue. This model is often referred to as an M/GI/1 system with *exhaustive service* and *multiple vacations*.

It is assumed for the M/GI/1 vacation system with exhaustive service that customer service periods are independent and identically distributed, and that vacation period are independent and identically distributed. Further, service period lengths and vacation period

lengths are assumed mutually independent and independent of the arrival process. Analogous to the results available for the classical M/GI/1 queue (with no vacations), the current literature Takagi (1987) provides only the probability generating function (pgf) of the ergodic queue length distribution and the Laplace-Stieltjes transform (LST) of the ergodic waiting time distribution for customers when they exist.

Little information regarding the stochastic processes (e.g., server's activity over time, customer departures from the system) that govern the behavior of the M/GI/1 vacation system with exhaustive service is available. However, Fuhrmann (1985) reveals an important decomposition property which shows that the ergodic customer waiting time is given as the sum of two independent random variables. This decomposition consists of the waiting time for the classical M/GI/1 queue (with no vacations) and the forward recurrence time of the vacation period. Doshi (1986) extends the waiting time decomposition of Furhmann to GI/GI/1 vacation systems by using sample path arguments. Kielson and Servi (1986) further generalize the waiting time decomposition to GI/G/1 systems by formalizing arguments presented by Gelenbe and Iasnogorodski (1980).

In the *M/GI/1 vacation system with gated service*, the server upon returning from vacation, services all customers queued at the time of return and then begins another vacation. All customers arriving subsequent to the server's return are held in the queue for service in the period following the end of the next vacation. If the server returns from vacation to find the system empty, another vacation begins immediately, and continues in this manner until upon return from vacation at least one customer is in the queue. Customer service times are independent and identically distributed and are drawn from a general distribution. Similarly, vacation periods are independent and identically distributed and are drawn from a general distribution. Further, customer service times and server vacation times ase mutually independent and independent of the arrival process.

Leibowitz (1961) and Takagi (1987) treat the M/GI/1 vacation system with gated service and offer the pgf of the ergodic queue length and the LST of the ergodic customer waiting time distribution when they exist. The waiting time distribution does not appear to have the decomposition property found in exhaustive service systems.

In *M/GI/1 vacation systems with E-limited service*, the server begins a vacation when either a prespecified number m of customers are served, or the system is emptied which ever occurs first. If the server returns from vacation to find the queue empty, another vacation begins immediately; the server continues in this manner until upon return from vacation, at least one customer is queued. As in previous models, customer service times are independent, identically distributed, and drawn from a general distribution; vacation periods are independent, identically distributed, and drawn from a general distribution. Customer service times and server vacation times are mutually independent and are independent of the arrival process.

It is clear that for the M/GI/1 vacation system with E-limited service, $m = \infty$ corresponds to *exhaustive service*; $m = 1$ is designated as simply *limited service*. Lee (1983) provides an analysis of E-limited service systems that leads to the ergodic queue length pgf at customer service or vacation period completion times. Lee's analysis leads to somewhat complicated expressions for the pgf; no corresponding LST of the ergodic customer waiting time is presented.

The *M/GI/1 vacation system with Bernoulli schedules* consists of a server that will, upon completion of a customer service that leaves the queue not empty, begin another customer service with fixed probability p, or begin a vacation with probability 1-p. If a

service completion leaves the queue empty, a vacation begins immediately. Similar to the server scheduling disciplines above, if the server finds the queue empty upon returning from vacation, then another vacation begins. This operation continues, as before, until the server returns from vacation to find the queue not empty. Customer service times are independent, identically distributed, and drawn from a general distribution; vacation periods are independent, identically distributed, and drawn from a general distribution. Again, customer service times and server vacation times are mutually independent, and are independent of the arrival process. It is clear that for the M/GI/1 vacation system with Bernoulli schedules, the *exhaustive* and *limited* service disciplines are obtained by setting p equal to 1 and 0 respectively.

The Bernoulli schedule service discipline, introduced by Kielson and Servi (1986), was first examined in vacation systems having non-renewal type arrival streams. Their analysis addresses the waiting time decomposition (discussed for exhaustive service above) and investigates stochastic bounds on the ergodic waiting time distribution. Takagi (1987) provides a formula (without development) for the LST of the ergodic waiting time distribution. Takagi's result is obtained by extending arguments used in analyzing other M/GI/1 vacation systems. The pgf for the ergodic queue length (as seen at arbitrary times), to the author's knowledge does not appear in the literature. Ramaswamy and Servi (1986) develop simple expressions for the joint conditional distribution of the busy period and system occupancy at the beginning of a busy period. They also provide expressions for the ergodic occupancy distribution at busy period initiation epochs.

The *M/GI/1 vacation system with G-limited service* is defined as follows: Let m be a prespecified number, and let L_n^* denote the number of messages queued when the server returns form the nth vacation. Upon returning from the nth vacation,the server will serve $\min(L_n^*, m)$ customers, and then begin the next vacation. Customer service times and

vacation periods are independent, identically distributed, and drawn from general distributions. It is clear that with m = 1, G-limited service reduces to simple *limited service*, while m = ∞ corresponds to *gated service.*

Hashida (1981) analyzes the G-limited service system ergodic queue length at the epochs of the server's return from vacation. Takagi (1987) provides extensions to Hashida's work that yields the ergodic queue length pgf at customer departure epochs. Takagi also provides the ergodic customer waiting time LST.

M/GI/1 vacation systems with decrementing service operate in the following manner. When the server returns from vacation to find at least one queued customer, the server will serve customers until the queue occupancy is one less than the number of customers queued at the last vacation completion. As in previous models, if the server returns from vacation to find the queue empty, then another vacation begins immediately. Customer service times as well as server vacation periods are independent, identically distributed, and drawn from general distributions. Service times and vacation times are mutually independent and are independent of the arrival process.

A generalization of the decrementing service discipline is found in the *M/GI/1 vacation system with G-decrementing service.* In this model, the server continues serving until : 1) the number of customers in the system is reduced to (a prespecified) number m less than the number queued at the end of the most recent vacation, or 2) the system becomes empty, whichever occurs first. For m = 1 the system reduces to *decrementing service*, while m = ∞ indicates *gated service.*

Takagi (1987), by extending analyses of other vacation systems, develops both the

ergodic queue length pgf and the ergodic waiting time distribution LST for the G-decrementing service model. As is true with most Takagi results, his analysis here is developed from classical queueing and transform arguments, and provides only limited insight as to the stochastic behavior of the system.

The M/GI/1 vacation systems discussed above (beginning with exhaustive service systems and ending with G-decrementing service systems) represent the most thoroughly investigated M/GI/1/L vacation systems appearing in the literature. These systems together form a small subset in the class of all M/GI/1/L vacation systems with Markov schedules.

The vacation systems discussed above share a set of special characteristics that allow these systems to be analyzed using relatively simple probability arguments. In particular, each of the above systems: 1) serves customers one at a time, 2) has independent, identically distributed customer service times, 3) has independent, identically distributed vacation periods, 4) has service times and vacation times that are mutually independent, and 5) has infinite queue capacity (i.e., $K = \infty$). When any of the five special characteristics is not present, the analysis of M/GI/1/L vacation systems becomes more difficult.

The analyses of the systems discussed above are remarkably similar to analyses of the different variations of the classical M/GI/1 queue (without vacations) where certain "special tricks" are exploited to yield desired results. As is the case with variations of the M/GI/1 queue, the analysis of each vacation system discussed above is largely unique, and is not investigated as a special case of some common model.

For M/GI/1 vacation systems that service customers in a one-at-a-time fashion, it is well known Kleinrock (1976) that the ergodic queue length as seen by departing customers is the same as the ergodic queue length as seen by an outside observer (that is, ergodic

system occupancy at arbitrary times). This fortunate circumstance is exploited throughout the literature, and hence, only for systems that serve customers one at a time have ergodic queue length pgf's been reported in the literature.

There are many simple M/GI/1 vacation systems (e.g., batch service systems) for which the ergodic queue length distribution as seen by departing customers and the ergodic queue length distribution seen at arbitrary times are not the same. As we will show, such systems can often be analyzed within the more powerful framework associated with Markov schedules in the same level of detail as the simpler one at a time service systems.

In the study of M/GI/1/L vacation systems with Markov schedules that follows, the system performance measures that are consider to be of principal importance are: 1) ergodic queue length at arbitrary times, 2) ergodic queue length as seen by customers departing the system, 3) ergodic queue length as seen by the server upon returns from vacation, and 4) ergodic customer waiting times. Developing the probability distributions of these four performance measures for all possible Markov schedule disciplines is formidable (likely impossible). Thus, the focus here is on investigating the underlying probability structure of M/GI/1/L vacation systems with Markov schedules. An understanding of this structure offers a mechanism for investigating ergodic queue lengths and ergodic waiting times within a common framework.

In Chapter 2., a formal development of the probability structure underlying M/GI/1/L vacation systems with Markov schedules is offered. This development is exposed in a "bottom-up" fashion. That is, a stochastic process that governs the behavior of such vacation systems is constructed from more fundamental stochastic processes that govern server activity , queue length, and customer arrivals. The probability structure on the

stochastic process that governs system behavior is shown to be semi-regenerative, with an underlying Markov renewal process whose probability structure is easily characterized.

The model presented in Chapter 2. is shown to be general enough to accommodate systems with finite or infinite queue capacities, state dependent customer service times and vacation periods, and irregular (state dependent) service disciplines. The models allow for the formal characterization of queue length distributions (both transient and ergodic) as seen by departing customers, the server returning from vacation, and at arbitrary times.

While this model and its probability structure accommodate a wide variety of M/GI/1/L vacation systems, obtaining specific formulae useful for engineering calculations from the model is another matter. However, all results appearing in the literature previously discussed may be obtained in a systematic fashion by particularizing the model of Chapter 2. In addition a number of results, not previously reported, are revealed through this systematic particularization.

Chapter 3. addresses application of the M/GI/1/L vacation model with Markov schedules to three different server scheduling disciplines: 1) Bernoulli schedules, 2) E-limited service, and 3) limited batch service. In Sections 3.1 and 3.2, the ergodic behavior of the M/GI/1 vacation systems with Bernoulli schedules and E-limited service are respectively investigated . Here, it is shown the general model of Chapter 2., when particularized to model two well studied systems, gives formulae that agree with those reported in the literature. In addition to developing the ergodic queue length pgf's and ergodic waiting time LST's for Bernoulli schedule and E-limited systems, some new ergodic occupancy results are revealed from the probability structure of the general model.

Section 3.3 considers the M/GI/1 vacation system with single batch service. To the

author's knowledge, this system is not investigated in the available literature. This system is a departure from those systems commonly investigated in that it does not operate with a one at a time service discipline. Consequently, the analysis here appeals to the general semi-regenerative structure of the system in order to develop formulae for ergodic system queue length pgf's. All results presented in this section are, to the author's knowledge, new.

It is not possible in a reasonable space to present all useful formulae that are easily obtained by particularizing the general model of Chapter 2. Thus, Chapter 3. seeks only to demonstrate some of the power and flexibility of the general model and its underlying probability structure.

Chapter 4. offers conclusions drawn from the current research effort, and areas of future research. Particular emphasis is given to possible extensions of the general model of Chapter 2. that address multiple queue, single server systems (polling systems) with Markov switching. Also discussed are qualitative results that may be obtainable from general probability structures, and the value of such qualitative results.

Before closing this introductory chapter, it is appropriate to briefly consider some of the important results reported in the queueing literature that address vacation systems other than M/GI/1/L systems. While the focus of the work offered here is given to studying systems having sophisticated server scheduling disciplines and simple (Poissonian) arrival processes, other investigators emphasize the converse. For example Lucantoni, Meier-Hellstern, and Neuts (1988) consider a vacation system having exhaustive service and a class of non-renewal arrival processes. In this system, the server scheduling discipline is simple while the customer arrival process is a rather sophisticated Markov Arrival Process

(MAP).

Kielson and Servi (1986) examine oscillating random walk models for GI/G/1 vacation systems with Bernoulli schedules; Servi (1986) examines D/GI/1 vacation systems. In these two works the authors investigate vacation systems with simple server scheduling disciplines, and more complicated (renewal-type) arrival processes. Additional works of a similar nature are reviewed by Doshi (1986) and Takagi (1987); readers seeking further review of non-M/GI/1 vacation systems should consult these surveys.

2. The M/GI/1/L Vacation System with Markov Schedules

This chapter provides a formal characterization of the stochastic behavior of the M/GI/1/L vacation system with Markov schedules. To the author's knowledge, the concepts associated with Markov schedules do not appear in the available literature and are identified here for the first time. Markov schedules define a class of server scheduling disciplines that include the scheduling disciplines reviewed in Chapter 1 as a subset. The focus of this chapter is directed towards revealing the generality of Markov schedules, and exploiting this generality to develop a common stochastic framework in which the queueing behavior of most M/GI/1/L vacation systems can be investigated.

In the sections to follow, a "bottom-up" approach is taken in developing the stochastic process that describes the queueing characteristics of M/GI/1/L vacation systems with Markov schedules. This stochastic process is constructed as the joint of more fundamental stochastic processes on which probability structures of practical significance are easily defined.

In Section 2.1, the server switching marked point process is introduced. This stochastic process governs the server's activity over time. In Section 2.2, the server activity marked point process is first introduced and is then used in constructing the joint queue length / server activity process. The joint queue length / server activity is a continuous-time stochastic process that marginally characterizes the system occupancy as seen by an observer outside the system.

In Section 2.3, the probability structure on the queue length / server activity marked point process under Markov schedules is developed. As will be shown, the queue length

/server activity marked point process is embedded within the joint queue length / server activity at convenient stopping times. This embedded marked point process is shown to be Markov renewal.

Section 2.4 presents results that characterize the joint queue length /server activity process for M/GI/1/L vacation systems with Markov schedules as semi-regenerative. The well known theory of semi-regenerative processes is used to characterize system queueing behavior, both transient and stationary. Particular emphasis is given to developing ergodic queue length distributions at stopping times and at arbitrary times.

2.1 The server switching marked point process.

Consider an M/GI/1/L vacation system having a Poisson arrival stream of rate λ. In vacation systems, the server's activity is divided exclusively between customer service periods and vacation periods. The server switching marked point process characterizes the server's activity over time by:

1) identifying the times (epochs) at which the server either completes a service period or completes a vacation period.

2) marking each epoch with a two-tuple indicating:

(i) epoch type ("s-type" for service completion, and "v-type" for vacation completion), and

(ii) number of epochs occurring since the last epoch that was marked by a different type (i.e., a count of the number of consecutive s-type or consecutive v-type epochs).

A realization, denoted by ϕ , of the server switching marked point process is shown in Figure 2.1. Note that associated with each epoch shown in Figure 2.1 is a two-tuple

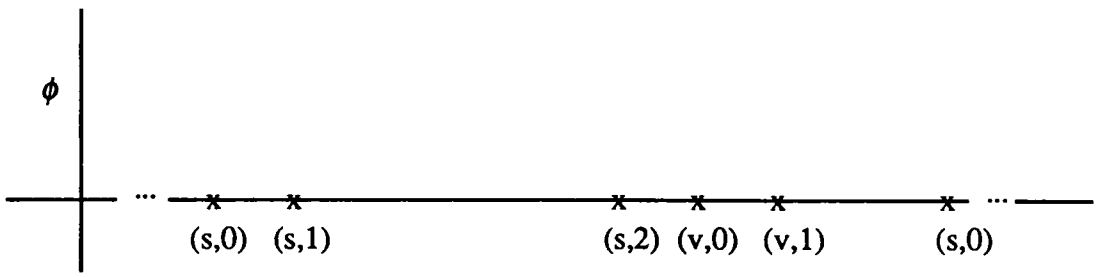

Figure 2.1 A realization of the server switching marked point process.

"mark" indicating epoch type (s or v) and a count of consecutive epochs of the same type.

Now, consider a more formal description of the server switching marked point process. Let ϕ be a realization of the server switching marked point process given by

$$\phi = \{h_m, \phi_m : m \in Z^+\}$$

with

$$\phi_m \in R^+ \text{ and } h_m \in F \times Z^+.$$

Here

R^+ is the set of nonnegative reals,

Z^+ is the set of nonnegative integers, and

$F = \{s,v\}$.

$\phi_m \in R^+$ denotes the time of the m<u>th</u> server switching epoch in the realization ϕ, while h_m marks the m<u>th</u> epoch by type and count on the "mark space" $F \times Z^+$ of the server switching marked point process. For convenience, let $\hat{E} = F \times Z^+$.

Denote by $\Phi_{\hat{E}}$ the set of all such realization ϕ. The probability space $(\Phi_{\hat{E}}, \sigma(\Phi_{\hat{E}}), P)$, with $\sigma(\Phi_{\hat{E}})$ a σ - algebra on $\Phi_{\hat{E}}$ and P a probability measure on $\sigma(\Phi_{\hat{E}})$, defines the server switching marked point process. For each $m \in Z^+$, define the mapping $T_m: \Phi_{\hat{E}} \rightarrow R^+$ as

$$T_m(\phi) = \phi_m,$$

and for each $m \in Z^+$ define the mapping $H_m: \Phi_{\hat{E}} \rightarrow \hat{E}$ as

$H_m(\phi) = h_m$.

The random variable T_m represents the time of the mth server switching event. The random process $T = \{T_m : m \in Z^+\}$ is a (random) point process and is referred to as the *server switching point process*. The random process $H = \{H_m : m \in Z^+\}$ is referred to as the *server switching marked process*. Henceforth, the server switching marked point process will be designated by (H,T).

At this juncture, a specific probability structure on the (H,T) process is not identified. Rather, the focus is here shifted to the construction of the joint queue length / server activity process in which is embedded the (H,T) process. As will be shown, the probability structure on this joint process is more easily characterized than that on the (H,T) process.

2.2 The queue length / server activity marked point process.

Intuition suggests that for any non-trivial vacation system, system occupancy (queue length) and server activity are interdependent. For M/GI/1/L vacation systems, queue length over time is governed by the Poissonian character of customer arrivals and the nature of the server scheduling discipline (as reflected by the server switching marked point process). In this subsection, a continuous time random process is introduced that, together with an embedded marked point process, allows characterization of the system queue length.

Define the random variable $n_t(\phi) \in I \subset Z^+$ as the queue length at time $t \in R^+$. Let

the vector valued random variable $h_t(\phi) \in \hat{E}$ be defined as

$$h_t(\phi) = H_\alpha(\phi) \quad \forall\, t \in R^+,$$

where $\alpha, m \in Z^+$ and, $\alpha = \sup(m \le t)$. Here, the random variable $h_t(\phi)$ indicates the *server's activity* at an arbitrary time t. It is now feasible to define a joint queue length server activity random variable $X_t(\phi)$. For convenience, let $E = \hat{E} \times I$, and define $X_t(\phi) \in E$ as

$$X_t(\phi) = (n_t(\phi), h_t(\phi)) \qquad \forall\, t \in R^+.$$

When the context is clear, the ϕ argument will be omitted in expressions for random variables dependent upon the server's switching activity.

Consider now the stochastic process X_{R^+} given by

$$X_{R^+} = \{X_t : t \in R^+\}$$

which defines the joint queuelength / server activity process. It is assumed that X_{R^+} is a right continuous process. This process, together with its underlying probability structure are the focus of the developments to follow. For general M/GI/1/L vacation systems, the probability structure on X_{R^+} is formidable. However, when the server scheduling discipline belongs to the (yet to be defined) class of Markov schedules, the probability structure on X_{R^+} is manageable.

Characterization of the probability structure on the X_{R^+} process for M/GI/1/L vacation systems with Markov schedules is carried out by first characterizing the probability structure of a particular marked point process embedded within X_{R^+}. The probability

structure on this embedded process will serve in part to formalize the definition of the class of server scheduling disciplines called Markov schedules.

Consider the server switching point process $T = \{T_m : m \in Z^+\}$ introduced in Sec. 2.1. Recall that T_m represents the time of the mth server switching epoch (either a service period completion or a vacation completion). Let X be the stochastic process embedded within X_{R^+} at the instants immediately following the epochs of the server switching point process T. It follows that X is the process given by

$$X = \{X_m : m \in Z^+\}$$

where,

$$X_m = X_{T_m} = (n_{T_m}, h_{T_m}).$$

Here, it is convenient to identify the *embedded queue length process N* given by

$$N = \{N_m : m \in Z^+\}$$

where,

$$N_m = n_{T_m}$$

Thus, it follows that the embedded process X is given by

$$X = (N, H)$$

the joint of the *embedded queue length process* and the *marked process* of the server switching marked point process. Thus, it follows that X_m is the two-tuple given by

$$X_m = (N_m, H_m).$$

At this juncture it can be observed that under the server scheduling disciplines of Chapter 1, the epochs of the server switching point process T are stopping times for the X_{R^+} process which implies that X forms a Markov chain on E. As will be shown in the following subsection, the X process forms a Markov chain for the entire class of Markov schedules.

Next, consider the stochastic process (X,T) formed as the joint of the embedded queue length / server activity process X and the server switching point process T. It follows that (X,T) forms a marked point process. This marked point process is readily recognized as an extension of the server switching point process (H,T) where,

$$(X, T) = (N, H, T)$$

Here, a particular realization of the ϕ of the (X,T) process is given by

$$\phi = \{(x_m, \phi_m): m \in Z^+\} \in \Phi_E$$

where,

$x_m \in E$

$\phi_m \in R^+$ (defined as before)

Φ_E is the set of all realizations.

The (X, T) process, designated as the *queue length / server activity marked point process* plays an essential role in developing a general stochastic structure for M/GI/1/L

vacation systems with Markov schedules. Exposition of a probability structure on an (X, T) process corresponding to Markov schedules is accomplished by examining certain probability structures on the constituent components of (X, T). As will be shown in the following subsection, the probability structure on (X, T) directly implies the probability structure on X_{R}..

2.3 Probability structure of the queue length / server activity marked point process (X,T).

Having in the preceding section defined for a set of stochastic processes that conveniently characterize the queueing behavior of vacation systems, it is possible to now offer a formal definition for Markov schedules vacation systems in terms the probability structure on these stochastic processes.

As a matter of notational convenience, define $i,j \in E$ in terms of their respective queue length and server activity components where,

$$i = (i_N, i_H) \quad \text{and} \quad j = (j_N, j_H)$$

with

$$i_N, j_N \in I \quad \text{and} \quad i_H, j_H \in \hat{E}.$$

Here, a set of conditions that are used to formally define the class of Markov schedules is introduced.

Condition 1.

The server scheduling discipline is such that for all $i,j \in E$, and $m \in Z^+$,

$$P\{H_{m+1} = j_H | X_0,\ldots,X_m,T_0,\ldots,T_m\} = P\{H_{m+1} = j_H | X_m\}$$

For convenience, define $g(i,j)$ as

$$g(i,j) = P\{H_{m+1} = j_H | X_m = i\} \quad \forall\, i,j \in E \tag{2.1}$$

Condition 2.

Customer service periods and server vacation periods are such that for all $i,j \in E$, $m \in Z^+$, and $t \in R^+$

$$P\{T_{m+1} - T_m \le t | H_{m+1},X_0,\ldots,X_m,T_0,\ldots T_m\} = P\{T_{m+1} - T_m \le t | H_{m+1} = j_H, X_m = i\}$$

whenever one-step transitions from state i to state j exist. For convenience, define F(i,j,t) as

$$F(i,j,t) = \begin{cases} P\{T_{m+1} - T_m \le t | H_{m+1} = j_H, X_m = i\}, \text{ for } g(i,j) \neq 0 \\ 0, \text{ for } g(i,j) = 0 \end{cases}$$

$$\forall\, i,j \in E, m \in Z^+, t \in R^+. \tag{2.2}$$

Condition 3.

The system occupancy (queue length) is such that for all $i,j \in E$, $m \in Z^+$, and $t \in R^+$,

$$P\{N_{m+1} = j_N | X_0, \ldots, X_m, H_{m+1}, T_0, \ldots, T_m, T_{m+1} - T_m = t\}$$
$$= P\{N_{m+1} = j_N | X_m, H_{m+1}, T_{m+1} - T_m = t\}$$

Define G(i,j,t) as

$$G(i,j,t) = P\{N_{m+1} = j_N | X_m = i, H_{m+1} = j_H, T_{m+1} - T_m = t\}$$
$$\forall i,j \in E, m \in Z^+, t \in R^+$$

(2.3)

Consider now the following definition for Markov schedules.

Definition 2.1

An M/GI/1/L vacation system having a queue length / server activity marked point process (X, T) satisfying Conditions 1, 2, and 3 above is said to have a server scheduling discipline belonging to the class of Markov schedules.

[]

The Markovian nature of the queueing behavior for vacation systems satisfying Conditions 1, 2, and 3. arises since these conditions are sufficient to assure that all service period completions and all vacation period completions are stopping times for the joint queue length / server activity process X_{R^+}. The importance of the distributions g(i,j), F(i,j,t), and (Gi,j,t) is that they are fundamental information that is usually taken as given for the study of particular M/GI/1/L vacation systems. It is easily reasoned that the vacation systems review in Chapter 1 each satisfy Conditions 1, 2, and 3.

Given the definition for the class of Markov schedules, it is now possible to

characterize the probability structure on the queue length / server activity marked point process (X, T) for M/GI/1/L vacation systems having Markov schedules. Consider the following definition Cinlar (1975).

Definition 2.2

For each $m \in Z^+$, let Y_m be a random variable taking values in the countable set D, and let U_m be a random variable taking values in R^+ such that $0 = U_0 < U_1 < U_2 \ldots$ The stochastic process $(Y, U) = \{Y_m, U_m : m \in Z^+\}$ is said to be Markov renewal with state space D provided that

$$P\{Y_{m+1} = j, U_{m+1} - U_m \le t \mid Y_0, \ldots, Y_m, U_0, \ldots, U_m\} = P\{Y_{m+1} = j, U_{m+1} - U_m \le t \mid Y_m\}$$

$$\forall\, m \in Z^+, j \in D, t \in R^+ \qquad (2.4)$$

(Y,U) is said to be homogeneous when

$$P\{Y_{m+1} = j, U_{m+1} - U_m \le t \mid X_m\} = Q(i,j,t) \qquad \forall\, i,j \in D, t \in R^+$$

independent of m. []

The probability structure on M/GI/1/L vacation systems having Markov schedules is formalized with the following proposition.

Proposition 2.3

An M/GI/1/L vacation system with Markov schedules has a queue length / server activity marked point process (X,T) that is Markov renewal on the state space E.

Proof:

Note that the law of total probability together with Bayes rule imply that $\forall\, m \in Z^+, j \in E, t \in R^+$

$$\begin{aligned}
&P\{X_{m+1} = j, T_{m+1} - T_m \le t \mid X_0,\ldots,X_m,T_0,\ldots,T_m\} \\
&\quad = \int_0^t P\{H_{m+1} = j_H \mid X_0,\ldots,X_m,T_0,\ldots,T_m\} \\
&\qquad \cdot P\{N_{m+1} = j_N \mid H_{m+1} = j_H, X_0,\ldots,X_m,T_0,\ldots,T_m,T_{m+1} - T_m = u\} \\
&\qquad \cdot dP\{T_{m+1} - T_m \le u \mid H_{m+1} = j_H, X_0,\ldots,X_m,T_0,\ldots,T_m\}
\end{aligned} \tag{2.5}$$

Since the system under consideration is M/GI/1/L with Markov schedules, Conditions 1, 2, and 3 hold. Thus, substituting eqns. (2.1), (2.2), and (2.3) into eq. (2.5) shows that $\forall\, m \in Z^+, j \in E, t \in R^+$

$$\begin{aligned}
&P\{X_{m+1} = j, T_{m+1} - T_m \le t \mid X_0,\ldots,X_m,T_0,\ldots,T_m\} \\
&\quad = \int_0^t P\{H_{m+1} = j_H \mid X_m\} \cdot P\{N_{m+1} = j_N \mid H_{m+1}, X_m, T_{m+1} - T_m = u\} \\
&\qquad \cdot dP\{T_{m+1} - T_m \le u \mid H_{m+1}, X_m\}
\end{aligned} \tag{2.6}$$

The right side of eq. (2.6) can be rewritten as

$$\begin{aligned}
&\int_0^t P\{H_{m+1} = j_H \mid X_m\} \cdot P\{N_{m+1} = j_N \mid H_{m+1}, X_m, T_{m+1} - T_m = u\} \\
&\cdot dP\{T_{m+1} - T_m \le u \mid H_{m+1}, X_m\} = P\{X_{m+1} = j, T_{m+1} - T_m \le t \mid X_m\}
\end{aligned} \tag{2.7}$$

Thus (X,T), by Definition 2.2, forms a Markov renewal process on E.

[]

When (X,T) is a Markov renewal process, the family of probabilities

$$Q(t) = \{Q(i,j,t) : i,j \in E, t \in R^+\}$$

is called the semi-Markov kernel over E. Note that for all $i,j \in E$, the mapping $t \rightarrow Q(i,j,t)$ has all properties of a probability distribution function except that

$$Q(i,j) = \lim_{t \rightarrow \infty} Q(i,j,t)$$

in general is not necessarily one. However, it follows directly from Definition 2.2 that

$$\sum_{j \in E} Q(i,j) = 1, \quad \forall\, i \in E \tag{2.8}$$

which leads to the following proposition.

Proposition 2.4

For an M/GI/1/L vacation system with Markov schedules, the marked process X associated with the queue length / server activity marked point process (X,T) forms a Markov chain on the state space E.

Proof:

Since the vacation system under consideration is M/GI/1/L with Markov schedules, it follows from Proposition 2.3 that the queue length / server activity marked point process (X,T) is Markov renewal. That X forms a Markov chain on E follows directly from (2.8) and the definition of a Markov chain.

[]

In the developments that follow, only M/GI/1/L vacation systems having Markov schedules are considered; thus, the queue length / server activity process (X,T) is always taken as a Markov renewal process. Given that (X,T) is Markov renewal, it is important

for results to be developed later that state classifications of the queue length / server activity process (X,T) be identified.

In (X,T) let $W_1^j, W_2^j, \ldots$ be the times between successive visits to state $j \in E$. If S_0^j represents the time of the first visit to state j, then

$$S_{m+1}^j = S_m^j + W_{m+1}^j, \quad \forall\, m \in Z^+$$

define the times of the visits to state j. It follows from Definition 2.2 that the sequence

$$S^j = \left\{S_m^j - S_0^j : m \in Z^+\right\}$$

forms a renewal process.

Definition 2.5

State $j \in E$ is said to be *recurrent* if, in the renewal process S^j, $W_m < \infty$ for each m a.s.; otherwise, state j is called transient. State $j \in E$ is said to be *periodic* with period δ if, in the renewal process S^j, the random variables $W_1^j, W_2^j, \ldots$ take values in the set $\{0, \delta, 2\delta, \ldots\}$ and δ is the largest such number. If no such δ exists, then j is said to be aperiodic.

[]

A state $j \in E$ in the (X, T) process is recurrent if and only if j is a recurrent state in the underlying Markov chain X. Thus, in order to address the question of recurrence in the states of (X, T), only the Markov chain X need be investigated. A state $j \in E$ in the (X, T) process is periodic if and only if the distribution of the time between two consecutive visits to state j is arithmetic with span δ.

Note that there exist Markov schedules such that it is possible for j to be periodic for (X, T) without being periodic for the embedded Markov chain X. Conversely, j can for some Markov schedules be periodic for the embedded chain X and aperiodic for the (X, T) process. Cinlar (1975) offers a number of criteria suitable for testing the periodicity and/or recurrence of the states of a Markov renewal processes. Those criteria, while not presented here, are suitable to classify the states of the queue length / server activity marked point process (X,T).

In most practical situations, the Markov schedules of interest are limited to those schedules leading to an embedded queue length / server activity process X that is irreducible (i.e., any state in E can be reached from any other state in E); this will become more clear in Chapter 3. When the corresponding vacation system is stable, $P\{n_t < \infty : t \in R^+\} = 1$, and X is irreducible, then (X, T) is characterized as being aperiodic and recurrent (see Cinlar 75). It is, however, emphasized that less practical Markov schedules leading to reducible X processes are easily accommodated within the probability structure on (X, T) considered thus far.

Having, for M/GI/1/L vacation systems with Markov schedules, characterized the (X, T) process as Markov renewal with semi-Markov kernel Q(t), it is possible to characterize the probability structure of a projection of the queue length / server activity marked point process onto a subspace of the state space E. The importance of characterizing such a projection arises when studying queue length as seen at particular epoch types.

Consider the realization $\phi = \{(x_m, \phi_m) : m \in Z^+\} \in \Phi_E$ (defined previously). Following the reasoning of Disney and Kiessler (1987), for $A \subset E$ let the sequence

$\phi^A = \{(x_m^A, \phi_m^A) : m \in Z^+\}$ be a subsequence of ϕ consisting of all pairs (x_m, ϕ_m) for which $x_m \in A$. Should it happen that the number of m such that $x_m \in A$ is finite, let m^* be the largest of these m, and for all $m > m^*$, let $x_m^A = \Delta$ and $\phi_m^A = +\infty$. Hence, the sequence ϕ^A is defined for all $m \in Z^+$. Now, for $k \in Z^+$, define $L_k^A : \Phi_E \to Z^+ \cup \{+\infty\}$ as

$$L_0^A(\phi) = \begin{cases} \inf\{m \in Z^+ : x_m \in A\} & \{m \in Z^+ : x_m \in A\} \neq \varnothing \\ +\infty & \text{otherwise} \end{cases}$$

and for k = 1, 2, ...,

$$L_k^A(\phi) = \begin{cases} \inf\{m > L_{k-1}^A(\phi) : x_m \in A\} & \{m > L_{k-1}^A : x_m \in A\} \neq \varnothing \\ +\infty & \text{otherwise} \end{cases}$$

For $k \in Z^+$, define

$$X_k^A(\phi) = x_k^A = X_{L_k^A(\phi)}(\phi)$$

and

$$S_k^A(\phi) = \phi_k^A = T_{L_k^A(\phi)}(\phi)$$

The stochastic process $(X^A, S^A) = \{(X_k^A, S_k^A) : k \in Z^+\}$ is the delayed Markov renewal process formed by embedding the (X, T) process at visits to the set $A \subset E$. (Note that when the set A consists of a single state (i.e., A={i}, $i \in E$), then (X^A, S^A) forms an delayed ordinary renewal process. Takagi (1987) recognized a special case of this fact and employed this special case in a number of his arguments.)

It is a simple matter to construct an ordinary Markov renewal process from the delayed Markov renewal process (X^A, S^A). The following proposition is offered without proof.

Proposition 2.6

For $k \in Z^+$, let $T_k^A = S_k^A - S_0^A$. The $(X^A, T^A) = \{(X_k^A, T_k^A) : k \in Z^+\}$ process is Markov renewal on the state space A. []

If given the semi-Markov kernel Q(t) for (X, T), it is possible to construct the semi-Markov kernel for (X^A, T^A).

Theorem 2.7

Let $Q_A(t)$ be the semi-Markov kernel for the (X^A, T^A) process. $Q_A(t)$ is given in terms of Q(t) by

$$Q_A(i,j,t) = Q(i,j,t) + \sum_{k=2}^{\infty} \sum_{i_1 \in B} \cdots \sum_{i_{k-1} \in B} \int_0^t \cdots \int_0^{t-u_1-\cdots-u_{k-1}} Q(i,i_1,du_1) \cdots Q(i_{k-1},j,t-u_1-\cdots-u_{k-1}) \tag{2.9}$$

where $i,j \in A$, $t \in R^+$, and $B=A^c$.

Proof:

See Disney and Kiessler (1987). []

The usefulness of Theorem 2.7 is demonstrated when considering the joint queue length / server activity process embedded only at service period completion epochs or only

at vacation period completion epochs. Let

$S = \{i \in E : i_H = (s, \cdot)\}$, and

$V = \{i \in E : i_H = (v, \cdot)\}$.

That is, $S \subset E$ denotes the set of all states corresponding to s-type epochs while $V \subset E$ denotes the set of all states corresponding to v-type epochs. Note that $E = S \cup V$ and $S \cap V = \varnothing$; thus, S and V together partition the state space E. Since E is at most countable, it is possible to express the semi-Markov kernel Q(t) as a matrix partitioned in blocks according to the S,V partition of E. That is

$$Q(t) = \begin{bmatrix} Q_{SS}(t) & Q_{SV}(t) \\ Q_{VS}(t) & Q_{VV}(t) \end{bmatrix} \tag{2.10}$$

Note that at this juncture, no particular ordering of states is specified, and when $K=\infty$ the submatrix blocks of (2.9) are each necessarily infinite dimensional.

Corollary 2.8

a) The joint queue length / server activity process embedded at service period completion epochs (X^S, T^S) is Markov renewal and has a semi-Markov kernel $Q_S(t)$ given by

$$Q_S(t) = Q_{SS}(t) + \sum_{k=0}^{\infty} \left[Q_{SV} * Q_{VV}^{(k)} * Q_{VS} \right](t) \tag{2.11}$$

b) The joint queue length / server activity process embedded at vacation period completion epochs (X^V, T^V) is Markov renewal and has a semi-Markov kernel $Q_V(t)$ given by

$$Q_V(t) = Q_{VV}(t) + \sum_{k=0}^{\infty} \left[Q_{VS} * Q_{SS}^{(k)} * Q_{SV} \right](t) \tag{2.12}$$

Here, $*$ is the convolution operator, and $Q_{(\bullet\bullet)}^{(k)}(t)$ is the k-fold convolution of $Q_{(\bullet\bullet)}(t)$

with itself.

Proof:

The fact that in part a) (X^S, T^S) and in part b) (X^V, T^V) are Markov renewal follows directly from Proposition 2.6. Since S and V partition E, eq. (2.11) in part a) and eq. (2.12) in part b) are both matrix representations of eq. (2.9) of Thm. 2.7 with S and V serving for A and B.

Corollary 2.8 assures that the queueing behavior of M/GI/1/L vacation systems with Markov schedules retain a Markov renewal structure when the system is examined at service completion epochs, vacation completion epochs, or both service completion and vacation completion epochs. While the semi-Markov kernel Q(t) is often easily formulated, the semi-Markov kernels $Q_s(t)$ and $Q_v(t)$ are usually formidable (as indicated by the complexity of eqns. (2.11) and (2.12)) and are difficult to formulate . The relationship between Q(t), $Q_s(t)$, and $Q_v(t)$ will be examined for some specific vacation systems in Chapter 3.

The characterization of (X^S, T^S) and (X^V, T^V) as Markov-renewal offers additional insight into the behavior of vacation systems. That is, (X^S, T^S) characterizes both the queue length embedded at departures from the system and the point process governing the customer departure stream. The (X^S, T^S) process marginally characterizes in part the backlog of customers awaiting service when the server returns from vacation. Ergodic results, when they exist , are obtained by examining the stationary distributions on the Markov chains X^S and X^V in the usual way.

2.4 Probability structure for the joint queue length / server activity process X_R..

Having identified the queue length / server activity marked point process (X, T) as Markov renewal under Markov schedules, characterization of the probability structure on the joint queue length / server activity X_R. requires introduction of the so called *Markov renewal equations.*

As defined in the previous subsection, Q(t) is the semi-Markov kernel for (X, T) on the state space E. Let f(t) and b(t) be vectors whose elements f(i,t) and b(i,t) are nonnegative functions that are bounded on finite intervals of t and are bounded in i. Suppose that b(t) is a known function and that f(t) is an unknown function. Then, the equation

$$f(i,t) = b(i,t) + \sum_{j \in E} \int_0^t Q(i,j,du) f(j,t-u) \qquad (2.13)$$

is called a Markov renewal equation. The set of Markov renewal equations on the state space E is given conveniently in matrix form as

$$f(t) = b(t) + (Q * f)(t) \qquad (2.14)$$

The solution to eq. (2.14) requires introduction of the *Markov renewal kernel* R(t) of $(X, T) \in E$. Here, elements of the Markov renewal kernel are referred to as *Markov renewal functions* and are given by

$$R(i,j,t) = E\left[\sum_{m=0}^{\infty} 1_{\{j\}}(X_m) 1_{[0,t]}(T_m) \mid X_0 = i\right] \qquad i,j \in E,\ t \in R^+$$

(2.15)

where, $1_{(\bullet)}$ denotes an indicator function. Each Markov renewal function R(i,j,t) can be written in terms of elements of Q(t), the semi-Markov kernel of (X, T), by

$$R(i,j,t) = \sum_{m=0}^{\infty} P\{X_m = j, T_m \le t \mid X_0 = i\}$$
$$= \sum_{m=0}^{\infty} Q^{(m)}(i,j,t) \qquad (2.16)$$

where,

$$Q^{(0)}(i,j,t) = \begin{cases} 1, & i = j \\ 0, & \text{otherwise} \end{cases}$$

and

$$Q^{(m)}(i,j,t) = \sum_{k \in E} \int_0^t Q(i,k,du) Q^{(m-1)}(i,j,t-u)$$

From a computational perspective, the Markov renewal kernel is formulated using Laplace-Stieltjes transforms. Let $\mathrm{Re}\ \sigma \ge 0$, and for all $i,j \in E$

$$Q_\sigma(i,j) = \int_0^\infty e^{-\sigma t} Q(i,j,dt),$$

and

$$R_\sigma(i,j) = \int_0^\infty e^{-\sigma t} R(i,j,dt).$$

It follows easily by taking the Laplace-Stieltjes transform of (2.16) that R_σ is given as the minimal solution to

$$R_\sigma(U - Q_\sigma) = U$$

where U is an identity matrix. In practice, it is most difficult to formulate the Markov renewal kernel R(t). This difficulty is especially evident when the state space E is not finite.

It is well known Cinlar (1975) that the solution to the Markov renewal equation of (2.14) is given by

$$f(t) = (R * b)(t) + c(t) \qquad (2.17)$$

where c(t) is a vector whose elements are functions of the same class the elements of f(t) and

$$c(t) = (Q * c)(t) \qquad (2.18)$$

Generally, (2.17) is not unique. However, when the server switching point process T has infinite lifetime (i.e., $\sup_m(T_m) = \infty$ a.s.), then $c(i,t) = 0 \ \forall \ i \in E$ and

$$f(t) = (R * b)(t) \qquad (2.19)$$

solves (2.14) uniquely.

Before identifying the probability structure on the joint queue length / server activity process $X_{R^+} = \{X_t : t \in R^+\}$, a definition is needed.

Definition 2.9

Let $Z = \{Z_t : t \geq 0\}$ be a stochastic process with topological space D. Suppose that the function $t \to Z_t(\omega)$ is right continuous and left hand limits exist for almost all ω. The process Z is said to be semi-regenerative if there exists a Markov renewal process $(Y,U) = \{Y_m, U_m : m \in Z^+\}$ having infinite lifetime satisfying the following:

a) for each $m \in Z^+$, U_m is a stopping time for Z,

b) for each $m \in Z^+$, Y_m is determined by $\{Z_u : u \leq U_m\}$,

c) for each $m \in Z^+$, $n \geq 1$, $0 \leq t_1 \leq t_2 \leq \ldots \leq t_n$, and function f defined on D^n and positive,

$$E_i\left[f\left(Z_{T_m+t_1},\ldots,Z_{T_m+t_n}\right) \mid Z_u : u \leq U_m\right] = E_j\left[f\left(Z_{t_1},\ldots,Z_{t_n}\right)\right] \text{ on } \{Y_m = j\}$$

where,

E_P E_j refer to expectations given the initial state for the Markov chain Y.

[]

It is now a simple matter to identify the probability structure on the joint queue length / server activity process X_{R^+}.

Proposition 2.10

Consider an M/GI/1/L vacation system having Markov schedules such that the server switching point process $T = \{T_m : m \in Z^+\}$ has an infinite lifetime. Then, the joint queue length / server activity process $X_{R^+} = \{X_t : t \in R\}$ is semi-regenerative.

Proof:

Note that the server switching point process $T = \{T_m : m \in Z^+\}$ has infinite lifetime if and only if the queue length /server activity marked point process

$$(X, T) = \{X_m, T_m : m \in Z^+\}$$

has infinite lifetime. Since the arrival stream to the system is Poisson, it follows that

$$P\{X_t = j \mid T_1 = u, X_1 = i\} = P\{X_{t-u} = j \mid X_0 = i\}$$

for all $i,j \in E$ and $u,t \in R^+$ where $u < t$ (i.e., (X, T) is an embedded Markov renewal process). Thus, $T = \{T_m : m \in Z^+\}$ is a set of stopping times for X_{R^+} and condition a) is satisfied. Further, since $X_{T_m} = X_m$, we have that X_m is determined by $\{X_u : u \le T_m\}$ satisfying condition b). That condition c) is satisfied follows from the Markov renewal property of (X, T).

[]

The importance of identifying the semi-regenerative structure on X_{R^+} is that the well known theory of semi-regenerative processes can be used to examine the system occupancy distribution as seen by an "outside observer". To this end, the two theorems to follow provide distributional results that are powerful tools for examining the queueing behavior of M/GI/1/L vacation systems having Markov schedules.

Theorem 2.11

Consider an M/GI/1/L vacation system having Markov schedules such that the server switching point process $T = \{T_m : m \in Z^+\}$ has infinite lifetime. For any $A \subset E$, all $i \in E$, and all $t \in R^+$,

$$P\{X_t \in A \mid X_0 = i\} = \sum_{k \in E} \int_0^t R(i,k,du)\, P\{X_{t-u} \in A, T_1 > t - u \mid X_0 = k\} \tag{2.20}$$

Proof:

It follows from Proposition 2.10 that the vacation system under consideration is such

that the joint queue length / server activity process X_{R^+} is semi-regenerative on the state space E. Following the logic of Kohlas (1982), note that

$$\{X_t \in A \mid X_0 = i\} = \{X_t \in A, T_1 > t \mid X_0 = i\} \cup \left\{ \cup_{k \in E} \{X_t \in A, T_1 > t, X_1 = k \mid X_0 = i\} \right\}.$$

Using the regeneration property,it now follows that

$$P\{X_t \in A \mid T_1 = s, X_1 = k\} = P\left\{X_{t-T_1} \in A \mid X_0 = k\right\}.$$

Thus,

$$P\{X_t \in A \mid X_0 = i\} = P\{X_t \in A, T_1 > t \mid X_0 = i\} + \sum_{k \in E} \int_0^t Q(i,k,du) P\{X_{t-u} \in A \mid X_0 = k\} \tag{2.21}$$

which is a Markov renewal equation . It follows from eq. (2.19) that the solution to (2.21) is given by (2.20)

[]

While the result given by eq. (2.10) of Thm. 2.11 is rather general, (2.20) offers little promise as a computational tool. This is true since computing the Markov renewal kernel R(t) is, in practice, a formidable if not impossible task. If, however, attention is restricted to ergodic queueing behavior (when it exists) a more computationally attractive result is available.

Theorem 2.12

Consider an M/GI/1/L vacation system having Markov schedules such that the queue length / server activity marked point process (X,T) is irreducible, aperiodic, and recurrent on the state space E. Let $A \subset E$, ν be an invariant measure for the Markov chain X, and $m(k) = E[T_1 | X_0 = k]$. Suppose that $\nu m = \sum_{k \in E} \nu(k) m(k) < \infty$.

Then,

$$\lim_{t \to \infty} P\{X_t \in A | X_0 = i\} = \frac{1}{\nu m} \sum_{k \in E} \nu(k) \int_0^{\infty} P\{X_t \in A, T_1 > t | X_0 = k\} dt \tag{2.22}$$

provided that $t \to P\{X_t \in A, T_1 > t | X_0 = k\}$ is Riemann integrable $\forall\, k \in E$.

Proof:

Since it follows from Proposition 2.10 that the vacation system under consideration is such that the joint queue length / server activity process $X_{R_\cdot}$ is semi-regenerative on the state space E, proof here is the same as Cinlar's (1975) proof of Theorem 10.6.12 .

[]

Note that (2.22), unlike (2.21) does not require the computation of the Markov renewal kernel R(t). However, (2.22) requires computation of ν a stationary measure on X; computation of such a measure is in principle simple so long as K is finite. When K is infinite, computation of ν is more difficult. This situation is considered for specific vacation systems in Chapter 3.

Theorem 2.12 leads directly to a characterization of the ergodic queue length distribution (when it exist) for the class of Markov schedules identified in the hypothesis of the theorem.

Corollary 2.13

Consider an M/GI/1/L vacation system having Markov schedules such that the queue length / server activity marked point process (X,T) is irreducible, aperiodic, and recurrent on the state space E. Let $A \subset E$, ν be an invariant measure for the Markov chain X, and $m(k) = E[T_1 | X_0 = k]$. Suppose that $\nu m = \sum_{k \in E} \nu(k) m(k) < \infty$. Then for each $j_N \in Z^+$,

$$\lim_{t \to \infty} P\{n_t = j_N\} = \frac{1}{\nu m} \sum_{j_H \in \hat{E}} \left[\sum_{k \in E} \nu(k) \int_0^\infty P\{X_t = j | T_1 > t, X_0 = k\} P\{T_1 > t | X_0 = k\} \, dt \right]$$

(2.23)

Proof:

Recall that $X_t = (n_t, h_t) \in E$. It follows directly that for all $j_N \in Z^+$ and $t \in R^+$,

$$P\{n_t = j_N\} = \sum_{j_H \in \hat{E}} P\{n_t = j_N, h_t = j_H\},$$

and since $\sum_{j_H \in \hat{E}} P\{n_t = j_N, h_t = j_H\}$ is a convergent series of all positive terms, we have that

$$\lim_{t \to \infty} P\{n_t = j_N\} = \lim_{t \to \infty} \sum_{j_H \in \hat{E}} P\{n_t = j_N, h_t = j_H\} = \sum_{j_H \in \hat{E}} \lim_{t \to \infty} P\{n_t = j_N, h_t = j_H\},$$

(2.24)

Thus, substituting (2.22) into (2.24) gives the desired result.

[]

A convenient probability structure that underlies M/GI/1/L vacation systems with

Markov schedules has now been identified. This structure, as revealed by Proposition 2.3 and Proposition 2.10, offers a common framework in which all M/GI/1/L vacation systems having Markov schedules may be examined. Some observations regarding the Markov renewal / semi-regenerative nature of such systems are in order.

The three conditions that identify those sever scheduling disciplines belonging to the class of Markov schedules are relatively general. Note that these conditions do not specify the order in which queued customers are serviced (e.g., first come first served, etc.); in fact, customers may be served in batches.

The three conditions defining Markov schedules admit a variety of server scheduling disciplines where customer service period distributions and/or server vacation period distributions are dependent upon the arrival process; the ordinary M/GI/1 queue is an example of such a vacation system. Here, the server vacations while the queue is idle and terminates its vacation immediately upon arrival of a customer to the empty queue.

Theorems 2.11 and 2.12 together with Corollary 2.13 offer powerful tools for analyzing characteristics of the queue length distribution for M/GI/1/L vacation systems having Markov schedules. This set of results is valid for systems having either finite or infinite queue capacities. For systems having finite queue capacities, numerical results are, in principle, readily calculated. Systems having infinite queue capacities are usually more difficult to analyze.

In the Chapter 3, the common framework developed in Chapter 2. is employed examine the queue length distributions for systems having finite and systems having infinite queue capacities. The difficulties that arise in obtaining specific numerical results will be made

clear in the developments of Chapter 3.

3. Example M/GI/1/L Vacation Systems with Markov Schedules

In this chapter, the general probability structure underlying M/GI/1/L vacation systems with Markov schedules is particularized to examine the queueing behavior of example vacation systems. The purpose of examining these example systems is threefold.

First, we seek to validate the general probability structure of Chapter 2 by examining the queueing behavior of a pair of previously studied vacation systems and comparing the results of this examination to known results. Second, we seek to demonstrate the usefulness of the general probability structure by providing previously unreported results associated with well studied systems. Finally, we seek to demonstrate the usefulness of the general probability structure by examining the queueing behavior of previously unstudied vacation systems.

In meeting this threefold purpose, the full generality of the probability structure of Chapter 2 is not exploited; rather, only ergodic results will be examined. Unless otherwise indicated, the the queue length / server activity process (X,T) is assumed to be irreducible and to possess a stationary distribution.

Chapter 3 is divided into three sections. Section 3.1 investigates the queueing behavior of the M/GI/1 vacation system with Bernoulli schedules, Sec. 3.2 investigates the behavior of the M/GI/1 vacation system with E-limited service, while Sec 3.3 investigates the M/GI/1 vacation system with batch service. The systems considered here appear in order of increasing complexity.

Note that the three systems considered each have infinite queue capacity (i.e., $L = \infty$). Such systems are, in principle, more difficult to analyze; note that Theorem 2.12 requires a stationary measure for a Markov chain having a countably infinite state space. Particularization of the general model of Chapter 2 to vacation systems having infinite queue capacities, in most instances, yields only probability generating functions (pgf's) for queue length. The nature of computational difficulties associated with infinite queue capacity systems will be made clear as the example systems are analyzed in their respective sections.

3.1 M/GI/1 vacation systems with Bernoulli schedules.

Consider again the M/GI/1 vacation system with Bernoulli schedules, first introduced in Chapter 1. Recall that the "Bernoulli schedule" server scheduling discipline requires that upon completion of a customer's service that leaves the queue not empty, the server will either begin serving the next customer in line with fixed probability p, or will begin a vacation with fixed probability 1-p. Upon a service period completion that leaves the system empty, the server begins a vacation period begins immediately.

At the end of a vacation period , the server arrives to find the queue either empty or not. Recall that if the server returns to find the queue not empty, a service period begins immediately; if the server returns to find the queue empty, another vacation period begins immediately. Customers are served in order of arrival. Exhaustive service and limited service server scheduling disciplines are obtained as special cases of the Bernoulli schedule by setting p to 1 and 0 respectively.

For M/GI/1 vacation systems with Bernoulli schedules, it is assumed that the lengths of customer service periods are independent, identically distributed with distribution S(t), and the lengths of server vacation periods are independent, identically distributed with distribution V(t). It is further assumed that the lengths of customer service periods and the lengths of server vacation periods are mutually independent. The Poisson stream of customers arriving to the system is assumed to have rate λ.

Since the queueing behavior of M/GI/1 vacation systems with Bernoulli schedules is to be examined within the general framework established in the previous chapter, it is necessary to show that Bernoulli schedules belong to the class of Markov schedules. That is, Bernoulli schedules must satisfy Conditions 1, 2, and 3. However, before verifying that Conditions 1, 2, and 3 hold, it is helpful to reexamine the mark space of the server switching marked point process (H,T).

When examining M/GI/1 vacation systems with Bernoulli schedules, the full generality of the model introduced in Chapter 2 is not required. In particular, let the mark space $\hat{E}$ of the server switching marked point process (H,T) be restricted to $\hat{E} = F$. Under Bernoulli schedules, this simplification of the mark space is appropriate since for any given epoch of (H,T), the type (s-type or v-type) of this epoch depends only upon the most recent previous epoch.

Let the joint queue length / server activity process X_{R^+} and the queue length / server activity marked point process (X,T) be defined as in Chapter 2. Since the queue capacity is infinite (i.e., $L = \infty$), it follows that the state space E for the system under consideration is given by

$$E = \hat{E} \times Z^+ = F \times Z^+.$$

Recall that $i,j \in E$ are expressible in terms of queue length and server activity components where,

$$i = (i_N, i_H) \text{ and } j = j(j_N, j_H)$$

with

$$i_N, j_N \in Z^+ \text{ and } i_H, j_H \in F.$$

Under Bernoulli schedules, the type of the next epoch of (X,T) depends only upon the present epoch of (X,T). Thus, it follows simply that for all $i,j \in E$ and $m \in Z^+$,

$$P\{H_{m+1} = j_H \mid X_0, \ldots, X_m, T_0, \ldots, T_m\} = P\{H_{m+1} = j_H \mid X_m\},$$

which implies that Bernoulli schedules satisfy Condition 1. As in (2.2), let

$$g(i,j) = P\{H_{m+1} = j_H \mid X_m\} \qquad \forall\, i,j \in E.$$

It follows that for systems operating with Bernoulli schedules,

$$g(i,j) = \begin{cases} 1, & j_H = v, i_N = 0, i_H \in F \\ 1-p, & j_H = v, i_N \neq 0, i_H = s \\ 1, & j_H = s, i_N \neq 0, i_H = v \\ p, & j_H = s, i_N \neq 0, i_H = s \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

It is true that for all vacation systems that the server is either serving customers or on vacation . For Bernoulli schedules, the time between any two contiguous epochs of (X,T) must be either a customer service period with distribution S(t) or as a server vacation period

with distribution V(t). Given two contiguous epochs, the server's activity between them is recognized as either a vacation period or a services period, conditional upon the type of the most recent of the two epoch and the number of customers queued at the older of the two epochs. That is, customer service periods and server vacation periods are such that for all $m \in Z^+$,

$$P\{T_{m+1} - T_m \le t \mid H_{m+1}, X_0, \ldots, X_m, T_0, \ldots, T_m\} = P\{T_{m+1} - T_m \le t \mid H_{m+1}, X_m = i\}$$

whenever j is "reachable in one step" from i. Hence, Bernoulli schedules satisfy Condition 2. Using the notation of (2.2), we have for all $i,j \in E$, $m \in Z^+$, $t \in R^+$ that

$$F(i,j,t) = \begin{cases} P\{T_{m+1} - T_m \le t \mid H_{m+1} = j_H, X_m = i\}, & g(i,j) \neq 0 \\ 0, & g(i,j) = 0 \end{cases}$$

Since inter-epoch times represent either vacation periods or service periods, it follows that

$$F(i,j,t) = \begin{cases} S(t), & j_H = s, i \in E, j \text{ one} - \text{step reachable from } i \\ V(t), & j_H = v, i \in E, j \text{ one} - \text{step reachable from } i \\ 0, & \text{otherwise} \end{cases} \qquad (3.2)$$

Since the customer arrival stream is Poisson, it is clear that the interarrival times are exponentially distributed and thus, have the memoryless property. Hence, for all $i,j \in E$, $m \in Z^+$, and $t \in R^+$,

$$P\{N_{m+1} = j_N | H_{m+1}, X_0, \ldots, X_m, T_0, \ldots, T_m, T_{m+1} - T_m = t\}$$
$$= P\{N_{m+1} = j_N | H_{m+1} = j_H, X_m = i, T_{m+1} - T_m = t\}.$$

which satisfies Condition 3. Following the notation of (2.3), we have for all $i,j \in E$, $m \in Z^+$, $t \in R^+$

$$G(i,j,t) = P\{N_{m+1} = j_N | H_{m+1} = j_H, X_m = i, T_{m+1} - T_m = t\}$$

and it follows that

$$G(i,j,t) = \begin{cases} \dfrac{e^{-\lambda t}(\lambda t)^{j_N - i_N + 1}}{(j_N - i_N + 1)!}, & i_N \geq 1,\ j_N \geq i_N - 1,\ j_H = s \\ \dfrac{e^{-\lambda t}(\lambda t)^{j_N - i_N}}{(j_N - i_N)!}, & i_N \geq 0,\ j_N \geq i_N,\ j_H = v \\ 0, & \text{otherwise} \end{cases} . \quad (3.3)$$

Since Conditions 1, 2, and 3 are satisfied for this example system, Bernoulli schedules belong to the class of Markov schedules. Thus, by Proposition 2.3 M/GI/1 vacation systems with Bernoulli schedules have a Markov renewal queue length / server activity marked point process (X,T), and have a semi-regenerative joint queue length / server activity process X_{R^+}. Given (3.1), (3.2), and (3.3), it is a simple matter to calculate the semi-Markov kernel Q(t) for the (X,T) process where,

$$Q(i,j,t) = \int_0^t g(i,j)G(i,j,u)F(i,j,du) \quad \forall\ i,j \in E,\ t \in R^+ \quad (3.4)$$

Substituting (3.1), (3.2), and (3.3) into (3.4) it follows that

$$Q(i,j,t) = \begin{cases} \int_0^t \frac{(\lambda u)^{j_N - i_N}}{(j_N - i_N)!} e^{-\lambda u} V(du), & j_H = v, i_N = 0, i_H \in F \\ (1-p)\int_0^t \frac{(\lambda u)^{j_N - i_N}}{(j_N - i_N)!} e^{-\lambda u} V(du), & j_H = v, i_N \neq 0, i_H = s \\ \int_0^t \frac{(\lambda u)^{j_N - i_N + 1}}{(j_N - i_N + 1)!} e^{-\lambda u} S(du), & j_H = s, i_N \neq 0, i_H = v \\ p\int_0^t \frac{(\lambda u)^{j_N - i_N + 1}}{(j_N - i_N + 1)!} e^{-\lambda u} S(du), & j_H = s, i_N \neq 0, i_H = s \\ 0, & \text{otherwise} \end{cases} \quad (3.5)$$

Having particularized the semi-Markov kernel Q(t) to reflect M/GI/1 vacation systems with Bernoulli schedules, it is feasible to investigate the ergodic queueing behavior of this system. (Recall that ergodic results exist here since it is assumed that (X,T) is irreducible and all states of E are recurrent.) In particular, the ergodic distribution of queue length as seen immediately following customer service completions, the ergodic distribution of queue length as seen immediately following the server's return from vacation, and the ergodic distribution of queue length as seen at an arbitrary time are investigated. Nonergodic queueing behavior is not examined here.

From Proposition 2.4 we have that the marked process X associated with the queue length / server activity marked point process (X,T) forms a Markov chain. Since X is embedded at all customer service completions and all server vacation completions, the

ergodic distribution of queue length as seen by either customers upon departure or the server upon return from vacation is simply the stationary distribution for the chain X. Note that by Corollary 2.13, the ergodic distribution of queue length as seen at an arbitrary time requires the stationary distribution for X. The three distributions of interest above are each determined, in part, by a stationary measure on the chain X.

Because the state space E is countably infinite, solving for a stationary measure on X is formidable. (This situation is analogous to studying the ergodic queue length distribution of the M/GI/1 queue without vacations where generating functions are used to study queue length distributions.) Thus, in what follows, pgf's of queue length are developed.

Let Q be the collection of one-step transition probabilities associated with the Markov chain X. Here, $Q = \lim_{t \to \infty} Q(t)$. Let the state space E be partitioned, as in Corollary 2.8, such that $E = S \cup V$ and $S \cap V = \varnothing$ where,

$S = \{i \in E : i_H = s\}$, and

$V = \{i \in E : i_H = v\}$.

The equations yielding the stationary distribution on X, when partitioned according to the state space partition described above, are written in matrix form as

$$[\pi_S \; \pi_V] = [\pi_S \; \pi_V] \begin{bmatrix} Q_{SS} & Q_{SV} \\ Q_{VS} & Q_{VV} \end{bmatrix} \tag{3.6}$$

where,

$$Q = \begin{bmatrix} Q_{SS} & Q_{SV} \\ Q_{VS} & Q_{VV} \end{bmatrix}$$

and $[\pi_S \; \pi_V]$ is the stationary distribution on X. Here, for $\alpha \in F$,

$$\pi_\alpha = [\pi_\alpha(0)\ \pi_\alpha(1)\ \pi_\alpha(2)\ \ldots]$$

where,

$$\pi_\alpha(j) = \lim_{m \to \infty} P\{N_m = j, H_m = \alpha\},\ j = 0, 1, 2, \ldots$$

From (3.5) and (3.6) it follows that

$$Q_{ss} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \ldots \\ ps_0 & ps_1 & ps_2 & ps_3 & ps_4 & \ldots \\ 0 & ps_0 & ps_1 & ps_2 & ps_3 & \ldots \\ 0 & 0 & ps_0 & ps_1 & ps_2 & \ldots \\ 0 & 0 & 0 & ps_0 & ps_1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{3.7}$$

$$Q_{sv} = \begin{bmatrix} v_0 & v_1 & v_2 & v_3 & v_4 & \ldots \\ 0 & qv_0 & qv_1 & qv_2 & qv_3 & \ldots \\ 0 & 0 & qv_0 & qv_1 & qv_2 & \ldots \\ 0 & 0 & 0 & qv_0 & qv_1 & \ldots \\ 0 & 0 & 0 & 0 & qv_0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{3.8}$$

$$Q_{vs} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \ldots \\ s_0 & s_1 & s_2 & s_3 & s_4 & \ldots \\ 0 & s_0 & s_1 & s_2 & s_3 & \ldots \\ 0 & 0 & s_0 & s_1 & s_2 & \ldots \\ 0 & 0 & 0 & s_0 & s_1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{3.9}$$

and

$$Q_{vv} = \begin{bmatrix} v_0 & v_1 & v_2 & v_3 & v_4 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{3.10}$$

where, q = 1-p and for j = 0, 1, 2, ...,

$$s_j = \int_0^{\infty} \frac{(\lambda t)^j e^{-\lambda t}}{j!} S(dt), \tag{3.11}$$

and

$$v_j = \int_0^{\infty} \frac{(\lambda t)^j e^{-\lambda t}}{j!} V(dt). \tag{3.12}$$

Substituting (3.7) through (3.10) into (3.6), it follows that for j = 0, 1, 2, ...,

$$\pi_s(j) = p\sum_{k=1}^{j+1} \pi_s(k) s_{j-k+1} + \sum_{k=1}^{j+1} \pi_v(k) s_{j-k+1}, \tag{3.13}$$

and

$$\pi_v(j) = \pi_s(0) v_j + q\sum_{k=1}^{j} \pi_s(k) v_{j-k} + \pi_v(0) v_j. \tag{3.14}$$

Now, define the following geometric transforms:

$$\Pi_s(z) = \sum_{j=0}^{\infty} \pi_s(j) z^j, \tag{3.15}$$

and

$$\Pi_v(z) = \sum_{j=0}^{\infty} \pi_v(j) z^j. \tag{3.16}$$

Equation (3.15) together with (3.13) gives

$$\Pi_s(z) = \sum_{j=0}^{\infty} z^j \left(p\sum_{k=1}^{j+1} \pi_s(k) s_{j-k+1} + \sum_{k=1}^{j+1} \pi_v(k) s_{j-k+1} \right) \tag{3.17}$$

while, (3.16) together with (3.14) gives

$$\Pi_V(z) = \sum_{j=0}^{\infty} z^j \left(\pi_S(0)\, v_j + q \sum_{k=1}^{j+1} \pi_S(k) v_{j-k} + \pi_V(0)\, v_j \right). \tag{3.18}$$

Distributing the outer sum on the right side of (3.17) and interchanging the order of summations allows (3.17) to be rewritten as

$$\begin{aligned} \Pi_S(z) = {} & p\pi_S(1)\sum_{j=0}^{\infty} s_j z^j + zp\pi_S(2)\sum_{j=0}^{\infty} s_j z^j + z^2 p\pi_S(3)\sum_{j=0}^{\infty} s_j z^j + \ldots \\ & + \pi_V(1)\sum_{j=0}^{\infty} s_j z^j + z\pi_V(2)\sum_{j=0}^{\infty} s_j z^j + z^2\pi_V(3)\sum_{j=0}^{\infty} s_j z^j + \ldots \end{aligned} \tag{3.19}$$

Similarly, distributing the outer sum on the right side of (3.18) and interchanging the order of summation allows (3.18) to be rewritten as

$$\Pi_V(z) = \pi_V(0)\sum_{j=0}^{\infty} v_j z^j + \pi_S(0)\sum_{j=0}^{\infty} v_j z^j + qz\pi_S(1)\sum_{j=0}^{\infty} v_j z^j + qz^2\pi_S(2)\sum_{j=0}^{\infty} v_j z^j + \ldots \tag{3.20}$$

Now, define $\tilde{S}(z)$ and $\tilde{V}(z)$ such that

$$\tilde{S}(z) = \sum_{j=0}^{\infty} s_j z^j, \tag{3.21}$$

and

$$\tilde{V}(z) = \sum_{j=0}^{\infty} v_j z^j \tag{3.22}$$

Expressions for $\tilde{S}(z)$ and $\tilde{V}(z)$ are found by substituting (3.11) and (3.12) into (3.21) and (3.22) respectively. Thus,

$$\tilde{S}(z) = \sum_{j=0}^{\infty} z^j \int_0^{\infty} \frac{(\lambda t)^j e^{-\lambda t}}{j!} S(dt) \tag{3.23}$$

and

$$\tilde{V}(z) = \sum_{j=0}^{\infty} z^j \int_0^{\infty} \frac{(\lambda t)^j e^{-\lambda t}}{j!} V(dt) \tag{3.24}$$

Passing the summation within the integral in both (3.23) and (3.24), it follows directly that

$$\tilde{S}(z) = \int_0^{\infty} e^{-(\lambda - \lambda z)t} S(dt) \tag{3.25}$$

and

$$\tilde{V}(z) = \int_0^{\infty} e^{-(\lambda - \lambda z)t} V(dt) \tag{3.26}$$

where, (3.25) and (3.26) are recognized as Laplace-Stieltjes transforms, of S(t) and V(t) respectively, evaluated with the transform operators equal to $(\lambda - \lambda z)$.

Now, (3.21) and (3.22) respectively allow (3.19) and (3.20) to be written as

$$\Pi_s(z) = \frac{\tilde{S}(z)}{z}\left(p\Pi_s(z) + \Pi_v(z) - \left(p\pi_s(0) + \pi_v(0)\right)\right) \tag{3.27}$$

and

$$\Pi_v(z) = \tilde{V}(z)\left(q\Pi_s(z) + \left(p\pi_s(0) + \pi_v(0)\right)\right) \tag{3.28}$$

Equations (3.27) and (3.28) can be solved simultaneously for both $\Pi_s(z)$ and $\Pi_v(z)$. As will be shown, for appropriate boundary conditions $\Pi_s(z)$ and $\Pi_v(z)$ are pgf's where, $\Pi_s(z)$ is the pgf for the queue length distribution seen by customers departing the system and $\Pi_v(z)$ is the pgf for the queue length distribution as seen by the server upon return from vacation.

As a matter of convenience, rearrange (3.27) and (3.28) such that

$$\frac{z}{\tilde{S}(z)}\Pi_s(z) = -\left(p\pi_s(0) + \pi_v(0)\right) + p\Pi_s(z) + \Pi_v(z) \tag{3.29}$$

and

$$\frac{1}{\tilde{V}(z)}\Pi_s(z) = \left(p\pi_s(0) + \pi_v(0)\right) + q\Pi_s(z) \tag{3.30}$$

It follows by adding (3.29) to (3.30) that

$$\Pi_s(z) + \Pi_v(z) = \frac{z}{\tilde{S}(z)}\Pi_s(z) + \frac{1}{\tilde{V}(z)}\Pi_v(z), \tag{3.31}$$

and solving (3.31) for $\Pi_v(z)$ yields

$$\Pi_v(z) = \frac{\tilde{V}(z)\left(\tilde{S}(z) - z\right)}{\tilde{S}(z)\left(1 - \tilde{V}(z)\right)}\Pi_s(z) \tag{3.32}$$

Substituting (3.32) into (3.29) and solving for $\Pi_s(z)$ gives

$$\Pi_s(z) = \frac{(p\pi_s(0) + \pi_v(0))(\tilde{V}(z) - 1)\tilde{S}(z)}{z - (p + q\tilde{V}(z))\tilde{S}(z)}. \qquad (3.33)$$

Substituting (3.33) into (3.32) and simplifying gives

$$\Pi_v(z) = \frac{(p\pi_s(0) + \pi_v(0))(z - \tilde{S}(z))\tilde{V}(z)}{z - (p + q\tilde{V}(z))\tilde{S}(z)}. \qquad (3.34)$$

Note that while (3.33) and (3.34) are geometric transforms, neither is necessarily a pgf. The value of the constant $(p\pi_s(0) + \pi_v(0))$ determines whether or not either (3.33) or (3.34) is a pgf; the value of $(p\pi_s(0) + \pi_v(0))$ required to make (3.33) a pgf is generally different than the value of $(p\pi_s(0) + \pi_v(0))$ required to make (3.34) a pgf. In the developments to follow, we shall show how the values of the constant $(p\pi_s(0) + \pi_v(0))$ is determined so that (3.33) and (3.34) become pgf's.

Next, consider $\Pi(z)$ the pgf for the distribution of the queue length as seen immediately following either customer service completions or server vacation completions. It is clear that $\Pi(z) = \Pi_s(z) + \Pi_v(z)$; thus, it follows from (3.31) that

$$\Pi(z) = \frac{z}{\tilde{S}(z)}\Pi_s(z) + \frac{1}{\tilde{V}(z)}\Pi_v(z) \qquad (3.35)$$

Substituting (3.33) and (3.34) into (3.35) and simplifying yields

$$\Pi(z) = \frac{(p\pi_s(0) + \pi_v(0))(z\tilde{V}(z) - \tilde{S}(z))}{z - (p + q\tilde{V}(z))\tilde{S}(z)} \qquad (3.36)$$

Note that any pgf taken in the limit as z approaches 1 from inside the unit circle is itself 1; that is, $\lim_{z \uparrow 1} \Pi(z) = 1$. Therefore, it follows that

$$1 = \lim_{z \uparrow 1} \frac{(p\pi_s(0) + \pi_v(0))(z\tilde{V}(z) - \tilde{S}(z))}{z - (p + q\tilde{V}(z))\tilde{S}(z)} \tag{3.37}$$

Note that (3.37) is of an indeterminant. Here, let the "prime" diacritical mark indicate differentiation with respect to z, and by L'Hopital's rule we have that

$$\lim_{z \uparrow 1} \frac{(p\pi_s(0) + \pi_v(0))(z\tilde{V}(z) - \tilde{S}(z))}{z - (p + q\tilde{V}(z))\tilde{S}(z)} = \lim_{z \uparrow 1} \frac{(p\pi_s(0) + \pi_v(0))(z\tilde{V}'(z) + \tilde{V}(z) - \tilde{S}'(z))}{1 - ((p + q\tilde{V}(z))\tilde{S}'(z) + \tilde{S}(z)q\tilde{V}'(z))} \tag{3.38}$$

It follows from (3.37) and (3.38) that

$$1 = \frac{(p\pi_s(0) + \pi_v(0))(\tilde{V}'(1) + 1 - \tilde{S}'(1))}{1 - (\tilde{S}'(1) + q\tilde{V}'(1))}. \tag{3.39}$$

Let $\bar{S}$ denote the expected length of a customer service period and $\bar{V}$ denote the expected length of a server vacation period. Note that $\tilde{S}'(1) = \lambda\bar{S}$ and $\tilde{V}'(1) = \lambda\bar{V}$. Rearranging (3.39) gives

$$(p\pi_s(0) + \pi_v(0)) = \frac{1 - \rho - q\lambda\bar{V}}{1 - \rho + \lambda\bar{V}} \tag{3.40}$$

where, $\rho = \lambda\bar{S}$, as usual, defines the traffic intensity. Equation (3.40) gives the

appropriate value of $(p\pi_s(0) + \pi_v(0))$ so that $\Pi(z)$ is a pgf. We now have that

$$\Pi(z) = \frac{1 - \rho - q\lambda\bar{V}}{1 - \rho + \lambda\bar{V}} \cdot \frac{(z\tilde{V}(z) - \tilde{S}(z))}{z - (p + q\tilde{V}(z))\tilde{S}(z)} \tag{3.41}$$

The case of exhaustive service is now examined; setting p to 1 in (3.41) gives

$$\Pi(z) = \frac{1 - \rho}{1 - \rho + \lambda\bar{V}} \cdot \frac{(z\tilde{V}(z) - \tilde{S}(z))}{z - \tilde{S}(z)}, \tag{3.42}$$

which agrees with the results of Fujiki and Gambe (1980).

The usefulness of (3.41) is limited since it addresses none of the system performance measures discussed earlier. However (3.42), by agreeing with the results of Fujiki and Gambe, (3.41) offers a partial validation of this particularization of the general model of Chapter 2.

At this juncture, it is possible to examine the pgf of the queue length as seen by customers immediately following departure from the system, and the pgf of the queue length as seen by the server immediately following returns from vacation. Here, it is convenient to employ Corollary 2.8 which characterizes the probability structure of the joint queue length / server activity process embedded at either service period completion epochs or at vacation period completion epoch.

First, we will examine the pgf of the distribution for the queue length as seen immediately following customer departures from the system. Let (X^s, T^s) be defined as

in Corollary 2.8. It is clear that the queue length distribution, as seen by customers departing the system, is given by $\hat{\pi}_S$ the stationary distribution on the Markov chain X^S. Here, $\hat{\pi}_S$ satisfies

$$\hat{\pi}_S = \hat{\pi}_S Q_S \tag{3.43}$$

where,

$$Q_S = \lim_{t \to \infty} Q_S(t).$$

It follows from (2.11) that the transition matrix Q_S for the chain X^S is given by

$$Q_S = Q_{SS} + \sum_{k=0}^{\infty} (Q_{SV} Q_{VV}^k Q_{VS}) \tag{3.44}$$

However, from (3.6) we have that

$$\begin{aligned} \pi_S &= \pi_S Q_{SS} + \pi_V Q_{VS} \\ \pi_V &= \pi_S Q_{SV} + \pi_V Q_{VV} \end{aligned} \tag{3.45}$$

Solving (3.45) simultaneously yields

$$\pi_S = \pi_S \left(Q_{SS} + \sum_{k=0}^{\infty} (Q_{SV} Q_{VV}^k Q_{VS}) \right); \tag{3.46}$$

hence, from (3.44) it is clear that π_S is a stationary measure on the chain X^S. It now follows that π_S and $\hat{\pi}_S$ are equivalent up to a multiplicative constant; hence, the pgf for π_S differs from the pgf for $\hat{\pi}_S$ by a multiplicative constant. This reasoning reveals that if the

constant $(p\pi_s(0) + \pi_v(0))$ is chosen such that $\lim_{z\uparrow 1} \Pi_s(z) = 1$, then $\Pi_s(z)$ is the pgf for the queue length as seen by customers departing the system. Further, by interchanging the roles of S and V in the above argument, $\Pi_v(z)$ is recognized as the pgf for the queue length as seen by the server immediately following returns from vacation when the constant $(p\pi_s(0) + \pi_v(0))$ is chosen such that $\lim_{z\uparrow 1} \Pi_v(z) = 1$.

It is possible now to determine the value of $(p\pi_s(0) + \pi_v(0))$ such that $\Pi_s(z)$ becomes the pgf for the queue length as seen by customers departing the system. Since $\lim_{z\uparrow 1} \Pi_s(z) = 1$, we have from (3.33) that

$$1 = \lim_{z\uparrow 1} \frac{(p\pi_s(0) + \pi_v(0))(\tilde{V}(z) - 1)\tilde{S}(z)}{z - (p + q\tilde{V}(z))\tilde{S}(z)} \tag{3.47}$$

Since (3.47) is of an indeterminant form, L'Hopital's rule is required to evaluate the limit; hence,

$$\lim_{z\uparrow 1} \frac{(p\pi_s(0) + \pi_v(0))(\tilde{V}(z) - 1)\tilde{S}(z)}{z - (p + q\tilde{V}(z))\tilde{S}(z)}$$

$$= \lim_{z\uparrow 1} \frac{(p\pi_s(0) + \pi_v(0))((\tilde{V}(z) - 1)\tilde{S}'(z) + \tilde{S}(z)\tilde{V}'(z))}{1 - ((p + q\tilde{V}(z))\tilde{S}'(z) + q\tilde{V}'(z)\tilde{S}(z))} \tag{3.48}$$

From, (3.47) and (3.48), it follows that

$$1 = \frac{(p\pi_s(0) + \pi_v(0))\tilde{V}'(1)}{1 - (\tilde{S}'(1) + q\tilde{V}'(1))} \tag{3.49}$$

and with ρ, $\bar{S}$, and $\bar{V}$ defined as before, (3.49) can be rearranged to show that

$$\left(p\pi_s(0) + \pi_v(0)\right) = \frac{1 - \rho - q\lambda\bar{V}}{\lambda\bar{V}}. \qquad (3.50)$$

Substituting (3.50) into (3.33) gives

$$\Pi_s(z) = \frac{1 - \rho - q\lambda\bar{V}}{\lambda\bar{V}} \cdot \frac{\left(\tilde{V}(z) - 1\right)\tilde{S}(z)}{z - \left(p + q\tilde{V}(z)\right)\tilde{S}(z)} . \qquad (3.51)$$

which is the desired pgf. To the author's knowledge, the pgf of (3.51) is a new result for M/GI/1 vacation systems with Bernoulli schedules. Note that with p = 1, (3.51) becomes

$$\Pi_s(z) = \frac{1 - \rho}{\lambda\bar{V}} \cdot \frac{\left(\tilde{V}(z) - 1\right)\tilde{S}(z)}{z - \tilde{S}(z)} \qquad (3.52)$$

which is the well known pgf for the queue length as seen by departing customers of M/GI/1 vacation systems with exhaustive service Takagi (1987).

It is a simple matter to determine the value of the constant $\left(p\pi_s(0) + \pi_v(0)\right)$ such that $\Pi_v(z)$ becomes the pgf of the queue length as seen by the server immediately following returns from vacation. Since $\lim_{z \uparrow 1} \Pi_v(z) = 1$, we have from (3.34) that,

$$1 = \lim_{z \uparrow 1} \frac{\left(p\pi_s(0) + \pi_v(0)\right)\left(z - \tilde{S}(z)\right)\tilde{V}(z)}{z - \left(p + q\tilde{V}(z)\right)\tilde{S}(z)} . \qquad (3.53)$$

Following an application of L'Hopitals rule, (3.53) reduces to

$$1 = \frac{(p\pi_s(0) + \pi_v(0))(z - \tilde{S}(1))}{z - q\tilde{V}(1) - \tilde{S}(1)}, \tag{3.54}$$

and from (3.54) it follows that

$$p\pi_s(0) + \pi_v(0) = \frac{1 - \rho - q\lambda\bar{V}}{1 - \rho}. \tag{3.55}$$

Substituting (3.55) into (3.34) yields

$$\Pi_v(z) = \frac{1 - \rho - q\lambda\bar{V}}{1 - \rho} \cdot \frac{(z - \tilde{S}(z))\tilde{V}(z)}{z - (p + q\tilde{V}(z))\tilde{S}(z)} \tag{3.56}$$

which completes the characterization of the pgf for the queue length as seen by the server immediately following returns from vacation. The pgf of (3.56) does not appear in the available literature, and thus, is new.

We now consider the pgf for the queue length as seen at arbitrary times. Development of this pgf appeals to Theorem 2.12 and Corollary 2.13 which address the stationary distribution of the joint queue length / server activity process $X_{\mathfrak{R}}$.. Since (X,T) is here assumed to be irreducible and to have a stationary distribution, it follows from (2.22) that for all j in E

$$\lim_{t\to\infty} P\{X_t = j\} = \frac{1}{\pi m}\sum_{k\in E}\pi(k)\int_0^\infty P\{X_t = j \mid T_1 > t, X_0 = k\}P\{T_1 > t \mid X_0 = k\}\,dt$$

(3.57)

where, π is the stationary distribution on the Markov chain X.

Here, some notation is introduced so as to simplify the development that follows. For all $j \in E$ let

$$\eta(j) = \lim_{t \to \infty} P\{X_t = j\}$$

and for $i,j \in E$ let

$$B(i,j) = \int_0^{\infty} P\{X_t = j \mid T_1 > t, X_0 = k\} P\{T_1 > t \mid X_0 = k\}\, dt$$

When the state space E is partitioned by S and V as in Corollary 2.8, it follows from (3.57) that the stationary distribution of the joint queue length / server activity process $X_{\mathfrak{R}^+}$ is given by

$$[\eta_S \; \eta_V] = \frac{1}{\pi m}[\pi_S \; \pi_V] \begin{bmatrix} B_{SS} & B_{SV} \\ B_{VS} & B_{VV} \end{bmatrix}. \qquad (3.58)$$

Here, for $\alpha \in F$,

$$\eta_\alpha = [\eta_\alpha(0) \; \eta_\alpha(1) \; \eta_\alpha(2) \; \ldots]$$

with

$$\eta_\alpha(j) = \lim_{t \to \infty} P\{n_t = j, h_t = \alpha\}, \quad j = 0, 1, 2, \ldots;$$

for all α, $\beta \in F$, and i,j = 0, 1, 2, ...

$$B_{\alpha\beta}(i,j) = \int_0^{\infty} P\{n_t = j, h_t = \beta \mid T_1 > t, N_0 = i, H_0 = \alpha\} \cdot P\{T_1 > t \mid N_0 = i, H_0 = \alpha\}\, dt \tag{3.59}$$

It is clear that whenever $T_1 > t$, it must be that $h_t = H_0$. Thus, for i,j = 0, 1, 2, ..., $B_{SV}(i,j) = B_{VS}(i,j) = 0$. It now follows from (3.59) that

$$B_{SV} = B_{VS} = [0]. \tag{3.60}$$

Further, (3.59) implies that

$$B_{SS} = \begin{bmatrix} a_0 & a_1 & a_2 & a_3 & a_4 & \cdots \\ 0 & d_0 & d_1 & d_2 & d_3 & \cdots \\ 0 & 0 & d_0 & d_1 & d_2 & \cdots \\ 0 & 0 & 0 & d_0 & d_1 & \cdots \\ 0 & 0 & 0 & 0 & d_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{3.61}$$

and

$$B_{VV} = \begin{bmatrix} a_0 & a_1 & a_2 & a_3 & a_4 & \cdots \\ 0 & c_0 & c_1 & c_2 & c_3 & \cdots \\ 0 & 0 & c_0 & c_1 & c_2 & \cdots \\ 0 & 0 & 0 & c_0 & c_1 & \cdots \\ 0 & 0 & 0 & 0 & c_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{3.62}$$

where for j = 0, 1, 2, ...

$$a_j = \int_0^{\infty} \frac{(\lambda t)^j e^{-\lambda t}}{j!}(1 - V(t))dt\,, \tag{3.63}$$

$$d_j = \int_0^{\infty} \frac{(\lambda t)^j e^{-\lambda t}}{j!}(1 - (pS(t) + qV(t)))dt\,, \tag{3.64}$$

and

$$c_j = \int_0^{\infty} \frac{(\lambda t)^j e^{-\lambda t}}{j!}(1 - S(t))dt\,. \tag{3.65}$$

Given the notation above, it is a straight forward matter to examine the pgf of the queue length distribution as seen at arbitrary times. Let κ be the row vector of queue length probabilities when the queue is observed at arbitrary times where

$$\kappa = [\kappa(0)\ \kappa(1)\ \kappa(2)\ ..] \tag{3.66}$$

with

$$\kappa(j) = \lim_{t\to\infty} P\{n_t = j\}, \qquad j = 0,1,2,\ldots \tag{3.67}$$

It follows directly from Corollary 2.13 that $\kappa(j)$ is given by

$$\kappa(j) = \eta_s(j) + \eta_v(j), \qquad j = 0,1,2,\ldots \tag{3.68}$$

Here, (3.68) taken together with (3.58), (3.61), and (3.62) shows that $\kappa(j)$ can be rewritten as

$$\kappa(j) = \frac{1}{\pi m}\left(\pi_s(0)a_j + \sum_{k=1}^{j} \pi_s(k)d_{j-k} + \pi_v(0)a_j + \sum_{k=1}^{j} \pi_v(k)c_{j-k} \right), \quad j = 0, 1, 2, \ldots \tag{3.69}$$

Defining $K(z)$ as the pgf of the queue length as seen at arbitrary times we have that

$$K(z) = \sum_{j=0}^{\infty} \kappa(j) z^j; \tag{3.70}$$

hence, it follows from (3.69) and (3.70) that

$$K(z) = \frac{1}{\pi m} \sum_{j=0}^{\infty} z^j \left(\pi_s(0)a_j + \sum_{k=1}^{j} \pi_s(k)d_{j-k} + \pi_v(0)a_j + \sum_{k=1}^{j} \pi_v(k)c_{j-k} \right) \tag{3.71}$$

Let $\Pi_s(z)$ and $\Pi_v(z)$ be defined by (3.15) and (3.16) respectively. Distributing the outer summation on the right side of (3.71) and then interchanging the order of summation in each term in the usual manner, (3.71) can be rewritten as

$$K(z) = \frac{1}{\pi m}\left(\pi_s(0)(A(z) - D(z)) + \pi_v(0)(A(z) - C(z)) + D(z)\Pi_s(z) + C(z)\Pi_v(z) \right) \tag{3.72}$$

where ,

$$A(z) = \sum_{j=0}^{\infty} a_j z^j = \int_0^{\infty} e^{-(\lambda - \lambda z)t}(1 - V(t))dt ,$$

$$D(z) = \sum_{j=0}^{\infty} d_j z^j = \int_0^{\infty} e^{-(\lambda - \lambda z)t}(1 - (pS(t) + qV(t)))dt ,$$

and

$$C(z) = \sum_{j=0}^{\infty} c_j z^j = \int_0^{\infty} e^{-(\lambda - \lambda z)t}(1 - S(t))dt .$$

Noting that $\lim_{z \uparrow 1} K(z) = 1$, and following much routine algebra (not shown here), (3.72) reduces to

$$K(z) = \frac{1 - \lambda \bar{S} - q\lambda \bar{V}}{\lambda \bar{V}} \cdot \frac{(\tilde{V}(z) - 1)\tilde{S}(z)}{z - (p + q\tilde{V}(z))\tilde{S}(z)} \qquad (3.73)$$

Equation (3.73) and (3.51) show that $K(z) = \Pi(z)$. That the pgf of the queue length as seen immediately following customer departures is the same as the pgf of the queue length as seen at arbitrary times is to be expected for this system. Klienrock (1975) shows that for queues with renewal type arrivals where customers are served one at a time, the queue length as seen immediately before arrivals and the queue length as seen immediately following departures are distributed the same. Wolff (1982) shows that the queue length as seen immediately before arrivals belonging to a Poisson stream is distributed the same as the queue length seen at an arbitrary time (sometimes referred to as the PASTA result). The Wolff result together with the Klienrock result imply that for queueing systems having Poisson arrivals and one at a time customer service, the queue length seen immediately following customer departures and the queue length as seen at an arbitrary time are

distributed the same; the M/GI/1 vacation system with Bernoulli schedules is such a system. (Takagi (1987) states this result as a theorem and provides a simple proof.)

Given the pgf of the queue length as seen at an arbitrary time $K(z)$, it is a simple matter to formulate the Laplace-Stieltjes transform of the ergodic customer waiting time for the M/GI/1 vacation system with Bernoulli schedules. Here, we appeal to the distributional form of Little's law as presented by Keilson and Servi (1988). The distributional form of Little's law is an ergodic result; thus, it is assumed that (X,T) is irreducible and possesses a stationary distribution.

Let $K(z)$ be the pgf of the queue length as seen at an arbitrary time, and let $W(\sigma)$ be the Laplace -Stieltjes transform of the waiting time T of an arbitrary customer. Here, $W(\sigma) = E[e^{-\sigma T}]$. The following Proposition, proven by Keilson and Servi (1988), is a statement of the distributional form of Little's law.

Proposition 3.1

Let an ergodic queueing system be such that

a) arrivals are Poisson of rate λ,

b) all arriving customers enter the system and remain in the system until served,

c) customers are served one at a time in order of arrival

d) newly arriving customers do not affect the waiting time of customers already in the system

Then, the distributional form of Little's law holds; that is,

$$K(z) = W(\lambda - \lambda z) \tag{3.74}$$

[]

Clearly, M/GI/1 vacation systems with Bernoulli schedules satisfy the conditions of Proposition 3.1. Recall that (3.25) and (3.26) show that

$$\tilde{S}(z) = S^*(\lambda - \lambda z) \tag{3.75}$$

and

$$\tilde{V}(z) = V^*(\lambda - \lambda z) \tag{3.76}$$

where $S^*(\sigma)$ and $V^*(\sigma)$ are Laplace-Stieltjes transforms of S(t) and V(t) respectively. Substituting (3.75) and (3.76) into (3.74) and making a change of variable indicated by (3.73) where $z = 1 - \frac{\sigma}{\lambda}$, we obtain the Laplace-Stieltjes transform of the waiting time T for an arbitrary customer. That is,

$$W(\sigma) = \frac{1 - \lambda(\bar{S} + q\bar{V})}{\bar{V}} \cdot \frac{(1 - V^*(\sigma))S^*(\sigma)}{\sigma - \lambda + \lambda(p + qV^*(\sigma))S^*(\sigma)} \tag{3.77}$$

The waiting time Laplace-Stieltjes transform of (3.77) concludes the analysis of M/GI/1 vacation systems with Bernoulli vacations as presented here. While queue length and waiting time distributions are not easily obtained, the general structure for M/GI/1/L vacation systems with Markov schedules allows, as is shown, development of queue length pgf at arbitrary times, embedded at departures and, embedded at vacation completions. Further, the waiting time Laplace-Stieltjes transform is easily obtained since the pgf of queue length as seen at arbitrary times is known.. Each of these transform results is important to performance analysis of M/GI/1 vacation systems with Bernoulli schedules since moments of the respective distributions can be calculated in the usual manner. However, no distribution moments are presented here.

3.2 M/GI/1 vacation systems with E-limited service.

Consider now the M/GI/1 vacation system with E-limited service introduced in Chapter 1. Recall that the "E-limited" server scheduling discipline requires that the server begins a vacation when either a prespecified number m^* of customers are served or the system is emptied, whichever occurs first. If the server returns from vacation to find the to find the queue empty, then another vacation begins immediately; the server continues in this manner until upon return from vacation, at least one customer is queued.

In this vacation system, it is assumed that the lengths of customer service periods are independent, identically distributed random variables having distribution S(t), and the lengths of server vacation periods are independent , identically distributed random variables having distribution V(t). Further, the lengths of service periods and vacation periods are assumes mutually independent. The Poisson stream of customers arriving to the system is assumed to have rate λ.

The queueing behavior of the M/GI/1 vacation system with E-limited service is to be examined within the general framework of vacation systems having Markov schedules. Thus, it is necessary to show that E-limited service is a server scheduling discipline belonging to the class of Markov schedules. That is, E-limited service must satisfy Conditions 1, 2, and 3 in order to be examined as a particularization the general model developed in Chapter 2. However before verifying that Conditions 1, 2, and 3 hold, it is helpful to re-examine the mark space of the server switching marked point process (H,T).

When examining M/GI/1 vacation systems with E-limited service, the full generality of the model introduced in Chapter 2 is not required. In particular, let the mark space $\hat{E}$ of the server switching marked point process (H,T) be restricted to the set

$$\hat{E} = (\{s\} \times \{1, 2, \ldots, m^*\}) \cup \{v\}.$$

This simplification of the mark space is convenient since the number of consecutive v-type epochs of (H,T) does not influence the server scheduling activity under E-limited service.

Let the joint queue length / server activity process X_{R_+} and the queue length / server activity marked point process (X,T) be defined as in Chapter 2. Since, in the system under consideration, the queue capacity is infinite (i.e., $L = \infty$), we have that the state space E is given by

$$E = \hat{E} \times Z^+.$$

Recall that $i,j \in E$ are vector quantities that consist of queue length and server activity components where,

$$i = (i_N, i_H) \text{ and } j = (j_N, j_H)$$

with,

$$i_N, j_N \in Z^+ \text{ and } i_H, j_H \in \hat{E}.$$

E-limited service requires that $i_H, j_H \in \hat{E}$ be two-tuples whenever $i,j \in E$ correspond to service completion epochs. That is, whenever $i,j \in E$ correspond to service completion epochs,

$$i_H = \left(i_{H_T}, i_{H_C}\right)$$

and

$$j_H = (j_{H_T}, j_{H_C})$$

where

$i_{H_T}, j_{H_T} = s$ and $i_{H_C}, j_{H_C} \in \{1,2,\ldots,m^*\}$ count the number of consecutive epochs of that are of s-type.

It is clear from the description of M/GI/1 vacation systems with E-limited service that the server's activity at the next epoch of (X,T) depends only on the queue length at the present epoch, the type of the present epoch, and the number of consecutive epochs of the same type up to and including the present epoch. Thus, for all $i,j \in E$

$$P\{H_{m+1} = j_H | X_0,\ldots,X_m,T_0,\ldots,T_m\} = P\{H_{m+1} = j_H | X_m\},$$

which satisfies Condition 1. Now, for all $i,j \in E$, let g(i,j) be defined as in (2.2). It follows that for all $m \in Z^+$,

$$g(i,j) = \begin{cases} 1, & j_{H_T} = s, j_{H_C} = i_{H_T} + 1, i_N \geq 1, i_{H_T} = s, i_{H_C} \leq m^* - 1 \\ & \text{or} \\ & j_{H_T} = s, j_{H_C} = 1, i_N \geq 1, i_H = v \\ & \text{or} \\ & j_H = v, i_N = 0, i_H = v \\ & \text{or} \\ & j_H = v, i_N \geq 0, i_{H_T} = s, i_{H_C} = m^* \\ 0, & \text{otherwise} \end{cases}$$

(3.78)

It is known that for all vacation systems, the server is either on vacation or is serving

customers. From the description of E-limited service, it follows that the time between any two contiguous epochs of of the queue length / server activity marked point process must be either a customer service period distributed S(t) or a server vacation period distributed V(t), both conditional upon the server's activity at the more recent of the two epochs, the number of customers queued at the older of the two epochs, and the server's activity at the older of the two epochs. That is , customer service periods and server vacation periods are such that for all $m \in Z^+$,

$$P\{T_{m+1} - T_m \le t \,|\, H_{m+1}, X_0, \ldots, X_m, T_0, \ldots, T_m\} = P\{T_{m+1} - T_m \le t \,|\, H_{m+1}, X_m\}$$

whenever, in the Markov chain X, state j is "one-step reachable" from state i where $i,j \in E$. Hence, E-limited service satisfies Condition 2. Following the notation of (2.2), we have for all $i,j \in E$ and $t \in R^+$ that

$$F(i,j,t) = \begin{cases} S(t), & j_H = s, g(i,j) = 1 \\ V(t), & j_H = s, g(i,j) = 1 \\ 0, & \text{otherwise} \end{cases} . \qquad (3.79)$$

Recognizing that the customer arrival stream is Poisson, it is clear that the customer interarrival times are exponentially distributed and thus have the memoryless property. It follows that for all $i,j \in E$, $m \in Z^+$, and $t \in R^+$,

$$\begin{aligned} P\{N_{m+1} = j_N \,|\, H_{m+1}, X_0, \ldots, X_m, T_0, \ldots, T_m, T_{m+1} - T_m = t\} \\ = P\{N_{m+1} = j_N \,|\, H_{m+1} = j_H, X_m = i, T_{m+1} - T_m = t\} \end{aligned}$$

which indicates that E-limited service satisfies Condition 3. Following the notation of (2.3) we have that

$$G(i,j,t) = \begin{cases} \dfrac{e^{-\lambda t}(\lambda t)^{j_N - i_N + 1}}{(j_N - i_N + 1)!}, & j_N \geq i_N - 1, j_{H_T} = s, i_N \geq 1 \\ \dfrac{e^{-\lambda t}(\lambda t)^{j_N - i_N}}{(j_N - i_N)!}, & j_N \geq i_N, j_H = v, i_N \geq 0 \\ 0, & \text{otherwise} \end{cases} . \qquad (3.80)$$

It is clear that since Conditions 1, 2, and 3 are satisfied for this example system that E-limited service belongs to the class of Markov schedules. Thus, by Proposition 2.3 the M/GI/1 vacation system with E-limited service has a queue length / server activity process (X,T) that is Markov renewal and has a joint queue length server activity process X_{R^+} that is semi-regenerative. It is now a simple matter to calculate the semi-Markov kernel Q(t) associated with (X,T). Substituting (3.78), (3.79), and (3.80) into (3.4) it follows that

$$
Q(i,j,t) = \begin{cases}
\displaystyle\int_0^t \frac{(\lambda u)^{j_N - i_N}}{(j_N - i_N)!} e^{-\lambda u} V(du), \\
\qquad j_N \geq i_N, j_H = v, i_N = 0, i_{H_T} = s, i_{H_c} < m^* \\
\qquad\qquad \text{or} \\
\qquad j_N \geq i_N, j_H = v, i_N > 0, i_{H_T} = s, i_{H_c} = m^* \\
\qquad\qquad \text{or} \\
\qquad j_N = 0, j_H = v, i_N = 0, i_H = v \\
\displaystyle\int_0^t \frac{(\lambda u)^{j_N - i_N + 1}}{(j_N - i_N + 1)!} e^{-\lambda u} S(du), \\
\qquad j_N \geq i_N - 1, j_{H_T} = s, j_{H_c} = i_{H_c} + 1, i_N > 0, i_{H_T} = s, i_{H_c} < m^* \\
\qquad\qquad \text{or} \\
\qquad j_N \geq i_N - 1, j_{H_T} = s, j_{H_c} = 0, i_N > 0, i_H = v \\
0, \qquad \text{otherwise}
\end{cases}
$$

(3.81)

Having particularized the semi-Markov kernel Q(t) to reflect M/GI/1 vacation systems with E-limited service, it is feasible to investigate the ergodic queueing behavior of this system. (Recall that the queue length / server activity marked point process (X,T) is assumed irreducible and that all states are recurrent.) In particular, the ergodic queue length distribution as seen immediately following customer service completions and the ergodic queue length distribution as seen by the server immediately following returns from vacation are investigated. Nonergodic results are not considered.

From Proposition 2.4 we have that the marked process X associated with the queue length / server activity marked point process (X,T) forms a Markov chain. Let Q be the

collection of one-step transition probabilities for the Markov chain X. Here, $Q = \lim_{t \to \infty} Q(t)$. Now, for $\alpha \in \{1,2,\ldots,m^*\}$, let

$$S_\alpha = \{i \in E : i_{H_T} = s, i_{H_C} = \alpha\}$$

and let

$$V = \{i \in E : i_H = v\}.$$

Clearly, V and S_α, $\alpha = 1,2,\ldots,m^*$, partition the state space E. The equations yielding the stationary distribution on X, when partitioned as the state space E, are written in matrix form as

$$\left[\pi_{S_1}\ \pi_{S_2}\ \pi_{S_3}\ \ldots\ \pi_{S_{m^*-1}}\ \pi_{S_{m^*}}\ \pi_V\right]$$

$$= \left[\pi_{S_1}\ \pi_{S_2}\ \pi_{S_3}\ \ldots \pi_{S_{m^*-1}}\ \pi_{S_{m^*}}\ \pi_V\right] \cdot \begin{bmatrix} 0 & Q_{S_1S_2} & 0 & \ldots & 0 & 0 & Q_{S_1V} \\ 0 & 0 & Q_{S_2S_3} & \ldots & 0 & 0 & Q_{S_2V} \\ 0 & 0 & 0 & \ldots & 0 & 0 & Q_{S_3V} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & Q_{S_{m^*-1}S_{m^*}} & Q_{S_{m^*-1}V} \\ 0 & 0 & 0 & \ldots & 0 & 0 & Q_{S_{m^*}V} \\ Q_{VS_1} & 0 & 0 & \ldots & 0 & 0 & Q_{VV} \end{bmatrix}$$

(3.82)

where,

$$Q = \begin{bmatrix} 0 & Q_{s_1 s_2} & 0 & \cdots & 0 & 0 & Q_{s_1 v} \\ 0 & 0 & Q_{s_2 s_3} & \cdots & 0 & 0 & Q_{s_2 v} \\ 0 & 0 & 0 & \cdots & 0 & 0 & Q_{s_3 v} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & Q_{s_{m^*-1} s_{m^*}} & Q_{s_{m^*-1} v} \\ 0 & 0 & 0 & \cdots & 0 & 0 & Q_{s_{m^*} v} \\ Q_{v s_1} & 0 & 0 & \cdots & 0 & 0 & Q_{vv} \end{bmatrix}$$

and $\left[\pi_{s_1}\ \pi_{s_2}\ \pi_{s_3} \cdots \pi_{s_{m^*-1}}\ \pi_{s_{m^*}}\ \pi_v\right]$ is the stationary distribution on X. Here, for $\alpha \in \{1,2,\ldots,m^*\}$,

$$\pi_{s_\alpha} = \left[\pi_{s_\alpha}(0)\ \pi_{s_\alpha}(1)\ \pi_{s_\alpha}(2) \ldots\right]$$

where,

$$\pi_{s_\alpha}(j) = \lim_{m \to \infty} P\{N_m = j, H_m = (s,\alpha)\}, \quad j = 0,1,2,\ldots$$

Also,

$$\pi_v = \left[\pi_v(0)\ \pi_v(1)\ \pi_v(2) \ldots\right]$$

where,

$$\pi_v(j) = \lim_{m \to \infty} P\{N_m = j, H_m = v\}, \quad j = 0,1,2,\ldots$$

From (3.81) and (3.82) it follows that for $k = 1,2,\ldots, m^* - 1$

$$Q_{VS_1} = Q_{S_k S_{k+1}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \cdots \\ s_0 & s_1 & s_2 & s_3 & s_4 & \cdots \\ 0 & s_0 & s_1 & s_2 & s_3 & \cdots \\ 0 & 0 & s_0 & s_1 & s_2 & \cdots \\ 0 & 0 & 0 & s_0 & s_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{3.83}$$

$$Q_{VV} = Q_{S_k V} = \begin{bmatrix} v_0 & v_1 & v_2 & v_3 & v_4 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{3.84}$$

and

$$Q_{S_{m-1}V} = \begin{bmatrix} v_0 & v_1 & v_2 & v_3 & v_4 & \cdots \\ 0 & v_0 & v_1 & v_2 & v_3 & \cdots \\ 0 & 0 & v_0 & v_1 & v_2 & \cdots \\ 0 & 0 & 0 & v_0 & v_1 & \cdots \\ 0 & 0 & 0 & 0 & v_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{3.85}$$

where, as in (3.11) and (3.12), for j = 0, 1, 2, ... ,

$$s_j = \int_0^{\infty} \frac{(\lambda t)^j e^{-\lambda t}}{j!} S(dt) \tag{3.86}$$

and

$$v_j = \int_0^{\infty} \frac{(\lambda t)^j e^{-\lambda t}}{j!} V(dt) \tag{3.87}$$

Substituting (3.83), (3.84), and (3.85) into (3.82) it follows that for j = 0, 1, 2, ...

$$\pi_{s_1}(j) = \sum_{k=1}^{j+1} \pi_v(k) s_{j-k+1}, \tag{3.88}$$

$$\pi_{s_{i+1}}(j) = \sum_{k=1}^{j+1} \pi_{s_i}(k) s_{j-k+1} \qquad i = 1, 2, \ldots, m^* - 1, \tag{3.89}$$

and

$$\pi_v(j) = \sum_{k=0}^{j} \pi_{s_{m^*}}(k) v_{j-k} + v_j \left(\pi_v(0) + \sum_{k=1}^{m^*-1} \pi_{s_k}(0) \right) \tag{3.90}$$

At this juncture it is convenient to introduce the ergodic probability $\pi_s(j)$ that the queue length is j and the epoch is s-type at the epochs of the *(X,T)*. Here,

$$\pi_s(j) = \lim_{m \to \infty} P\{N_m = j, H_m = s\}.$$

Clearly, $\pi_s(j)$ is the marginal probability given by

$$\pi_s(j) = \sum_{i=1}^{m^*} \pi_{s_i}(j), \qquad j = 0, 1, 2, \ldots \tag{3.91}$$

Now, define the following geometric transforms:

$$\Pi_{s_i}(z) = \sum_{j=0}^{\infty} \pi_{s_i}(j) z^j, \qquad i = 1, 2, \ldots, m^*, \tag{3.92}$$

$$\Pi_v(z) = \sum_{j=0}^{\infty} \pi_v(j) z^j, \tag{3.93}$$

and

$$\Pi_s(z) = \sum_{j=0}^{\infty} \pi_s(j) z^j, \tag{3.94}$$

Substituting (3.94) into (3.91), it follows that

$$\Pi_s(z) = \sum_{i=1}^{m^*} \Pi_{s_i}(z) \tag{3.95}$$

Substituting (3.88), (3.89) into (3.92), and substituting (3.90) into (3.93) respectively give

$$\Pi_{s_1}(z) = \sum_{j=0}^{\infty} z^j \sum_{k=1}^{j+1} \pi_v(k) s_{j-k+1}, \tag{3.96}$$

$$\Pi_{s_{i+1}}(z) = \sum_{j=0}^{\infty} z^j \sum_{k=1}^{j+1} \pi_{s_i}(k) s_{j-k+1} \qquad i = 1, 2, \ldots, m^* - 1 \tag{3.97}$$

and

$$\Pi_v(z) = \sum_{j=0}^{\infty} z^j \left(\sum_{k=0}^{j} \pi_{s_{m^*}}(k) v_{j-k} + v_j \left(\pi_v(0) + \sum_{k=1}^{m^*-1} \pi_{s_k}(0) \right) \right) \tag{3.98}$$

Interchanging the order of summation in the usual manner, (3.96), (3.97), and (3.98) can be rewritten respectively as

$$\Pi_{S_1}(z) = \frac{\tilde{S}(z)}{z}\left(\Pi_V(z) - \pi_V(0)\right), \tag{3.99}$$

$$\Pi_{S_{i+1}}(z) = \frac{\tilde{S}(z)}{z}\left(\Pi_{S_i}(z) - \pi_{S_i}(0)\right), \qquad i = 1,2,\ldots, m^* - 1 \tag{3.100}$$

and

$$\Pi_V(z) = \tilde{V}(z)\Pi_{S_{m^*}}(z) + \tilde{V}(z)\left(\pi_V(0) + \sum_{i=1}^{m^*-1}\pi_{S_i}(0)\right) \tag{3.101}$$

where, $\tilde{S}(z)$ and $\tilde{V}(z)$ are given by (3.25) and (3.26) respectively.

Substituting (3.99) and (3.100) into (3.95), it follows that

$$\Pi_S(z) = \frac{z}{\tilde{S}(z)}\left(\Pi_S(z) - \Pi_{S_1}(z)\right) + \frac{1}{\tilde{V}(z)}\Pi_V(z) - \pi_V(0) \tag{3.102}$$

Rearranging (3.96) and substituting into (3.102)yields

$$\Pi_S(z) = \frac{\tilde{S}(z)}{z - \tilde{S}(z)}\Pi_V(z) \tag{3.103}$$

Solving (3.99), (3.100), and (3.101) simultaneously for $\Pi_V(z)$ gives the transform relationship

$$\Pi_V(z) = \frac{\tilde{V}(z)}{\tilde{S}^{m^*}(z)\tilde{V}(z) - z^{m^*}}\left(\pi_V(0)\left(\tilde{S}^{m^*}(z) - z^{m^*}\right) + \sum_{i=1}^{m^*-1} z^i \pi_{S_i}(0)\left(\tilde{S}^{m^*-i}(z) - z^{m^*-i}\right)\right) \tag{3.104}$$

Now, substituting (3.104) into (3.103) it follows that

$$\Pi_S(z) = \frac{\tilde{V}(z)\tilde{S}(z)}{(z - \tilde{S}(z))\left(\tilde{S}^{m^*}(z)\tilde{V}(z) - z^{m^*}\right)} \cdot \left(\pi_V(0)\left(\tilde{S}^{m^*}(z) - z^{m^*}\right) + \sum_{i=1}^{m^*-1} z^i \pi_{S_i}(0)\left(\tilde{S}^{m^*-i}(z) - z^{m^*-i}\right)\right) \tag{3.105}$$

Note that while (3.104) and and (3.105) are geometric transforms, neither is necessarily a pgf. The respective values taken by the constants $\pi_V(0), \pi_{S_1}(0), \pi_{S_2}(0), \ldots, \pi_{S_{m^*-1}}(0)$ determine whether or not $X_S(z)$ or $X_V(z)$ is a pgf. The value of the constants $\pi_V(0), \pi_{S_1}(0), \pi_{S_2}(0), \ldots, \pi_{S_{m^*-1}}(0)$ required for $X_S(z)$ to become a pgf are generally different than the values required for $X_V(z)$ to become a pgf. Thus, interpretation of these constants must necessarily be considered within the context of the pgf in which they appear.

It is now a routine matter to examine both the pgf of the queue length as seen by customers immediately following departure from the system and the pgf of the queue length as seen by the server immediately following returns from vacation. Here, we have that when the constants $\pi_V(0), \pi_{S_1}(0), \pi_{S_2}(0), \ldots, \pi_{S_{m^*-1}}(0)$ are chosen such that $\lim_{z \uparrow 1} \Pi_S(z) = 1$, then the geometric transform $\Pi_S(z)$ becomes the pgf for queue length as seen by customers immediately following departure from the system. Similarly, we have that when the constants $\pi_V(0), \pi_{S_1}(0), \pi_{S_2}(0), \ldots, \pi_{S_{m^*-1}}(0)$ are chosen such that

$\lim_{z \uparrow 1} \Pi_V(z) = 1$, then the geometric transform $\Pi_V(z)$ becomes the pgf for the queue length as seen by the server immediately following returns from vacation. The fact that $\Pi_S(z)$ and $\Pi_V(z)$ are pgf's under the above specified boundary conditions follows from the same reasoning as was presented for the M/GI/1 vacation system with Bernoulli schedules; hence, a formal development of this results is not presented here.

The task of determining values for the constants $\pi_V(0), \pi_{S_1}(0), \pi_{S_2}(0), \ldots, \pi_{S_{m^*-1}}(0)$ such that $\Pi_S(z)$ and $\Pi_V(z)$ become pgf's is lengthy; application of Rouche's theorem and Lagrange's theorem lead to a set of m^* simultaneous equations that can be solved for the appropriate values of $\pi_V(0), \pi_{S_1}(0), \pi_{S_2}(0), \ldots, \pi_{S_{m^*-1}}(0)$. The application of Rouche's theorem in order to determine unknown constants for pgf's, similar in form to $\Pi_S(z)$ and $\Pi_V(z)$, is common within the queueing literature Takagi (1987); as a matter of convenience, a development of formulae suitable for determining the desired values of $\pi_V(0), \pi_{S_1}(0), \pi_{S_2}(0), \ldots, \pi_{S_{m^*-1}}(0)$ is not presented here. Rather, we are satisfied that (3.104) is the pgf of the queue length as seen by the server immediately following returns from vacation, and that (3.105) is the pgf of the queue length as seen immediately following customer departures. We have here that (3.105) agrees with Takagi (1987). To the author's knowledge, (3.104) does not appear in the published literature and thus is new.

Appealing to the discussion offered in the previous section, we may easily examine the pgf of queue length as seen at arbitrary times. Note that since the M/GI/1 vacation system with E-limited service has a Poisson arrival stream and customers served one at a time, the queue length as seen by customers immediately following departure from the system and

the queue length as seen at arbitrary times are distributed the same. Since geometric transform pairs are unique, it follows that the pgf of the queue length as seen at arbitrary times is given by (3.105).

Let T be the ergodic waiting time in the system for an arbitrary customer, and let $W(\sigma) = E[e^{-\sigma T}]$ be the Laplace-Stieltjes transform of the waiting time T. Clearly, the M/GI/1 vacation system with E-limited service satisfies the conditions of Proposition 3.1; thus, the distribution form of Little's law can be employed to provide an expression for $W(\sigma)$. Following the same reasoning as was considered for M/GI/1 vacation systems with Bernoulli schedules we have that

$$W(\sigma) = \Pi_S\left(1 - \frac{\sigma}{\lambda}\right). \tag{3.106}$$

The waiting time Laplace-Stieltjes transform of (3.106) concludes the analysis of the M/GI/1 vacation system with E-limited service as considered here. While the queue length and waiting time distributions are not easily obtained, the general structure of the M/GI/1/L vacation system with Markov schedules, as is shown, allows development of pgf's for the queue length as seen at arbitrary times, as seen at customer departures, and as seen at vacation completions. Further, the distributional form of Little's law holds for this system allowing the Laplace-Stieltjes transform of customer waiting time in the system to be formulated from the pgf of queue length as seen at arbitrary times. Each of these transform results is important to the performance analysis of the M/GI/1 vacation system with E-limited service since moments of their respective distributions can be calculated in the usual manner.

The M/GI/1 vacation system with E-limited service considered above serves to

demonstrate that the general model of Chapter 2 can be particularized to reflect systems having server scheduling disciplines more sophisticated than Bernoulli schedules. The complexity of E-limited service is reflected by the mark space $\hat{E}$ of the server switching marked point process. It is worth noting that while $\hat{E}$ is more complicated under E-limited service than it is under Bernoulli schedules, the procedures for developing queue length pgf's for these two server scheduling disciplines are remarkably similar. This similarity is a happy benefit of our analyses originating from the common stochastic framework developed in Chapter 2.

It would seem natural to, at this juncture, investigate special cases of M/GI/1 vacation systems with E-limited service. Clearly, setting $m^* = \infty$ indicates *exhaustive service* while setting $m^* = 1$ indicates simple *limited service*. Recall that exhaustive service was investigated in a straight forward manner as a special case of Bernoulli schedules. Investigating exhaustive service as a special case of E-limited service requires development of limiting arguments relative to m^*. Such limiting arguments are obviously unnecessary to investigate exhaustive service and are considered beyond the scope of this work.

Investigating limited service as a special case of E-limited service is a relatively simple matter Takagi (1987). However, limited service also appears as a special case of *limited batch service* and thus, is developed in the section to follow.

3.3 The M/GI/1 vacation system with limited batch service.

We now introduce the M/GI/1 vacation system with *limited batch service*. To the author's knowledge, the limited batch service is not investigated in the available literature;

thus, the analysis here is thought to be new. This system is a modification of the M/GI/1 vacation system with limited service where customers are served in "batches" of a fixed size k^*. The server scheduling discipline for a system with limited batch service requires that the server begin a vacation period following the completion of service for each batch of customers. If upon return from vacation the server finds fewer than k^* customers queued (i.e., an incomplete batch), another vacation begins immediately. The server continues to operate in this manner until upon return from vacation there are at least k^* customers (i.e., at least one batch) queued.

When the server returns from vacation to find at least k^* customers queued, he begins service on the batch of customers formed by the first k^* customers in line. That is, batches are served in order of arrival. Clearly, limited batch service reduces to the simple *limited service* server scheduling discipline introduced in Chapter 1 when $k^*=1$.

For the M/GI/1 vacation system with limited batch service, it is assumed that the lengths of batch service periods are independent, identically distributed with distribution B(t) and the lengths of the server vacation periods are independent, identically distributed with distribution V(t). It is further assumed that the lengths of batch service periods and the lengths of server vacation periods are mutually independent and are independent of the arrival process. The Poisson stream of customers arriving to the system is assumed to have rate λ.

Since the queueing behavior is the M/GI/1 vacation system with limited batch service is to be examined within the framework of the previous chapter, it is necessary to show that the limited batch service server scheduling discipline belongs to the class of Markov schedules. That is, limited batch service must satisfy Conditions 1, 2, and 3. However, before verifying that Conditions 1, 2, and 3 hold, it is helpful to reexamine the mark space of the server switching marked point process (H,T).

When examining the M/GI/1 vacation system with limited batch service, the full generality of the model developed in Chapter 2 is not required. In particular, let the mark space $\hat{E}$ be be restricted such that $\hat{E} = F$. Under limited batch service this simplification of the mark space is convenient since for any given epoch of (H,T), the type (s-type or v-type) of this epoch depends only upon the most recent previous epoch.

Let the joint queue length / server activity process X_R. and the queue length / server activity marked point process be defined as in Chapter 2. Since the queue capacity here is infinite (i.e., $L = \infty$), it follows that the state space E for the M/GI/1 vacation system with limited batch service is given by

$$E = \hat{E} \times Z^+ = F \times Z^+$$

Recall that $i,j \in E$ are expressible in terms of queue length and server activity components where,

$$i = (i_N, i_H) \text{ and } j = (j_N, j_H)$$

with

$$i_N, j_N \in Z^+ \text{ and } i_H, j_H \in F.$$

Under limited batch service,the type of the next epoch of (X,T) depends only upon the present epoch of (X,T). Thus, it follows that for all $i,j \in E$ and $m \in Z^+$,

$$P\{H_{m+1} = j_H | X_0, \ldots, X_m, T_0, \ldots, T_m\} = P\{H_{m+1} = j_H | X_m\}$$

which implies that limited batch service satisfies Condition 1. Let g(i,j) be defined is in (2.1). It follows that

$$g(i,j) = \begin{cases} 1, & j_H = s,\ i_N \geq k^*,\ i_H = v \\ & \text{or} \\ & j_H = v,\ i_N < k^*,\ i_H = v \\ & \text{or} \\ & j_H = v,\ i_N \in Z^+,\ i_H = s \\ 0, & \text{otherwise} \end{cases} \tag{3.107}$$

Here, we have that the server is either serving a batch of customers or is on vacation. For limited batch service, the time between any two contiguous epochs of (X,T) must correspond to either a batch service period with distribution B(t) or a server vacation period with distribution V(t). Given two contiguous, the server's activity in the period between them is recognized as either a vacation period or a service period conditioned upon the type of the more recent of the two epochs and the number of customers queued at the older of the two epochs. That is, both batch service periods and server vacation periods are such that for all $m \in Z^+$,

$$P\{T_{m+1} - T_m \leq t \mid H_{m+1}, X_0, \ldots, X_m, T_0, \ldots, T_m\} = P\{T_{m+1} - T_m \leq t \mid H_{m+1}, X_m\}$$

Thus, limited batch service satisfies Condition 2. Following the notation of (2.2) we have that for all $i,j \in E$, $m \in Z^+$, and $t \in R^+$

$$F(i,j,t) = \begin{cases} B(t), & j_H = s,\ i \in E,\ g(i,j) = 1 \\ V(t), & j_H = v,\ i \in E,\ g(i,j) = 1 \\ 0, & \text{otherwise} \end{cases} . \tag{3.108}$$

Since the customer arrival stream is Poisson, it is clear that the interarrival times are

exponentially distributed and, thus, have the memoryless property. Hence, it follows that for all $i,j \in E$, $m \in Z^+$, and $t \in R^+$,

$$\begin{aligned} P\{N_{m+1} = j_N | H_{m+1}, X_0, \ldots X_m, T_0, \ldots T_m, T_{m+1} - T_m = t\} \\ = P\{N_{m+1} = j_N | H_{m+1}, X_m, T_{m+1} - T_m = t\} \end{aligned}$$

which implies that limited batch service satisfies Condition 3. Following the notation of (2.3), G(i,j,t) is given by

$$G(i,j,t) = \begin{cases} \dfrac{(\lambda t)^{j_N - i_N + k^*} e^{-\lambda t}}{(j_N - i_N + k^*)!}, & j_N \geq i_N - k^*,\ j_H = s,\ i_N \geq k^*,\ i_H = v \\ \dfrac{(\lambda t)^{j_N - i_N} e^{-\lambda t}}{(j_N - i_N)!}, & j_N \geq i_N,\ j_H = v,\ i_N \in Z^+,\ i_H \in F \\ 0, & \text{otherwise} \end{cases} \quad . \quad (3.109)$$

Since Conditions 1, 2, and 3 are satisfied for this example, limited batch service belongs to the class of Markov schedules. Thus, by Proposition 2.3, any M/GI/1 vacation system with limited batch service, has a Markov renewal queue length / server activity marked point process (X,T) and has a semi-regenerative joint queue length / server activity process X_{R^+}. Substituting (3.107), (3.108), and (3.109) into (3.4) it follows that the semi-Markov kernel Q(t) corresponding to (X,T) is given by

$$Q(i,j,t) = \begin{cases} \int_0^t \frac{(\lambda u)^{j_N - i_N + k^*}}{(j_N - i_N + k^*)!} e^{-\lambda u} B(du), & j_N \geq i_N - k^*,\ j_H = s,\ i_N \in Z^+,\ i_H = v \\ \int_0^t \frac{(\lambda u)^{j_N - i_N}}{(j_N - i_N)!} e^{-\lambda u} V(du), & j_N \geq i_N,\ j_H = v,\ i_N \in Z^+,\ i_H = \in F \\ 0, \quad \text{otherwise} \end{cases} \tag{3.110}$$

for all $i,j \in E$.

Having particularized the semi-Markov kernel Q(t) to model M/GI/1 vacation systems with limited batch service, it is now feasible to investigate ergodic queueing behavior of such systems. (Recall that ergodic results exist here since it is assumed that (X,T) is irreducible and that all states of E are recurrent.) In particular, the ergodic distribution of queue length as seen immediately following batch service completions, the ergodic queue length distribution as seen immediately following the server's returns from vacation, and the ergodic queue length distribution as seen at arbitrary times are investigated. Nonergodic results are not considered in what follows.

From Proposition 2.4, we have that the marked process X associated the queue length / server activity marked point process (X,T) forms a Markov chain. Since X is embedded at all batch service completions and all server vacation completions, the ergodic distribution of queue length as seen at these epochs is simply the stationary distribution for the Markov chain X. Note that by Corollary 2.13, the ergodic distribution of queue length as seen at arbitrary times requires the stationary distribution for X. Hence, the three queue length distribution of interest for system performance analysis each require that a stationary

measure on X be calculated.

Since here the state space E is countably infinite, solving for a stationary measure on X is formidable. Thus, in the development to follow pgf's corresponding to the three queue length distributions of interest are examined.

Let Q be the collection of one-step transition probabilities associated with the Markov chain X. Here, $Q = \lim_{t \to \infty} Q(t)$. Let the state space E be partitioned as is Corollary 2.8, such that $E = S \cup V$ and $S \cap V = \varnothing$ where,

$$S = \{i \in E : i_H = s\}$$

and

$$V = \{i \in E : i_H = v\}.$$

The equations yielding the stationary distribution on X, when partitioned according to the state space partition described above, are written in matrix form as

$$[\pi_s \; \pi_v] = [\pi_s \; \pi_v] \begin{bmatrix} Q_{ss} & Q_{sv} \\ Q_{vs} & Q_{vv} \end{bmatrix} \tag{3.111}$$

where,

$$\pi_\alpha(j) = \lim_{t \to \infty} P\{n_t = j, \; h_t = \alpha\}.$$

It follows from (3.110) that

$$Q = \begin{bmatrix} Q_{ss} & Q_{sv} \\ Q_{vs} & Q_{vv} \end{bmatrix}$$

and $[\pi_s\ \pi_v]$ is the stationary distribution on X. Here, for $\alpha \in F$,

$$\pi_\alpha = [\pi_\alpha(0)\ \pi_\alpha(1)\ \pi_\alpha(2)\ \ldots]$$

where,

$$Q_{ss} = [0], \tag{3.112}$$

$$Q_{sv} = \begin{bmatrix} v_0 & v_1 & v_2 & \cdots & v_{k^\bullet-1} & v_{k^\bullet} & v_{k^\bullet+1} & \cdots \\ 0 & v_0 & v_1 & \cdots & v_{k^\bullet} & v_{k^\bullet+1} & v_{k^\bullet+2} & \cdots \\ 0 & 0 & v_0 & \cdots & v_{k^\bullet+1} & v_{k^\bullet+2} & v_{k^\bullet+3} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & \cdots & v_0 & v_1 & v_2 & \cdots \\ 0 & 0 & 0 & \cdots & 0 & v_0 & v_1 & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & v_0 & \cdots \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{3.113}$$

$$Q_{vs} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots \\ s_0 & s_1 & s_2 & \cdots & s_{k^\bullet-1} & s_{k^\bullet} & s_{k^\bullet+1} & \cdots \\ 0 & s_0 & s_1 & \cdots & s_{k^\bullet-2} & s_{k^\bullet-1} & s_{k^\bullet} & \cdots \\ 0 & 0 & s_0 & \cdots & s_{k^\bullet-3} & s_{k^\bullet-2} & s_{k^\bullet-1} & \cdots \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{3.114}$$

and

$$Q_{VV} = \begin{bmatrix} v_0 & v_1 & v_2 & \cdots & v_{k^*-1} & v_{k^*} & v_{k^*+1} & \cdots \\ 0 & v_0 & v_1 & \cdots & v_{k^*} & v_{k^*+1} & v_{k^*+2} & \cdots \\ 0 & 0 & v_0 & \cdots & v_{k^*+1} & v_{k^*+2} & v_{k^*+3} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & \cdots & v_0 & v_1 & v_2 & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{3.115}$$

where, for j = 0, 1, 2, ...,

$$s_j = \int_0^\infty \frac{(\lambda t)^j e^{-\lambda t}}{j!} B(dt) \tag{3.116}$$

and

$$v_j = \int_0^\infty \frac{(\lambda t)^j e^{-\lambda t}}{j!} V(dt). \tag{3.117}$$

Substituting (3.112) through (3.115) into (3.111), we find that for j = 0, 1, 2, ...

$$\pi_s(j) = \sum_{k=k^*}^{k^*+j} \pi_v(k) s_{j-k+k^*}. \tag{3.118}$$

and

$$\pi_v(j) = \gamma(j) + \sum_{k=0}^{j} \pi_s(k) v_{j-k} \tag{3.119}$$

where,

$$\gamma(j) = \begin{cases} \sum_{k=0}^{j} \pi_V(k)\, v_{j-k}, & j < k^* \\ \sum_{k=0}^{k^*} \pi_V(k)\, v_{k^*-k+j}, & j \geq k^*. \end{cases}$$

Now, define the following geometric transforms:

$$\Pi_S(z) = \sum_{j=0}^{\infty} \pi_S(j) z^j \tag{3.120}$$

and

$$\Pi_V(z) = \sum_{j=0}^{\infty} \pi_V(j) z^j. \tag{3.121}$$

Substituting (3.118) into (3.120) gives

$$\Pi_S(z) = \sum_{j=0}^{\infty} z^j \left(\sum_{k=k^*}^{k^*+j} \pi_V(k)\, s_{j-k+k^*} \right) \tag{3.122}$$

while substituting (3.119) into (3.121) gives

$$\Pi_V(z) = \sum_{j=0}^{\infty} z^j \left(\gamma(j) + \sum_{k=0}^{j} \pi_S(k)\, v_{j-k} \right). \tag{3.123}$$

Interchanging the order of summations respectively in (3.122) and (3.123) in the usual way allows the geometric transforms $\Pi_S(z)$ and $\Pi_V(z)$ to be written as

$$\Pi_S(z) = \frac{\tilde{B}(z)}{z^{k^*}}\Pi_V(z) - \frac{\tilde{B}(z)}{z^{k^*}}\sum_{j=0}^{k^*-1}\pi_V(j)z^j \tag{3.124}$$

and

$$\Pi_V(z) = \tilde{V}(z)\Pi_S(z) + \tilde{V}(z)\sum_{j=0}^{k^*-1}\pi_V(j)\,z^j \tag{3.125}$$

where, $\tilde{S}(z)$ and $\tilde{V}(z)$ are given by

$$\tilde{B}(z) = \sum_{j=0}^{\infty} z^j \int_0^{\infty} \frac{(\lambda t)^j e^{-\lambda t}}{j!} B(dt) = \int_0^{\infty} e^{-(\lambda - \lambda z)t} B(dt)$$

and

$$\tilde{V}(z) = \sum_{j=0}^{\infty} z^j \int_0^{\infty} \frac{(\lambda t)^j e^{-\lambda t}}{j!} V(dt) = \int_0^{\infty} e^{-(\lambda - \lambda z)t} V(dt)$$

Solving (3.124) and (3.125) simultaneously yields

$$\Pi_S(z) = \frac{\tilde{B}(z)\left(\tilde{V}(z) - 1\right)}{z^{k^*} - \tilde{B}(z)\tilde{V}(z)}\sum_{j=0}^{k^*-1}\pi_V(j)z^j \tag{3.126}$$

and

$$\Pi_V(z) = \frac{\tilde{V}(z)\left(z^{k^*} - \tilde{B}(z)\right)}{z^{k^*} - \tilde{B}(z)\tilde{V}(z)}\sum_{j=0}^{k^*-1}\pi_V(j)\,z^j \tag{3.127}$$

Note that while (3.126) and (3.127) are geometric transforms, neither is necessarily a pgf. The values taken by the constants $\pi_V(0),\pi_V(1),\ldots,\pi_V(k^* - 1)$ determine whether or not either (3.126) or (3.127) defines a pgf. The values of $\pi_V(0),\pi_V(1),\ldots,\pi_V(k^* - 1)$ required to make (3.126) a pgf are generally different than those values required to make (3.127) a pgf.

Applying the same reasoning as was presented for the M/GI/1 vacation with Bernoulli schedules, we have that when $\pi_V(0),\pi_V(1),\ldots,\pi_V(k^* - 1)$ are chosen such that

$$1 = \lim_{z\uparrow 1} \frac{\tilde{B}(z)\left(\tilde{V}(z) - 1\right)}{z^{k^*} - \tilde{B}(z)\tilde{V}(z)} \sum_{j=0}^{k^*-1} \pi_V(j) z^j$$

then $\Pi_s(z)$ is the pgf for the queue length as seen immediately following batch service completions. Similarly, when $\pi_V(0),\pi_V(1),\ldots,\pi_V(k^* - 1)$ are chosen such that

$$1 = \lim_{z\uparrow 1} \frac{\tilde{V}(z)\left(z^{k^*} - \tilde{B}(z)\right)}{z^{k^*} - \tilde{B}(z)\tilde{V}(z)} \sum_{j=0}^{k^*-1} \pi_V(j) z^j$$

then $\Pi_V(z)$ is the pgf of the queue length as seen immediately following the server's returns from vacation. For the M/GI/1 vacation with limited batch service, determining values for the constants $\pi_V(0),\pi_V(1),\ldots,\pi_V(k^* - 1)$ such that $\Pi_s(z)$ and $\Pi_V(z)$ are pgf's is analogous to the task of determining constants for the pgf's developed for M/GI/1 vacation systems with E-limited service.

A routine application of Rouche's theorem to $\Pi_s(z)$ and $\Pi_V(z)$ yields a set of k^* simultaneous equations for each of the two pgf's. These two sets of k^* simultaneous

equations may be solved independently of one another for the respective values of $\pi_v(0),\pi_v(1),\ldots,\pi_v(k^* - 1)$ that are required for $\Pi_s(z)$ and $\Pi_v(z)$ to be pgf's. However, the focus of this section is directed towards showing that the M/GI/1/L vacation system with Markov schedules can be particularized so as to capture the queueing behavior of M/GI/1 vacation systems with limited batch service. As a matter of convenience we omit the extensive algebra required to obtain the sets of simultaneous equations described above, and let the pgf for the queue length as seen immediately following batch service completions and the pgf for the queue length as seen immediately following the server's return form vacations be written as in (3.126) and (3.127) respectively.

At this juncture, we are positioned to develop the pgf for queue length as seen at arbitrary times. Before developing this pgf , it is convenient to examine a special case of the M/GI/1 vacation system with limited batch service. While this special case is somewhat of an aside, the brief development to follow affords us the opportunity to, in part, verify (3.126) by particularizing to a simple system that appears in the literature.

Note that when $k^* = 1$ (i.e., batches consist of a single customer) we obtain the M/GI/1 vacation system with simple *limited service*, introduced in Chapter 1. It follows directly from (3.126) and (3.127) that, for limited service, the pgf for the queue length as seen immediately following customer service completions and the pgf as seen immediately following the server's return from vacation are given respectively by

$$\Pi_s(z) = \frac{\tilde{S}(z)\left(\tilde{V}(z) - 1\right)}{z - \tilde{S}(z)\tilde{V}(z)}\pi_v(0) \tag{3.128}$$

and

$$\Pi_v(z) = \frac{\tilde{V}(z)\left(z - \tilde{S}(z)\right)}{z - \tilde{S}(z)\tilde{V}(z)}\pi_v(0) \tag{3.129}$$

where, $\tilde{B}(z)$ is replaced by $\tilde{S}(z)$.

(Since we have defined both $\Pi_s(z)$ and $\Pi_v(z)$ to be pgf's, it follows from our previous reasoning that $\pi_v(0)$ takes different values in (3.128) and (3.129)

First, consider $\Pi_s(z)$ the pgf for the queue length as seen immediately following customer service completions . It is a simple matter to to evaluate the constant $\pi_v(0)$. Since we have that

$$\lim_{z \uparrow 1} \Pi_s(z) = 1,$$

an application of L'Hopital's rule shows that

$$1 = \lim_{z \uparrow 1} \frac{\pi_v(0)\left(\tilde{S}(z)\tilde{V}'(z) + \tilde{S}'(z)\left(\tilde{V}(z) - 1\right)\right)}{1 - \left(\tilde{S}(z)\tilde{V}'(z) + \tilde{S}'(z)\tilde{V}(z)\right)} \tag{3.130}$$

where the prime diacritical mark indicates differentiation with respect to z. Let $\bar{S}$ denote the expected length of a customer service period and $\bar{V}$ denote the expected length of a server vacation period. Recall that $\tilde{S}'(1) = \lambda\bar{S}$ and $\tilde{V}'(1) = \lambda\bar{V}$. Rearranging (3.130) gives

$$\pi_v(0) = \frac{1 - \rho - \lambda\bar{V}}{\lambda\bar{V}} \tag{3.131}$$

where, $\rho = \lambda\bar{S}$ defines the traffic intensity. Now substituting (3.131) into (3.130) gives

$$\Pi_s(z) = \frac{1-\rho-\lambda\bar{V}}{\lambda\bar{V}} \cdot \frac{\tilde{S}(z)(\tilde{V}(z)-1)}{z-\tilde{S}(z)\tilde{V}(z)}. \tag{3.132}$$

Now, consider $\Pi_v(z)$ the pgf of the queue length as seen immediately following the server's returns from vacation. With $\Pi_v(z)$ given as a generating function, it follows that

$$\lim_{z\uparrow 1} \Pi_v(z) = 1.$$

Thus, applying L'Hopital's rule we have that

$$1 = \lim_{z\uparrow 1} \frac{\pi_v(0)(\tilde{V}(z)(1-\tilde{S}'(z)) + \tilde{V}'(z)(z-\tilde{S}(z)))}{1-(\tilde{S}(z)\tilde{V}'(z) + \tilde{S}'(z)\tilde{V}(z))}. \tag{3.133}$$

It now follows from (3.133) that the value of $\pi_v(0)$ required for $\lim_{z\uparrow 1} \Pi_v(z) = 1$ is given by

$$\pi_v(0) = \frac{1-\rho-\lambda\bar{V}}{1-\rho}. \tag{3.134}$$

Substituting (3.134) into (3.129) we have that

$$\Pi_v(z) = \frac{1-\rho-\lambda\bar{V}}{1-\rho} \cdot \frac{\tilde{V}(z)(z-\tilde{S}(z))}{z-\tilde{S}(z)\tilde{V}(z)}. \tag{3.135}$$

Since the M/GI/1 vacation system with limited service requires that customers are served one at a time, we have the the queue length as seen immediately following customer service completions is distributed the same as the queue length as seen at arbitrary times. Thus, it follows that the pgf of the queue length as seen at arbitrary times is given by

(3.132). Further, under limited service, this system satisfies the conditions of Proposition 3.1 and the distributional form of Little's law holds. With $W(\sigma)$ defined as the Laplace-Stieltjes transform of the customer waiting time distribution, we have that

$$\Pi_s(z) = W(\lambda - \lambda z). \tag{3.136}$$

Thus, (3.136) together with (3.132) gives

$$W(\sigma) = \frac{1 - \rho - \lambda\bar{V}}{\bar{V}} \cdot \frac{S^*(\sigma)(1 - V^*(\sigma))}{\sigma + (S^*(\sigma)V^*(\sigma) - 1)} \tag{3.137}$$

where, as in (3.75) and (3.76), $S^*(\sigma)$ and $V^*(\sigma)$ are the Laplace-Stieltjes transforms of S(t) and V(t) respectively.

In the analysis of the M/GI/1 vacation system with limited service, (3.132) and (3.137) agree with results reported by Takagi (1987). The pgf for the queue length as seen immediately following the server's returns from vacation given by (3.135) appears to be new. At this juncture, discussion of the M/GI/1 vacation system with limited service is concluded and we return to the more general M/GI/1 vacation system with limited batch service.

The example vacation systems considered thus far serve customers in a one at at time fashion. For such systems, the pgf for the queue length as seen at arbitrary times is known to equal the pgf for the queue length as seen immediately following customer service completions. Thus, developing queue length pgf's for these vacation systems has required examining only the Markov chain X associated with the queue length / server activity marked point process (X,T). It is clear that when k^* is greater that 1 in limited batch

service systems, customers are not served in a one at a time fashion. Hence, for this system, we must appeal to Corollary 2.13 and the joint queue length / server activity process X_{R^+} in order to investigate the pgf for the queue length as seen at arbitrary times.

Let X_{R^+} and (X,T) be defined as in Theorem 2.12. Since (X,T) is here assumed to be irreducible and to have a stationary distribution, it follows from (2.22) that for all $j \in E$

$$\lim_{t\to\infty} P\{X_t = j\} = \frac{1}{\pi m}\sum_{k\in E}\pi(k)\int_0^\infty P\{X_t = j \mid T_1 > t, X_0 = k\}P\{T_1 > t \mid X_0 = k\}dt \tag{3.138}$$

where π is the stationary distribution on the Markov chain X. Here, some notation is introduced so as to simplify the development to follow. For all $j \in E$, let

$$\eta(j) = \lim_{t\to\infty} P\{X_t = j\}$$

and for all $i,j \in E$, let

$$B(i,j) = \int_0^\infty P\{X_t = j \mid T_1 > t, X_0 = k\}P\{T_1 > t \mid X_0 = k\}dt .$$

When the state space E is partitioned by the sets S and V as in Corollary 2.8, it follows from (3.138) that the stationary distribution of the joint queue length / server activity process X_{R^+} is given by

$$[\eta_S \; \eta_V] = \frac{1}{\pi m}[\pi_S \; \pi_V]\begin{bmatrix} B_{SS} & B_{SV} \\ B_{VS} & B_{VV} \end{bmatrix} \tag{3.139}$$

Here, for $\alpha \in F$,

$$\eta_\alpha = [\eta_\alpha(0)\ \eta_\alpha(1)\ \eta_\alpha(2)\ \ldots]$$

with

$$\eta_\alpha(j) = \lim_{t\to\infty} P\{n_t = j, h_t = \alpha\}, \quad j = 0,1,2,\ldots$$

and for all $\alpha, \beta \in F$, $i,j = 0,1,2,\ldots$

$$B_{\alpha\beta}(i,j) = \int_0^\infty P\{n_t = j, h_t = \beta \mid T_1 > t, N_0 = i, H_0 = \alpha\} \cdot P\{T_1 > t \mid N_0 = i, H_0 = \alpha\}dt \quad (3.140)$$

It is clear that whenever $T_1 > t$, it must be that $h_t = H_0$. Thus, for $i,j = 0,1,2,\ldots$,

$$B_{sv}(i,j) = B_{vs}(i,j) = 0.$$

It now follows from (3.140) that

$$B_{sv} = B_{vs} = [0]. \quad (3.141)$$

Further, (3.140) implies that

$$B_{ss} = \begin{bmatrix} a_0 & a_1 & a_2 & \cdots & a_{k^*-1} & a_{k^*} & a_{k^*+1} & a_{k^*+2} & \cdots \\ 0 & a_0 & a_1 & \cdots & a_{k^*-2} & a_{k^*-1} & a_{k^*} & a_{k^*+1} & \cdots \\ 0 & 0 & a_0 & \cdots & a_{k^*-3} & a_{k^*-2} & a_{k^*-1} & a_{k^*} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & \cdots & a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & 0 & 0 & \cdots & 0 & a_0 & a_1 & a_2 & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & a_0 & a_1 & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & a_0 & \cdots \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{3.142}$$

and

$$B_{vv} = \begin{bmatrix} a_0 & a_1 & a_2 & \cdots & a_{k^*-1} & a_{k^*} & a_{k^*+1} & a_{k^*+2} & \cdots \\ 0 & a_0 & a_1 & \cdots & a_{k^*-2} & a_{k^*-1} & a_{k^*} & a_{k^*+1} & \cdots \\ 0 & 0 & a_0 & \cdots & a_{k^*-3} & a_{k^*-2} & a_{k^*-1} & a_{k^*} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & \cdots & a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & 0 & 0 & \cdots & 0 & c_0 & c_1 & c_2 & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & c_0 & c_1 & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & c_0 & \cdots \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{3.143}$$

where for $j = 0, 1, 2, \ldots,$

$$a_j = \int_0^\infty \frac{(\lambda t)^j e^{-\lambda t}}{j!}(1 - V(t))dt \tag{3.144}$$

and

$$c_j = \int_0^\infty \frac{(\lambda t)^j e^{-\lambda t}}{j!}(1 - B(t))dt \tag{3.145}$$

Given the notation above, it is straight a forward matter to examine the queue length distribution as seen at arbitrary times. Let κ be the vector of queue length probabilities when the system is observed at arbitrary times. Here,

$$\kappa = [\kappa(0)\ \kappa(1)\ \kappa(2)\ ..] \tag{3.146}$$

with

$$\kappa(j) = \lim_{t\to\infty} P\{n_t = j\}, \quad j = 0,1,2,\ldots \tag{3.147}$$

It follows directly from Corollary 2.13 that $\kappa(j)$ is given by

$$\kappa(j) = \eta_s(j) + \eta_v(j), \quad j = 0,1,2,\ldots \tag{3.148}$$

Here, (3.148) taken together with (3.139), (3.142), and (3.143) shows that $\kappa(j)$ can be written as

$$\kappa(j) = \frac{1}{\pi m}\left(\delta(j) + \sum_{k=0}^{j}\pi_s(k)a_{j-k}\right), \quad \text{for } j = 0, 1, 2, \ldots \tag{3.149}$$

where,

$$\delta(j) = \begin{cases} \sum_{k=0}^{j}\pi_v(k)a_{j-k}, & j < k^* \\ \sum_{k=0}^{k^*-1}\pi_v(k)a_{j-k} + \sum_{k=k^*}^{j}\pi_v(k)c_{j-k}, & j \geq k^* \end{cases}.$$

Defining $K(z)$ as the pgf of the queue length as seen at arbitrary times, we have that

$$K(z) = \sum_{j=0}^{\infty} \kappa(j) z^j; \tag{3.150}$$

hence, it follows from (3.149) and (3.150) that

$$K(z) = \frac{1}{\pi m} \sum_{j=0}^{\infty} z^j \left(\delta(j) + \sum_{k=0}^{j} \pi_s(k) a_{j-k} \right) \tag{3.151}$$

Let $\Pi_s(z)$ and $\Pi_v(z)$ be given by (3.120) and (3.121) respectively. Distributing the outer summation on the right hand side of (3.151) and the interchanging the order of summations in the usual manner, (3.151) can be rewritten as

$$K(z) = \frac{1}{\pi m} \left(\tilde{A}(z)\Pi_s(z) + \tilde{C}(z)\Pi_v(z) + \left(\tilde{A}(z) - \tilde{C}(z)\right) \sum_{j=0}^{k-1} \pi_v(j) z^j \right) \tag{3.152}$$

where,

$$\tilde{A}(z) = \sum_{j=0}^{\infty} a_j z^j = \int_0^{\infty} e^{-(\lambda - \lambda z)t} (1 - V(t)) dt \tag{3.153}$$

and

$$\tilde{C}(z) = \sum_{j=0}^{\infty} c_j z^j = \int_0^{\infty} e^{-(\lambda - \lambda z)t} (1 - B(t)) dt \,. \tag{3.154}$$

Substituting (3.126) and (3.127) into (3.152) and rearranging, it follows that

$$K(z) = \left(\frac{1}{\pi m} \sum_{j=0}^{k^*-1} \pi_v(j) z^j \right) \cdot \frac{z^{k^*}(\tilde{A}(z) + \tilde{C}(z)(\tilde{V}(z) - 1)) - \tilde{B}(z)\tilde{A}(z)}{z^{k^*} - \tilde{B}(z)\tilde{V}(z)}. \qquad (3.155)$$

With (3.155) we have the pgf for the queue length as seen at arbitrary times and for the M/GI/1 vacation system with limited batch service, it remains only to evaluate the constants appearing in (3.155). A straight forward approach to evaluating these constants is to first examine πm. It follows from Theorem 2.12 that for the (X,T) process,

$$\pi m = \bar{V} \sum_{j=0}^{\infty} \pi_s(j) + \bar{V} \sum_{j=0}^{k^*-1} \pi_v(j) + \bar{B} \sum_{j=k^*}^{\infty} \pi_v(j) \qquad (3.156)$$

where $\bar{B}$ is the expected length of the service time for a batch of customers and $\bar{V}$ is the expected length of the server's vacations. Observe that

$$\sum_{j=0}^{\infty} \pi_s(j) = \lim_{m \to \infty} P\{H_m = s\} \qquad (3.157)$$

and

$$\sum_{j=0}^{\infty} \pi_v(j) = \lim_{m \to \infty} P\{H_m = v\} \qquad (3.158)$$

where, (3.157) gives the stationary probability that the queue length is observed immediately following a batch service completion and (3.158) gives the stationary probability that the queue length is observed immediately following the server's return form vacation. The probabilities of (3.157) and (3.15) will be treated by examining the pgf for the queue length as seen immediately following either batch service completions or the

server's returns from vacation.

Let $\pi(j)$ be the stationary probability that the queue length is j when observed immediately following either batch service completions or the server's returns from vacation. It follows that

$$\pi(j) = \pi_s(j) + \pi_v(j). \tag{3.159}$$

Define $\Pi(z)$ as the pgf of the queue length embedded at both batch service completions and vacation completions where

$$\Pi(z) = \sum_{j=0}^{\infty} \pi(j) z^j. \tag{3.160}$$

Substituting (3.159) into (3.160) yields

$$\Pi(z) = \sum_{j=0}^{\infty} z^j (\pi_s(j) + \pi_v(j)) \tag{3.161}$$

and it is easily recognized that

$$\Pi(z) = \Pi_s(z) + \Pi_v(z). \tag{3.162}$$

Here, $\Pi_s(z)$ and $\Pi_v(z)$ are the geometric transforms given by (3.126) and (3.127) respectively. Since $\Pi(z)$ is a pgf, it follows from (3.162) that

$$1 = \lim_{z \uparrow 1} (\Pi_s(z) + \Pi_v(z)) \tag{3.163}$$

For convenience, we adopt the notation that

$$\Pi_s(1) = \lim_{z \uparrow 1} \Pi_s(z)$$

and

$$\Pi_v(1) = \lim_{z \uparrow 1} \Pi_v(z).$$

It now follows that (3.156) can be rewritten as

$$\pi m = \bar{V}\Pi_s(1) + \bar{B}\Pi_v(1) + (\bar{V} - \bar{B})\sum_{j=0}^{k^*-1}\pi_v(j) \tag{3.164}$$

whenever the constants $\pi_v(0), \pi_v(1), \ldots, \pi_v(k^* - 1)$ are chosen such that

$$\lim_{z \uparrow 1} \Pi(z) = 1.$$

Here, substituting (3.126) and (3.127) into (3.162) we have that

$$\Pi(z) = \sum_{j=0}^{k^*-1}\pi_v(j)z^j \cdot \frac{z^{k^*}\tilde{V}(z) - \tilde{B}(z)}{z^{k^*} - \tilde{B}(z)\tilde{V}(z)} \quad . \tag{3.165}$$

By applying Rouche's theorem to (3.165) in the usual manner, it is a routine, but lengthy, matter to evaluate the the constants $\pi_v(0), \pi_v(1), \ldots, \pi_v(k^* - 1)$ such that $\lim_{z \uparrow 1} \Pi(z) = 1$. As a matter of convenience, evaluation of these constants is omitted.

In order to rewrite $K(z)$, the pgf of the queue length as seen at arbitrary times, in a convenient form, we introduce the following Laplace-Stieltjes transforms.

$$B^*(\sigma) = \int_0^\infty e^{-\sigma t} B(dt), \tag{3.166}$$

$$V^*(\sigma) = \int_0^\infty e^{-\sigma t} V(dt), \tag{3.167}$$

$$C^*(\sigma) = \int_0^\infty e^{-\sigma t}(1 - B(t))dt, \tag{3.168}$$

and

$$A^*(\sigma) = \int_0^\infty e^{-\sigma t}(1 - V(t))dt. \tag{3.169}$$

From elementary properties of Laplace-Stieltjes transforms, it is can be shown that

$$C^*(\sigma) = \frac{1}{\sigma} - \frac{B^*(\sigma)}{\sigma} \tag{3.170}$$

and

$$A^*(\sigma) = \frac{1}{\sigma} - \frac{V^*(\sigma)}{\sigma} \tag{3.171}$$

Recalling that

$$\tilde{B}(z) = B^*(\lambda - \lambda z),$$

$$\tilde{V}(z) = V^*(\lambda - \lambda z),$$

$$\tilde{C}(z) = C^*(\lambda - \lambda z),$$

and

$$\tilde{A}(z) = A^*(\lambda - \lambda z),$$

it follows that substituting (3.164) and (3.168) through (3.171) into (3.155) gives

$$K(z) = \frac{1}{(\lambda - \lambda z)\left((\bar{V} - \bar{S})\left(\Pi_S(1) + \sum_{j=0}^{k^*-1}\pi_V(j)\right) + \bar{S}\right)} \cdot \sum_{j=0}^{k^*-1}\pi_V(j)z^j$$

$$\cdot \frac{(V^*(\lambda - \lambda z) - 1)B^*(\lambda - \lambda z)(z^{k^*} + 1)}{z^{k^*} - B^*(\lambda - \lambda z)V^*(\lambda - \lambda z)} \qquad (3.172)$$

where, the constants $\pi_V(0), \pi_V(1), \ldots, \pi_V(k^* - 1)$ are chosen such that $\lim_{z \uparrow 1} \Pi(z) = 1$.

Since M/GI/1 vacation systems with limited batch service fail to satisfy the conditions of Proposition 3.1, the distributional form of Little's law does not hold here and no Laplace-Stieltjes transform for the customer waiting time distribution is available. However, given the queue length pgf of (3.172), it is a simple matter to calculate the expected customer waiting time by applying the customary form of Little's law. Let T be the waiting time for an arbitrary customer. Little's law requires that

$$E[T] = \frac{1}{\lambda} \lim_{z \uparrow 1} \frac{dK(z)}{dz}. \qquad (3.173)$$

The calculation indicated by (3.173) is tedious; since $K(z)$ is of indeterminate form, finding E[T] requires multiple application of L'Hopital's rule. The application of Little's law in queueing systems is well studied and is presented in most elementary queueing theory texts. Refinements of (3.173) are considered outside the scope of this work.

Excepting our examination of simple limited service systems, the results presented in this section are new. It is here worth restating that the ergodic queueing behavior of systems operating with limited batch service is not completely characterized by the Markov chain X. Investigation of the queue length distribution as seen at arbitrary times requires that the semi-regenerative nature of the joint queue length / server activity process X_{R^*} be exploited. For this reason, limited batch service represents the most sophisticated server scheduling discipline considered in this chapter. This completes our investigation of M/GI/1 vacation systems with limited batch service.

4. Conclusions and Recommendations for Future Research

Vacation systems represent an important class of queueing models having application in both computer communication systems and integrated manufacturing systems. By specifying an appropriate server scheduling discipline, vacation systems are easily particularized to model many practical situations where a server's effort is divided between primary and secondary customers.

The queueing literature reviewed in Chapter 1 offers performance analyses for M/GI/1 vacation systems operating under a variety of server scheduling disciplines. These analyses are not derived as particularizations of some general model for the M/GI/1/L vacation system. Rather, each author exploits certain "special tricks" that are uniquely applicable to the particular server scheduling discipline under investigation to yield desired results. The absence of a general model suggests that performance analysis of vacation systems must be considered on a case by case basis.

The development of a general stochastic framework that subsumes a wide variety of server scheduling disciplines (including those introduced in Chapter 1) for M/GI/1/L vacation systems is the focus of this research. In Chapter 2 we have identified a class of server scheduling disciplines that we denote as *Markov schedules*. Characterization of the class of Markov schedules is new. Chapter 2 provides a formal characterization of the stochastic behavior of M/GI/1/L vacation systems having Markov schedules.

A "bottom-up" approach has been taken in developing a stochastic process that describes the queueing characteristics of M/GI/1/L vacation systems with Markov schedules. This process, called the *joint queue length / server activity process*, is shown

to have embedded within it the *queue length / server activity marked point process*. The queue length / server activity marked point process is constructed from more fundamental stochastic processes on which probability structures of practical significance are easily defined.

Beginning with formal definitions for the *server switching point process* and the *queue length process*, Section 2.2 offers a detailed development of the queue length / server activity marked point process and identifies it as a stochastic process embedded within the joint queue length / server activity process at all service period completions and vacation period completions.

Section 2.3 presents the development of the probability structure on the queue length / server activity marked point process. Here, three conditions that define Markov schedules are presented. It is then shown that when the server scheduling discipline for an M/GI/1/L vacation system satisfies these conditions, the queue length / server activity marked point process is Markov renewal. Further, it is shown that the queue length /server activity marked point process also forms a Markov renewal process when embedded only at service period completion epochs or embedded only at vacation period completion epochs. This fact is exploited often when analyzing the example vacation systems of Chapter 3.

The probability structure on the joint queue length / server activity process is developed in Section 2.4. It is shown that when the server scheduling discipline for an M/GI/1/L vacation system belongs to the class of Markov schedules, service period completion times and vacation period completion times are stopping times for the joint queue length / server activity process, and consequently the process is semi-regenerative. Theorem 2.11 is the principal result of Chapter 2, offering a characterization of the queueing behavior for

M/GI/1/L vacation systems over all time. Theorem 2.12 and Corollary 2.13 provide formulae that are convenient computational tools for examining ergodic queueing behavior.

The probability structure associated with M/GI/1/L vacation systems having Markov schedules presented in Chapter 2 is new. Investigation of the joint queue length / server activity process forms the cornerstone of this research. The semi-regenerative nature of this process allows characterization of the ergodic queue length as seen at arbitrary times with computational formulae that are relatively simple. The joint queue length / server activity process can be particularized to capture most server scheduling disciplines investigated by other authors, and is sufficiently general to characterize more sophisticated server scheduling (e.g., batch service systems) that do not appear in the literature.

It is worth noting that Theorem 2.12 and Corollary 2.13 provide useful computational formulae that accommodate systems having either finite or infinite queue capacities. In the case of finite queue capacities, these results are easily applied. There is little literature available regarding vacation systems with finite queue capacities, and the results of Chapter 2 offer a powerful theory for analyzing such systems.

The general probability structure underlying M/GI/1/L vacation systems with Markov schedules is particularized, in Chapter 3, to examine the queueing behavior of three example vacation system. These example systems are presented so as to demonstrate both the validity and usefulness of the general probability structure developed in Chapter 2. The ergodic queueing behavior of these systems is examined by developing certain queue length probability generating functions (pgf's) that are of practical importance.

In Section 3.1, the ergodic queueing behavior of the M/GI/1 vacation system having Bernoulli schedules is examined. Expressions for the queue length pgf's as seen

immediately following service completions, immediately following vacation completions, and at arbitrary times are developed. The expressions for these three pgf's do not appear in the literature and are presumed new. However, exhaustive service is a special case of Bernoulli schedules; it is shown that when particularized to reflect exhaustive service, the expression for the pgf of queue length as seen at arbitrary times agrees with results found in the literature.

In Section 3.2, the ergodic queueing behavior of the M/GI/1 vacation system with E-limited service is examined. Expressions for the queue length pgf's as seen immediately following service completions, immediately following vacation completions, and at arbitrary times are developed. E-limited service is such that the pgf for queue length as seen immediately following service completions is the same as the pgf for queue length as seen at arbitrary times. Expressions for these pgf's agree with results presented by Takagi (1987). The expression for the pgf of queue length as seen immediately following the server's returns from vacation is new.

Bernoulli schedules and E-limited service belong to the class of server scheduling disciplines in which customers are served in a one at a time fashion. For queueing systems having such server scheduling disciplines, it is well known that the ergodic queue length as seen at arbitrary times can be studied by examining the Markov chain embedded within the joint queue length / server activity process immediately following service completions. Thus, the example systems considered of Sections 3.1 and 3.2 are analyzed via the Markov chain X embedded within the queue length / server activity marked point process (X,T). It is unnecessary to analyze systems having limited to one type service via the semi-regenerative joint queue length / server activity process X_{R^+}.

Unlike vacation systems having limited to one type service, the ergodic queue length as seen at arbitrary times is not the same as the ergodic queue length as seen immediately following service completions for batch service systems. The ergodic queue length as seen at arbitrary times for batch service systems, having server scheduling disciplines belonging to the class of Markov schedules, must be examined via the stationary distribution of the joint queue length / server activity process X_{R+}. To the author's knowledge, vacation systems having server scheduling disciplines other than the limited to one type do not appear in the literature.

In Section 3.3, we examine the ergodic queueing behavior of the M/GI/1 vacation system with *limited batch service*. Expressions for the queue length pgf's as seen immediately following service completions, immediately following vacation completions, and at arbitrary times are developed. The limited batch service server scheduling discipline does not appear in the literature, and to the author's knowledge its introduction here represents the first analysis of a vacation system having a server scheduling that is not of the limited to one type. Thus, the expressions for the three above mentioned pgf's are presumed new.

The M/GI/1 vacation system with limited batch service subsumes, as a special case, the simple *limited service* server scheduling discipline when batches are of size one. In this special case, the pgf of ergodic queue length as seen at arbitrary times agrees with results appearing in the literature Takagi (1987).

For each of the example vacation systems considered in Chapter 3, the procedure for developing the ergodic queue length pgf's is the same. Unlike the analyses appearing in the literature, no "special tricks" particular to any of the systems is required for these developments. The procedure for developing the queue length pgf's for these vacation

system begins with identification of the appropriate dimensions for the state space E of the joint queue length / server activity process X_R.. It is then straight forward to form a pair of simultaneous equations that when solved yield geometric transforms that, under the appropriate boundary conditions, are the queue length pgf's as seen immediately following service completions and immediately following vacation completions.

In the case of batch service we have that the queue length pgf as seen at arbitrary times is not the same as the queue length pgf as seen immediately following service completions, and the former pgf is obtained following a simple linear transformation on the stationary distribution of (X,T) the queue length / server activity marked point process.

It is important to note that the queue length pgf's developed in Chapter 3 are attainable since the vacation systems under consideration are such that: 1) the queue capacity is infinite, 2) customer service times are independent and identically distributed, 3) server vacation times are independent and identically distributed, and 4) the length of service periods and the length of vacation periods are mutually independent. When any of these four characteristics is relaxed for M/GI/1 vacation systems with Markov schedules, pgf's are difficult, if not impossible, to obtain. However, Theorem 2.12 and Corollary 2.13 still apply and offer a somewhat less convenient characterization of the ergodic queueing behavior.

The probability structure underlying M/GI/1/L vacation systems with Markov schedules, developed in Chapter 2, suggests a number of possibly interesting extensions to the present research. In the discussion to follow, no formal exposition of such extensions is presented; rather, possible future research topics, presented in no particular order of importance, are informally discussed.

Given the generality exhibited by the class of Markov schedules, it is a simple matter to identify for M/GI/1/L vacation systems many practical server scheduling disciplines belonging to the class of Markov schedules that do not, as yet, appear in the literature. In particular, batch service vacation systems, examined in this work only under the limited batch service server scheduling, are of significant practical importance. However, batch-type server scheduling disciplines require much further investigation. While such investigations are simply applications of the theory of M/GI/1/L with Markov schedules, it seems reasonable that there exist many important results to be discovered for particular vacation systems.

For our work thus far, the focus has been limited to simple vacation systems; that is, vacation systems where customers, upon completing service, depart the system never to return. There exist situations, however, where customers, upon completing service, may rejoin the queue and await further service. Such systems are here referred to as vacation systems with instantaneous feedback. In situations where the feedback mechanism is a Bernoulli switch, Disney and Keissler (1987), the probability structure of Chapter 2 appears to directly apply.

In situations where the feedback mechanism is more sophisticated than a Bernoulli switch, it may be possible to classify certain server scheduling disciplines as belonging to the class of Markov schedules; however, this may require extending E the state space of the queue length / server activity marked point process (X,T) beyond the present definition. In particular, more sophisticated feedback mechanisms may require extending the dimension of the random vector h_t to include additional random variables in order to satisfy Conditions 1, 2, and 3 and thus belong to the class of Markov schedules. The study of feedback vacation systems having Markov schedules is an open issue.

The notion of extending the definition of the queue length / server activity marked point process (X,T) and its state space E so that Markov schedules are defined for server scheduling disciplines of systems other than the simple vacation systems examined in this work suggests an approach to investigating single server, multiple queue systems. Examples of single server, multiple queue systems of practical importance are found in vacation systems having priority services and polling systems.

Single server, multiple queue systems, under their most general description, operate as follows. A fixed number of queues are attended by a single server. Customers arrive to each queue according to a stochastic process (that can be different for each queue). Customer service times are drawn from general distributions. Under specified conditions the server, upon completion of a customer service, will abandon further customer service to begin a walk of random length leading to some queue in the system. Upon completing a walk, the server will, under specified conditions, either begin a customer service or begin another walk.

For convenience, we here restrict our attention to single server, multiple queue system (consisting of M queues) having mutually independent arrival streams, and having server activity such that the lengths of all walk times and service times are mutually independent.

It is clear that the server's activity is divided exclusively between walking between queues and serving customers. Given this observation, it is a simple matter to extend our present definition of the mark space for the server switching marked point process to reflect the server's behavior in multiple queue systems. That is, server switching epochs are marked by either s_i or v_i , i = 1, 2, ..., M. Here, s_i indicates a service completion at queue

i while v_i indicates a walk completion ending at queue i.

It is also a simple matter to extend our present definition of the queue length process n_t to account for multiple queues. Here, let n_t be the M-vector of queue lengths at time t. This extension of the queue length process together with the extension to the server switching marked point process infer extensions to both the joint queue length / server activity process X_{R^+} and the queue length / server activity marked point process (X,T).

With the above described extensions, Conditions 1, 2, and 3 define a class of Markov schedules for multiple queue systems. It appears (though it has not been shown) that the results presented Chapter 2 are unaltered under the extension to single sever, multiple queue systems. That is, vacation systems having Markov schedules appear to be a special case of single server, multiple queue systems having Markov schedules.

Accepting that all results of Chapter 2 extend to single server, multiple queue systems, much research remains in order to quantify the queueing behavior of such systems. In the multiple queue environment, probability generating functions are necessarily multidimensional. Intuition suggests that construction of multidimensional transforms will be unfeasible for all but the most simple server scheduling disciplines. Hence, the value of transform results in the multiple queue environment may be limited. For this reason it would seem that future research efforts for single server, multiple queue systems should be directed towards exploiting the theory of Chapter 2 in developing qualitative system performance measures. The underlying probability structure associated with systems having Markov schedules appears promising for answering such qualitative questions as ordering server scheduling disciplines according to increasing system throughput, or comparing priority schedules to determine minimum server idle time.

Given the complexity of the state space E for the joint queue length / server activity process X_{R^+} for single server, multiple queue systems, it is perhaps unreasonable to seek performance measures that generate numbers. Characterization of the probability structure of stochastic processes associated with queueing systems having Markov schedules offers the promise of studying the performance of a wide variety of practical, non-elementary single server, multiple queue systems via qualitative measures. Much research remains in identifying how such qualitative measures can be developed from the underlying stochastic processes that govern the queueing behavior of these systems.

Finally, it should be noted that the characterization of customer waiting times is an open research issue. Sections 3.1 and 3.2 offer example systems where customers are served in a one at a time fashion. For such systems, it was shown that the distributional form of Little's law holds, and customer waiting times can be investigated via the pgf of queue length as seen at arbitrary times. Many systems of practical interest fail to satisfy the conditions of Proposition 3.1, and for these systems the distributional form of Little's law is of little value.

Much research is needed to characterize the customer waiting time distributions of systems for which the distributional form of Little's law does not apply. When the queue length / server activity process is Markov renewal, customer waiting times appear to be readily formulated as first passage times of the Markov renewal process. Thus, the underlying Markov renewal structure of vacation systems having Markov schedules may provide the framework necessary to study ergodic customer waiting times. Such analyses may extend to single server, multiple queue systems having Markov schedules, owing that these systems are characterized by an underlying Markov renewal process.

5. References

Cinlar, E. 1975. *Introduction to stochastic processes.* Englewood Cliffs, N.J.: Prentice-Hall.

Disney, R. L., and Kiessler, P. C. 1987. *Traffic processes in queueing networks: a Markov renewal approach.* Baltimore, MD: The Johns Hopkins University Press.

Doshi, B. T. 1986. Queueing systems with vacations - a survey. *Queueing Systems.* 1:29-66.

Fuhrmann, S. W. , and Cooper, R. B. 1985. Stochastic decompostions in the M/G/1 queue with generalized vacations. *Oper. Res.* 33:1117-29.

Fujiki, M., and Gambe, E. 1980. *Communication traffic theory* (in Japanese). Tokyo: Maruzen Co. Ltd.

Gelenbe, E., and Iasnogorodski, R. 1980. A queue with server of the walking type (autonomous service). *Ann. Inst. Poincare.* 16:63-73.

Kielson, J.,and Servi, L. 1986. Oscillating random walk models for GI/G/1 vacation systems with Bernoulli schedules. *J. Appl. Probab.* 23:709-802.

_____. 1988. On the distributional form of Little's law. Submitted to*Oper. Res. Ltrs.*

Kleinrock, L. 1975. *Queueing systems* Vol.1, *Theory.* New York: Wiley Interscience.

_____. 1976. *Queueing systems* Vol. 2, *Computer applications.* New York: Wiley Interscience.

Kohlas, J. 1982. *Stochastic methods of operations research.* New York: Cambridge university Press.

Lee, T. T., 1983. M/G/1/N queue with vacation and limited service discipline. Technical Momorandum, Bell Telephone Laboratories.

Leibowitz, M. A. 1961. An approximate method for treating a class of multiqueue problems. *IBM Journal of Research and Development.* 5:204-9.

Lucantoni, D., Meier-Hellestern, K., and Neuts, M. 1988. A single server queue with server vacations and a class of non-renewal arrival processes. U. of Arizona, Dept of Systems Engineering Working Paper.

Ramaswamy R.,and Servi, L. 1986. Busy period of M/G/1 vacation queues with Bernoulli schedules. Submitted to *IEEE Trans. Commun.*

Servi, L. 1986. D/G/1 queues with vacations. *Oper. Res.* 31:705-19.

Takagi, H. 1987. Queueing analysis of vacation models. IBM TRL Research Report TR87-0032.

Wolff, R. W. 1982. Poisson arrivals see time averages. Oper. Res. 30:223-31.

The vita has been removed from the scanned document