THESIS REPORT
Ph.D.

SYSTEMS
RESEARCH
CENTER

*Supported by the
National Science Foundation
Engineering Research Center
Program (NSFD CD 8803012),
Industry and the University*

Nonexhaustive Policies in Polling Systems and Vacation Models: Qualitative and Approximate Approach

by Tedijanto
Advisor: A.M. Makowski

Ph.D. 90-3

ABSTRACT

Title of dissertation: NONEXHAUSTIVE POLICIES IN POLLING SYSTEMS AND VACATION MODELS: QUALITATIVE AND APPROXIMATE APPROACHES

Tedijanto, Doctor of Philosophy, 1990

Dissertation directed by: Armand M. Makowski, Associate Professor, Electrical Engineering Department

We consider a polling system which consists of a number of queues attended by a single server. The server switches from one queue to another following a fixed cyclic order. Such a system finds a wide variety of applications in the computer, communication and manufacturing fields. Polling systems under the exhaustive service policy, where the server serves each queue until it becomes empty, have been extensively studied in the literature. This thesis studies nonexhaustive service policies, in particular the so-called limited and Bernoulli policies. Unlike the exhaustive policy, nonexhaustive policies usually do not lend themselves to exact analysis. We show that approaches based on heavy and light traffic analysis, and stochastic comparison techniques, can provide useful information about the performance of these policies.

In the first part of the thesis, we consider polling systems with a single queue, more commonly known as vacation models. In a fairly general setting, we prove heavy traffic limit theorems for vacation models under the Bernoulli and limited policies. We then establish light traffic results for vacation models with Poisson arrivals which are subsequently combined with the heavy traffic results to form the bases for interpolation approximations. Using stochastic comparison techniques, we identify some general conditions under which two service policies can be stochastically compared. In this framework, we establish bounds, monotonicity and comparison results for various service policies. The comparison between the limited and

Bernoulli policies represents a relatively harder problem and cannot be established in the general framework. However, we show that under more restrictive conditions, a weaker comparison in the increasing convex ordering indeed holds.

In the second part of the thesis, we study $M/GI/1$ polling systems. Taking advantage of a recently established decomposition result, we first derive a pseudo-conservation law for the Bernoulli policy which, in the homogeneous case, leads to closed-form formulae for some performance measures. As a by–product, we obtain a comparison result between the Bernoulli and limited policies in homogeneous polling systems. For the limited policy, we propose and study an approximation algorithm which is based on the interpolation approximation developed for the vacation models.

NONEXHAUSTIVE POLICIES IN POLLING SYSTEMS AND VACATION MODELS: QUALITATIVE AND APPROXIMATE APPROACHES

by

Tedijanto

Dissertation submitted to the Faculty of the Graduate School
of The University of Maryland in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
1990

Advisory Committee:

Associate Professor Armand M. Makowski, Chairman/Advisor
Assistant Professor Scott D. Carson
Associate Professor Clyde P. Kruskal
Associate Professor Prakash Narayan
Associate Professor A. Yavuz Oruç
Associate Professor Charles B. Silio

DEDICATION

To My Dear Parents

ACKNOWLEDGMENT

I wish to express my heartfelt gratitude to my advisor, Professor Armand M. Makowski, for his guidance and support throughout the course of this work, and for carefully reading the many iterations of this thesis. Without his ceaseless encouragement, the completion of this dissertation would not have been possible. I would also like to thank my fellow graduate student Subir Varma for introducing me to heavy and light traffic analysis, and for the numerous discussions which have contributed positively to this thesis. Special thanks go to Professors Scott D. Carson, Clyde P. Kruskal, Prakash Narayan, A. Yavuz Oruç and Charles B. Silio for serving in my advisory committee. I am also grateful to Systems Research Center for providing a stimulating environment in which I conducted this research. Finally, I am mostly indebted to Ming for her love, patience and encouragement throughout the years.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

BACKGROUND AND SUMMARY

1.1 Introduction

During the last decade, network architecture based on ring topology and token passing protocol has received widespread acceptance among designers of Local Area Network (LAN) [16]. Two notable examples of LANs employing such architecture are the so–called Token Bus and Token Ring LANs. A Token Bus LAN consists of a number of workstations connected to a single bidirectional communication bus. At any given time, only one station can use the bus to transmit messages to other stations. The stations are *logically* arranged to form a ring, i.e., each station stores the unique address of the station which is to be positioned immediately following it in the logical order. The last station in the order stores the address of the first one. A station, after finishing its message transmission (or otherwise having to relinquish its control of the bus under some prespecified conditions), releases a string of control characters called a *token* which is intended for the next station in the logical order. When this station detects the token, it seizes control of the bus and starts transmitting messages if some are ready for transmission. The operation of a Token Ring LAN is very similar to that of the Token Bus LAN; the main difference between these two LANs is that the stations in the Token Ring LAN are *physically* arranged to form a ring. Each station is connected to the next station in the ring by a unidirectional bus, and data is circulated around the ring on a series of station–to–station hops.

The Token Bus and Token Ring LANs constitute two examples of systems in which a single resource is shared by a number of users by circulating it among them according to a fixed cyclic order. Such systems, known as *polling systems*, arise quite naturally in the computer, communication and manufacturing areas. In queueing–theoretic formulation, polling systems are modeled by the so–called *cyclic–service queueing systems*. A cyclic–service queueing system consists of a set of N queues $Q_1, Q_2, \ldots, Q_N$ served by a single server. The server serves one queue at a time and travels among the queues following a fixed cyclic order: $Q_1, Q_2, \ldots, Q_N, Q_1, Q_2, \ldots$. A (possibly nonzero) delay called the *switchover time* is experienced by the server in going from one queue to the next. Here a queue represents a user, a customer in the queue represents a service demand from the corresponding user and a switchover time models the processing time involved in transferring the control of the resource from one user to the next. Since here we only consider queueing–theoretic approach to polling systems, in the sequel we shall use the terms 'polling system' and 'cyclic–service queueing system' interchangeably.

Polling systems have been studied since the 1950s; many variants have been introduced to model a wide variety of systems. Variants are usually distinguished from one another by the *service policy* used. Loosely speaking, the service policy (some authors prefer to call it the *switching policy*) determines how long the server stays in each queue. For instance, under the *exhaustive* service policy, the server serves each queue until the queue becomes empty. Under a *nonexhaustive* policy, the server may switch to the next queue even though some customers are waiting in the current queue. Numerous nonexhaustive policies have been proposed and studied in the literature. However, unlike the exhaustive policy, the performance of most nonexhaustive policies are not yet very well understood. In fact, for some of these policies, in particular the so–called *limited* policies, exact analysis does not appear to be feasible [82,83].

The purpose of this thesis is to study some nonexact approaches to the problem of evaluating nonexhaustive service policies. Specifically, we combine light and heavy traffic analysis with stochastic comparison techniques to obtain asymptotic and structural results which are then used as bases for approximation schemes. Most of the thesis is devoted to the limited and the so–called *Bernoulli* service policies, but we expect that similar approaches can also be applied to other service policies for which exact analysis is not feasible.

The motivation for studying nonexhaustive service policies resides in the fact that many real systems which are modeled by polling systems do not operate under an exhaustive service policy. This is due to a drawback of the exhaustive policy in that it allows a user (queue) with high traffic load to practically monopolize the resource. As the name would indicate, the limited service policy puts a fixed limit on each queue on the number of customers the server can serve consecutively within a visit to that queue. The Bernoulli policy also imposes a limit; the difference is the implementation of this limit as a random variable. We shall present a more precise description of these two service policies in the next section. As we shall see, each of these policies is parametrized by an N–dimensional vector, thus lending themselves to system optimization, i.e., the policy parameter can be appropriately chosen to maximize a given performance measure. The limited policy where the limit is set to one for all the queues is currently used in most Token Bus and Token Ring LANs. More recently, limited policies have also been used to study FDDI token ring protocol [74].

In the next section, we present a precise description of the systems considered in this thesis. In Section 3, we briefly survey the literature on polling systems. This survey is not intended as a complete account of the work done on the subject, but rather is meant to give the reader a general feel of the methods used by researchers in this field. We shall also point out the difficulties in analyzing nonexhaustive service policies. In Section 4, we summarize the results of this thesis in a chronological fashion.

1.2 Polling Systems and Vacation Models

In this section, we describe more precisely a polling system. We also introduce a *vacation model* which can be seen as a polling system with $N = 1$. We do not define much notation in this section; we shall do that at appropriate places in the thesis as we go along.

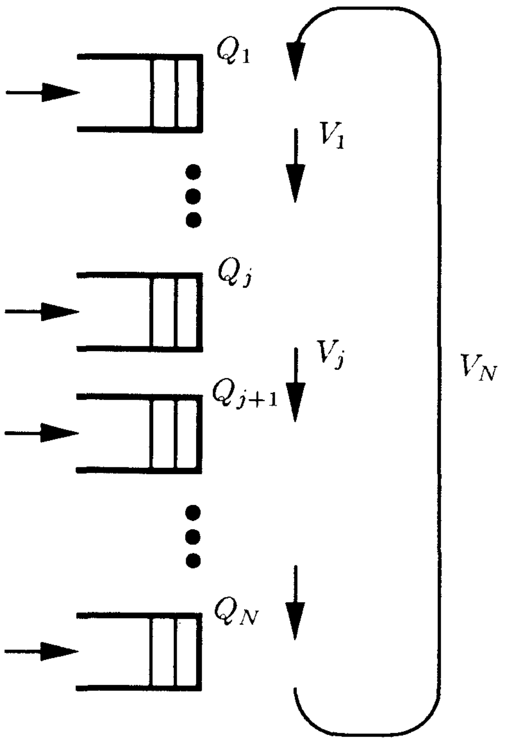

Figure 1.1: A Polling System

A polling system, depicted in Figure 1.1, consists of N infinite–capacity queues, denoted by $Q_1, Q_2, \ldots, Q_N$, and a single server which serves the queues in a fixed order: $Q_1, Q_2, \ldots, Q_N$, $Q_1, Q_2, \ldots$. For each $j = 1, \ldots, N$, we shall refer to a customer which arrives to Q_j as a type–j customer. The arrivals to Q_j are governed by the renewal process $\{A_j^{n+1},\ n = 0, 1, \ldots\}$ with the interpretation that A_j^{n+1} is the time between the arrival of the nth and the $(n+1)$st type–j customers. We take the convention that the customer with index 0 arrives at time 0. The nth customer of type–j brings with it a job which requires a random amount B_j^n of processing time. The server takes a random amount of time to switch from Q_j to $Q_{j+1\,(\text{mod}\,N)}$; the length of the nth of such time is given by a random variable V_j^n. We assume that the sequences $\{A_j^{n+1},\ n = 0, 1, \ldots\}$, $\{B_j^n,\ n = 0, 1, \ldots\}$ and $\{V_j^n,\ n = 0, 1, \ldots\}$,

$j = 1, \ldots, N$, are all mutually independent and each is constituted of i.i.d. random variables.

A limited service policy is parametrized by a vector of integer components $(m_1, \ldots, m_N)$ with $m_j \geq 1$, $j = 1, \ldots, N$. The integer m_j is the maximum number of customers that can be served in one visit of the server to Q_j. To be more precise, at every service completion at Q_j, the server switches to Q_{j+1} if and only if the queue is empty or the number of customers that have been served during the current visit is m_j.

A Bernoulli service policy is parametrized by a vector of probabilities $(p_1, \ldots, p_N)$ with $0 \leq p_j \leq 1$, $j = 1, \ldots, N$. At every service completion at Q_j, if the queue is not empty, the server serves the next available customer with probability p_j and switches to Q_{j+1} with probability $1 - p_j$. If the queue is empty at the service completion, the server switches to Q_{j+1}.

A polling system with $N = 1$ is more commonly known as a *vacation model*, in which case the switchover times are usually called *vacation times*. During these times, the server "switches" to the same queue and becomes temporarily unavailable to the customer—hence the name. Although vacation models have many applications of their own in the computer communications area [22,23], these models are usually considered in conjunction with polling systems. As we shall see in this thesis, vacation models can usually be studied more thoroughly because of their relative simplicity. In many instances, results obtained for vacation models can be extended to polling systems or otherwise used as building blocks for approximation schemes. In a polling system with an arbitrary number of queues, we can also find vacation models embedded in the system. This can be seen by observing a queue in isolation and looking at the time the server is away from that queue as a vacation time for that queue.

In the context of vacation models, the limited policy with parameter m limits to m the number of customers that can be served between any two consecutive vacations. Under the Bernoulli policy with parameter p, if the queue is nonempty at the end of each service, the server serves the next customer with probability p or goes on vacation with probability $1-p$.

The most important performance measure of interest in polling systems and vacation models is the waiting time of customers in individual queues, i.e., the time elapsed from the moment the customer enters the system to the moment it starts receiving service. Other performance measures such as the server cycle time (the time it takes the server to visit all the queues once) and the queue sizes are sometimes also considered.

1.3 Literature Survey

Polling systems have been studied since the late 1950s when Mack et al. [61,62] used a polling system with single–buffer queues to analyze a set of machines maintained by a patrolling repairman. Then in the 1970s, polling systems were used to study Polling Data Link Control—a system consisting of a number of geographically dispersed terminals connected to a host processor through a single multidrop line (Konheim and Meister [50] and Swartz [78]). In the 1980s, with the advent of low-cost and more powerful microcomputers, the use of LANs to interconnect these computing facilities and other peripheral devices has become more commonplace. The fact that many widely used LANs, especially those employing token passing schemes, can be modeled by polling systems has generated a great deal of interest among researchers in this class of systems. The reader is referred to a monograph by Takagi [81] for a comprehensive survey and a complete list of references on polling systems; see also his survey papers [82,83].

The Exhaustive and Gated Policies

One of the few nonexhaustive policies which have been analyzed thoroughly is the so–called *gated* policy. Under this policy, the server serves only those customers which have been waiting in a queue when the server *polls* that queue; those customers which might arrive during the current visit are served in the next visit. The exhaustive and gated policies are the service policies most extensively studied in the literature. This is due in part to the fact that they found applications to Polling Data Link Control in the 1970s and in part to the fact that they can be analyzed exactly. These two policies in the most general form (i.e., arbitrary N, nonhomogeneous queues and nonzero switchover times) were first solved in the mid 1970s by Eisenberg [25], Hashida [40], Aminetzah [1], and Ferguson and Aminetzah [30]. The treatment of these policies for the case where time is slotted (discrete) can be found in Swartz [78], and Rubin and de Moraes [70]. Earlier, the exhaustive and gated service policies had been analyzed in more restricted settings: Avi–Itzhak et al. [3] considered the case $N = 2$ with zero switchover times; Cooper and Murray [20] studied the case N arbitrary with zero switchover times; the case $N = 2$ with nonzero switchover times was investigated by Sykes [79] and Eisenberg [24]; Konheim and Meister [50] considered the case N arbitrary, nonzero switchover times and homogeneous queues.

In all of the above–mentioned studies, arrivals were assumed to be Poisson. In polling systems with Poisson arrivals, we can usually identify some embedded Markov chains. As it turned out, most performance measures of polling systems under either the exhaustive or gated policy can be solved by analyzing these Markov chains. Consider the Markov chains

$$C_j = \left\{(X_j^1(n), \ldots, X_j^N(n)),\ n = 0, 1, \ldots\right\}, \qquad j = 1, \ldots, N \tag{1.1}$$

where $X_j^i(n)$ is the number of customers in Q_i when the server polls Q_j for the nth time. Assuming stability, we let $F_j(z_1, \ldots, z_N)$, $j = 1, \ldots, N$, be the limiting probability generating function (PGF) of the Markov chain C_j, i.e.,

$$F_j(z_1, \ldots, z_N) = \lim_{n\to\infty} E\left[\prod_{i=1}^{N} z_i^{X_j^i(n)}\right], \qquad |z_i| \leq 1,\ i = 1, \ldots, N. \tag{1.2}$$

The key step in the analysis lies in obtaining (cyclic) recursive equations for $\{F_j,\ j = 1, \ldots, N\}$, i.e., expressing F_{j+1} in terms of F_j, $j = 1, \ldots, N-1$ and F_1 in terms of F_N. Taking the first order (partial) derivatives of both sides of these relationships, we then obtain a set of N^2 linear equations involving

$$f_j(i) = \left.\frac{\partial F_j(z_1, \ldots, z_N)}{\partial z_i}\right|_{(z_1,\ldots,z_N)=(1,\ldots,1)}, \qquad j, i = 1, \ldots, N. \tag{1.3}$$

Taking the second order (partial) derivatives yields another set of N^3 linear equations involving

$$f_j(i,k) = \left.\frac{\partial^2 F_j(z_1,\ldots,z_N)}{\partial z_i\, \partial z_k}\right|_{(z_1,\ldots,z_N)=(1,\ldots,1)}, \qquad j,i,k = 1,\ldots,N. \tag{1.4}$$

For the exhaustive and gated service policies, these sets of linear equations turn out to be closed and so can be solved numerically. Moreover, it can be shown that most performance measures of interest can be expressed in terms of the quantities defined in (1.3) and (1.4) above. For the special case where the queues are homogeneous, the sets of linear equations for the exhaustive and gated policies yield closed–form formulae for performance measures of interest.

The Limited Policy

Unfortunately for the limited policy, we cannot express most performance measures of interest in terms of $\{f_j(i),\ j,i = 1,\ldots,N\}$ and $\{f_j(i,k),\ j,i,k = 1,\ldots,N\}$ alone; other quantities are usually involved. Furthermore, taking derivatives of both sides of the relationships between F_{j+1} and F_j yields equations involving some other quantities besides $\{f_j(i),\ j,i = 1,\ldots,N\}$ and $\{f_j(i,k),\ j,i,k = 1,\ldots,N\}$. The only known case where these equations can be solved (i.e., there exist a sufficient number of equations for the unknowns) is the *single* service policy (i.e., a limited policy where $m_i = 1$ for $i = 1,\ldots,N$, also called the *limited–to–one* policy) with homogeneous queues. In this case, explicit formulae can actually be found, as shown by Watson [88] and Takagi [80]. The problem of analyzing the single service policy with $N = 2$ has been formulated as a *Riemann–Hilbert boundary value problem* and solved by Cohen and Boxma [18] (for zero switchover times), and Boxma [7]. Approximate analysis of the single service policy with arbitrary N and non-homogeneous queues has been done by Kuehn [51], and Boxma and Meister [11], among other people. Boxma and Meister based their approximations on the so–called *pseudo–conservation laws* which we shall discuss below. The limited policy with arbitrary limit parameters so far has hardly been analyzed; no exact results for this policy are available. Fuhrmann and Wang [34,35] proposed and studied approximation methods for this policy, extending the work by Boxma and Meister and exploiting upper bounds developed by Fuhrmann [32].

The Bernoulli Policy

The Bernoulli policy was introduced by Keilson and Servi [45] in the context of vacation models. The performance of polling systems under this service policy was approximated by Servi [72]. He considers N vacation models; each corresponds to a queue observed in isolation, with the time the server spends in other queues being taken as a vacation period for that queue. Assuming that a vacation time defined in this way is independent of the service times in the isolated queue, and assuming that we know how long the server stays in other queues per visit, we can use exact results for vacation models to compute how long the server stays in the isolated queue. Servi proposed an iterative algorithm based on this fact to compute the moments of the waiting times.

Pseudo–conservation Laws

Many recent approximation methods [11,26,28,34,35] for polling systems are based on the so–called pseudo–conservation (P–C) laws. A P–C law basically equates the weighted sum of the mean waiting times to a simple expression which depends only on the policy parameters and the first and second moments of the interarrival, service and switchover times distributions. This law extends to polling systems (with Poisson arrivals) the classical conservation law for $M/GI/1$ queues established by Kleinrock [49]. P–C laws have been established for other service policies. Ferguson and Aminetzah [30], and Watson [88] independently established this law for the exhaustive and gated policies; in [88], Watson also presented the law for the single policy. In [27,28], Everitt found explicit forms of the P–C law for variants of the limited policies. For these policies, however, the pseudo–conservation law still contains some unknowns which were identified as the second factorial moments of the numbers of customers served in one service period at various queues. An approximation to these unknowns is investigated in [27]. Fuhrmann [31] used decomposition results for vacation models established by Fuhrmann and Cooper [33] (see below) to derive P–C laws for homogeneous polling systems under the exhaustive, gated and single policies. Boxma and Groenendijk [10] extended Fuhrmann's arguments to establish a work decomposition result for polling systems with any service policy.

Vacation Models

Vacation models with Poisson arrivals under exhaustive policies were first thoroughly analyzed by Levy and Yechiali [58], although these systems have been used by Avi–Itzhak et al. [3] and Cooper [19] to analyze polling systems. Levy and Yechiali also considered a variant of a vacation model in which the server when becoming idle takes only one vacation and then waits for the arrival of the next customer (if it is not yet waiting in the queue). Doshi [22] later termed such a model a *single–vacation model*; he called our vacation model a *multi–vacation model.* Another related system is a single–server queue with the so–called *initial set–up times.* In this system, the server is turned off whenever it becomes idle. If the server is in this state when a customer arrives, it requires an initial set–up time before it can serve the customer. This system was introduced by Scholl and Kleinrock [71]. In all the above–mentioned systems, some kind of *decomposition results* were obtained. Basically, these results decompose a performance measure of interest such as waiting time or queue length into two terms: that of the corresponding $M/G/1$ system and another term which is associated with the vacations (or initial set–up times). Fuhrmann and Cooper [33] showed that these results hold for more general form of "vacations" using a unified approach.

Decomposition results have also been extended to $GI/GI/1$ vacation models. Gelenbe and Iasnogorodski [36] used arguments from the theory of complex variables. Doshi [21] employed sample path arguments similar to the ones used by Levy and Kleinrock [56] to show decomposition results for $M/G/1$ vacation models. Keilson and Servi showed decomposition results for $GI/GI/1$ vacation models under a Bernoulli policy by making use of arguments similar to those used by Gelenbe and

Iasnogorodski [36].

1.4 Thesis Summary

In this thesis, we study approaches which so far have been hardly used by researchers in this field, namely light and heavy traffic analysis, and stochastic comparison techniques. The thesis is divided into two parts: Part 1 consists of Chapters 2–5 while Part 2 comprises Chapters 6 and 7. In the first part, we consider vacation models; in the second part, the results obtained for vacation models are used to study polling systems. We conclude the thesis in Chapter 8 by making some observations and recommendations for possible future extensions to the obtained results. In the following, we summarize the main results of each chapter and indicate how they relate to one another.

In Chapter 2, we establish heavy traffic limit theorems for vacation models. Heavy traffic analysis is concerned with finding the limits of (normalized) quantities of interest as the traffic intensity approaches its critical value. Heavy traffic limit theorems are generally obtained by means of functional central limit theorems with the help of the theory of weak convergence. For the exhaustive and Bernoulli policies, we exploit a sample path representation of the vacation model in terms of the sample path of a standard $GI/GI/1$ queue. Using this representation, coupled with the so–called converging together theorem, we obtain the heavy traffic results for the exhaustive and Bernoulli policies from available results for the $GI/GI/1$ queue. For the limited policy, however, this sample path representation is not available, and so we are forced to take a different approach. We use a method which finds for each quantity of interest an auxiliary quantity which is more amenable to heavy traffic analysis, and at the same time behaves very closely to the original quantity of interest in heavy traffic.

In Chapter 3, we analyze vacation models in light traffic. Light traffic analysis studies the system behavior as the traffic rate approaches zero. In addition to the limits of (unnormalized) quantities of interest, light traffic analysis also yields the derivatives of the quantities of interest (with respect to the traffic rate) as the traffic rate goes to zero. We use a technique, developed by Reiman and Simon, where the computation of the nth derivative typically involves n customers in the system. The light and heavy traffic results are then combined to form the bases of an interpolation approximation method for vacation models, especially for those under the limited policy. Numerical results indicate that this approximation method performs very well.

In Chapter 4, we use stochastic ordering techniques to develop a framework in which two service policies can be compared. We compare two service policies by comparing some quantities of interest in a vacation model under one service policy to the corresponding quantities under the other service policy. We use an ordering called the *stochastic ordering* on random processes. We show that two service policies can be compared in this manner under some fairly general conditions. The setting in which the comparisons are made is also very general, as we make little assumptions on the probabilistic structure of the interarrival, service and vacation processes of the vacation models considered. Furthermore, for some quantities of interest, the comparisons can be made independently of the order in which the customers are served. We then show that various service policies in the literature can be stochastically compared using the general framework. We also establish stochastic monotonicity results for various classes of parametrizable service policies.

Unfortunately, the Limited and the Bernoulli policy cannot be compared in the general framework developed in Chapter 4. In Chapter 5, we show that these two policies can be compared in a weaker ordering, namely, the increasing convex ordering. This comparison proves to be significantly harder to establish; we prove the result for the special case where the vacation times are deterministic. The comparison is shown to hold both in the transient and steady–state regimes.

In Chapter 6 and 7, we consider polling systems under the Bernoulli and limited service policies, respectively. For the Bernoulli policy, we establish a pseudo–conservation law with the help of a work decomposition result obtained by Boxma and Groenendijk [9]. For the case where all the queues are homogeneous, the pseudo–

conservation law readily yields a closed–form formula for the (common) mean waiting time. For the limited policy, we propose and study an approximation scheme which exploits the interpolation approximation for vacation models developed in Chapter 3.

PART I

VACATION MODELS

CHAPTER 2

HEAVY TRAFFIC LIMIT THEOREMS FOR VACATION MODELS

2.1 Introduction

In this chapter, we establish heavy traffic limit theorems for vacation models under the exhaustive, Bernoulli and limited policies. In Chapter 3, these results are combined with light traffic results to form a basis for interpolation approximations.

The theory of heavy traffic analysis of queueing systems was pioneered by Kingman [47,48] who used the term heavy traffic to refer to queueing systems with traffic intensity less than but close to one. In [47,48], Kingman considers a sequence $\{G^r,\ r = 1, 2, \ldots\}$ of stable $GI/GI/1$ queues, each with traffic intensity $\rho^r < 1$ and limiting waiting time distribution W^r. He shows that $(1-\rho^r)W^r$ converges in distribution to a negative exponential distribution as $\rho^r \uparrow 1$ in some fashion. Extending Kingman's work, Prohorov [65] and Viskov [87] obtained the limit in distribution for the normalized version of the double sequence $\{W_n^r,\ n, r = 1, 2, \ldots\}$ of actual waiting times as n and r tend to infinity simultaneously. They consider the situations where $(\rho^r - 1)\sqrt{n}$ approaches c with (i) $-\infty < c < \infty$, (ii) $c = -\infty$, and (iii) $c = \infty$ as $n, r \to \infty$. By doing this, they expanded the notion of heavy traffic to include sequences of unstable queueing systems as well. In [89], Whitt considers a (single) sequence $\{\omega^r,\ r = 1, 2, \ldots\}$ of random functions defined by

$$\omega^r(t) = \frac{1}{\sqrt{r}} W^r_{\lfloor rt \rfloor}, \qquad 0 \le t \le 1, \tag{2.1}$$

where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x. He shows that under certain conditions, this sequence converges weakly to some diffusion process. The primary tools in his analysis are the so-called functional central limit theorems on the space $D[0,1]$. (Theorem A.4 in Appendix A). Here and in the sequel, we use $D[0,1]$ to denote the space of real-valued right-continuous functions on $[0,1]$ with left–hand limits, endowed with the Skorohod metric. The reader is referred to [17], [54] and [91] for surveys of heavy traffic results.

Since heavy traffic limit theorems for vacation models rely heavily on the available results for $GI/GI/1$ queues, we shall discuss these results in the next section. In Section 3, we use these results to establish the heavy traffic limit theorems for vacation models under the exhaustive service policy. In Section 4 and 5, we discuss the limit theorems for the vacation models under the Bernoulli and limited service policies, respectively.

For the reader's convenience, important theorems and definitions in weak convergence used throughout the discussion on heavy traffic are collected in Appendix A. A word on the notation used in this chapter: For a sequence $\{Y_n,\ n = 1, 2, \ldots\}$ of i.i.d. random variables, we use $\bar{Y}$ and σ_Y to denote the common mean and standard deviation of Y_n, respectively.

2.2 Heavy Traffic Limit Theorems for GI/GI/1 Queues

In this section we review heavy traffic results pertaining to $GI/GI/1$ queues. Following both historical and logical sequence, we shall discuss waiting time, queue size, and virtual waiting time processes in that order.

2.2.1 The Probabilistic Setting

Let $(\Omega, \mathcal{F}, P)$ be a probability space rich enough to support a sequence $\{G^r,\ r = 1, 2, \ldots\}$ of $GI/GI/1$ queues with first come first serve (FCFS) service discipline. For each $r = 1, 2, \ldots,\ G^r$ is represented by two independent sequences of i.i.d. nonnegative random variables, $\{A^r_n,\ n = 0, 1, \ldots\}$ and $\{B^r_n,\ n = 0, 1, \ldots\}$. For $n = 0, 1, \ldots,$ we interpret B^r_n as the nth service time and A^r_{n+1} as the time between the nth and the $(n+1)$th arrivals in G^r with the understanding that A^r_0 is the arrival time of the first customer. We assume that for each $r = 1, 2, \ldots,$ the first customer in G^r arrives at time 0 (i.e., $A^r_0 = 0$) to an empty queue.

For easy reference, we shall state here a definition borrowed from [52] which we use throughout this chapter.

Definition 2.1 (Condition A) *A double sequence $\{Y_n^r,\ n, r = 1, 2, \ldots\}$ of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ is said to satisfy Condition A if*

1. *For each $r = 1, 2, \ldots$, $\{Y_n^r, n = 1, 2, \ldots\}$ is an i.i.d. sequence of random variables with mean $\bar{Y}^r$ and standard deviation σ_Y^r.*

2. *$\bar{Y}^r \to \bar{Y} < \infty$ and $0 < \sigma_Y^r \to \sigma_Y$ with $0 < \sigma_Y < \infty$.*

3. *For some $\epsilon > 0$, $\sup_r E\left[|Y_n^r|^{2+\epsilon}\right] < \infty$.*

The sequences $\{A_n^r,\ n, r = 1, 2, \ldots\}$ and $\{B_{n-1}^r,\ n, r = 1, 2, \ldots\}$ are both assumed to satisfy Condition A. If we define

$$X_n^r = B_{n-1}^r - A_n^r, \qquad n, r = 1, 2, \ldots, \tag{2.2}$$

then clearly $\{X_n^r,\ n, r = 1, 2, \ldots\}$ also satisfies Condition A.

We impose the following conditions (C2.1) and (C2.2) where

(C2.1) $\bar{X}^r \sqrt{r}$ approaches C with $-\infty < C < 0$ as $r \to \infty$;

(C2.2) $0 < \bar{A}^r \to \bar{A}$ and $0 < \bar{B}^r \to \bar{B}$ with $0 < \bar{A}, \bar{B} < \infty$.

We define a continuous *reflection* mapping $f\colon D[0,1] \to D[0,1]$ by

$$f(y)(t) = \sup_{0 \le s \le t} [y(t) - y(s)], \qquad 0 \le t \le 1,\ y \in D[0,1], \tag{2.3}$$

and denote a Wiener process defined on $[0,1]$ by $\mathcal{W}$. We also let $\xrightarrow{\mathcal{D}}$ denote weak convergence.

2.2.2 The Actual Waiting Time Process

For each $r = 1, 2, \ldots$, let $\{Z_n^r,\ n = 0, 1, \ldots\}$ be the random walk associated with $\{X_n^r,\ n = 1, 2, \ldots\}$, i.e.,

$$Z_0^r = 0 \quad \text{and} \quad Z_n^r = \sum_{j=1}^{n} X_j^r, \qquad n = 1, 2, \ldots. \tag{2.4}$$

It can be shown [60] that the waiting time W_n^r of the nth customer ($n = 0, 1, \dots$) in G^r can be expressed as

$$W_0^r = 0 \quad \text{and} \quad W_n^r = \max\{Z_n^r - Z_k^r,\ k = 0, \dots, n\}, \qquad n = 1, 2, \dots. \tag{2.5}$$

For $r = 1, 2, \dots$, let ζ^r be $D[0,1]$-valued random functions defined by

$$\zeta^r(t) = \frac{1}{\sqrt{r}} \left(Z^r_{\lfloor rt \rfloor} - \bar{X}^r \lfloor rt \rfloor \right), \qquad 0 \le t \le 1 \tag{2.6}$$

and d^r be a deterministic function in $D[0,1]$ defined by

$$d^r(t) = \frac{\lfloor rt \rfloor \bar{X}^r}{\sqrt{r}}, \qquad 0 \le t \le 1. \tag{2.7}$$

From (2.1), (2.3) and (2.5)-(2.7), we have

$$\omega^r = f(\zeta^r + d^r), \qquad r = 1, 2, \dots. \tag{2.8}$$

We see that as an immediate consequence of Prohorov's theorem, $\zeta^r \xrightarrow{\mathcal{D}} \sigma_X \mathcal{W}$. Applying the continuous mapping theorem (Theorem A.1 in Appendix A) and noting from (C2.1) that

$$d^r(t) \to d(t) := C\, t, \qquad 0 \le t \le 1, \tag{2.9}$$

we have

$$\omega^r \xrightarrow{\mathcal{D}} f(\sigma_X \mathcal{W} + d). \tag{2.10}$$

2.2.3 The Queue Size Process

We have seen that the heavy traffic limit theorem for the actual waiting time process relies on the relationship (2.5) between the process and a random walk. This relationship allows us to use Prohorov's theorem directly. Unfortunately, such is not the case with the queue size process. For this process, we use a technique often used in heavy traffic theory where a modified process is analyzed in place of the process in question. This modified process is chosen such that its heavy traffic behavior can be analyzed with the help of either a random walk, a renewal counting process or the sum of random number of random variables. Furthermore, this modified process has the same heavy traffic limit as the original process.

To that end, for $r = 1, 2, \ldots,$ set

$$\tilde{A}^r(t) = \max\{n \geq 1 : A_0^r + \cdots + A_{n-1}^r \leq t\}, \qquad t \geq 0 \tag{2.11}$$

and

$$\tilde{B}^r(t) = \begin{cases} \max\{n \geq 1 : B_0^r + \cdots + B_{n-1}^r \leq t\} & \text{if } t \geq B_0^r \\ 0 & \text{otherwise,} \end{cases} \qquad t \geq 0. \tag{2.12}$$

We then define the modified queue size process $\{\hat{Q}^r(t),\ t \geq 0\}$ by

$$\hat{Q}^r(t) = \sup_{0 \leq s \leq t} [H^r(t) - H^r(s)], \qquad t \geq 0, \tag{2.13}$$

where

$$H^r(t) = \tilde{A}^r(t) - \tilde{B}^r(t), \qquad t \geq 0. \tag{2.14}$$

We can think of $\hat{Q}^r(t)$ as the queue size process of a modified $GI/GI/1$ queue where the server performs services one after another whether or not there is any customer in the queue. A "potential service completion" signifies a real departure epoch if the queue is not empty at that time; otherwise, nothing happens.

Define the following random functions taking values in $D[0,1]$ $(r = 1, 2, \ldots)$:

$$\begin{aligned} \alpha^r(t) &= \frac{1}{\sqrt{r}}\left[\tilde{A}^r(rt) - \frac{rt}{\bar{A}^r}\right], && (2.15) \\ \beta^r(t) &= \frac{1}{\sqrt{r}}\left[\tilde{B}^r(rt) - \frac{rt}{\bar{B}^r}\right], && (2.16) \\ \eta^r(t) &= \frac{1}{\sqrt{r}}\left[H^r(rt) - \left(\frac{1}{\bar{A}^r} - \frac{1}{\bar{B}^r}\right)rt\right] = \alpha^r(t) - \beta^r(t), && (2.17) \\ \chi^r(t) &= \frac{1}{\sqrt{r}}\hat{Q}^r(rt) = f(\eta^r + e^r)(t), && 0 \leq t \leq 1, (2.18) \end{aligned}$$

where f is the reflection mapping (2.3) and e^r is the element of $D[0,1]$ defined by

$$e^r(t) = \left(\frac{1}{\bar{A}^r} - \frac{1}{\bar{B}^r}\right)\sqrt{r}\,t, \qquad 0 \leq t \leq 1. \tag{2.19}$$

By the central limit theorem for renewal processes (Theorem A.6 in Appendix A), we have

$$\alpha^r \xrightarrow{\mathcal{D}} \frac{\sigma_A}{\bar{A}^{3/2}}\mathcal{W}_1 \tag{2.20}$$

and

$$\beta^r \xrightarrow{\mathcal{D}} \frac{\sigma_B}{\bar{B}^{3/2}} \mathcal{W}_2, \tag{2.21}$$

where $\mathcal{W}_1$ and $\mathcal{W}_1$ are independent versions of the Wiener process. Using this independence we can show that

$$\eta^r \xrightarrow{\mathcal{D}} \gamma \mathcal{W}_3, \tag{2.22}$$

where $\gamma^2 = \sigma_A^2/\bar{A}^3 + \sigma_B^2/\bar{B}^3$ and $\mathcal{W}_3$ is a Wiener process. Applying the continuous mapping theorem, and noting from (C2.1) that

$$e^r(t) \rightarrow \frac{C}{\bar{A}\bar{B}} t = C\, t =: e(t), \qquad 0 \leq t \leq 1, \tag{2.23}$$

we then conclude that

$$\hat{\chi}^r \xrightarrow{\mathcal{D}} f(\gamma \mathcal{W}_3 + e). \tag{2.24}$$

For the queue size process of the $GI/GI/1$ system, define χ^r similarly to $\hat{\chi}^r$. It can be shown [41] that

$$m(\chi^r, \hat{\chi}^r) \xrightarrow{P} 0, \tag{2.25}$$

where m is the Skorohod metric defined on $D[0,1]$ and $\xrightarrow{P}$ denotes convergence in probability. Applying the converging together theorem (Theorem A.2 in Appendix A), we get

$$\chi^r \xrightarrow{\mathcal{D}} f(\gamma \mathcal{W}_3 + e). \tag{2.26}$$

2.2.4 The Virtual Waiting Time Process

As with the queue size process, we have to resort to a modified process to obtain heavy traffic limit theorems for the virtual waiting time process. For a $GI/GI/1$ queue, the virtual waiting time process is identical to the workload process which can be described as follows. Define for $r = 1, 2, \ldots$

$$T^r(t) = \sum_{i=1}^{\tilde{A}^r(t)} B_{i-1}^r, \qquad t \geq 0. \tag{2.27}$$

This process represents the total amount of work in terms of servicing time that has arrived up to time t. Since the amount of work is depleted with rate one if any is present, the workload at time t is given by

$$U^r(t) = \sup_{0 \le s \le t} [Y^r(t) - Y^r(s)], \qquad t \ge 0, \tag{2.28}$$

where

$$Y^r(t) = T^r(t) - t, \qquad t \ge 0. \tag{2.29}$$

We shall not study directly the heavy traffic behavior of $U^r(t)$ but instead, we consider

$$\begin{aligned} \hat{Y}^r(t) &= \sum_{i=1}^{\tilde{A}^r(t)} X_i^r \\ &= T^r(t) - \sum_{i=1}^{\tilde{A}^r(t)} A_i^r, \qquad t \ge 0. \end{aligned} \tag{2.30}$$

Define the following random functions taking values in $D[0,1]$ $(r = 1, 2, \ldots)$:

$$\begin{aligned} \psi^r(t) &= \frac{1}{\sqrt{r}} Y^r(rt), && (2.31) \\ v^r(t) &= \frac{1}{\sqrt{r}} U^r(rt) = f(\psi^r)(t), && (2.32) \\ \hat{\psi}^r(t) &= \frac{1}{\sqrt{r}} \hat{Y}^r(rt), && (2.33) \\ \tau^r(t) &= \frac{1}{\sqrt{r}} \tilde{A}^r(rt)\bar{X}^r, && (2.34) \\ \phi^r(t) &= \frac{1}{\sqrt{r}} \left[\hat{Y}^r(rt) - \tilde{A}^r(rt)\bar{X}^r\right] = \hat{\psi}^r(t) - \tau^r(t), \qquad 0 \le t \le 1. && (2.35) \end{aligned}$$

By the central limit theorem for random sums of random variables (Theorem A.5 in Appendix A), we have

$$\phi^r \xrightarrow{\mathcal{D}} \frac{\sigma_X}{\sqrt{A}} \mathcal{W}. \tag{2.36}$$

Kyprianou [52, Lemma 1] shows that $\tau^r \xrightarrow{\mathcal{D}} \tau$ where τ is defined by

$$\tau(t) = \frac{C}{A} t, \qquad 0 \le t \le 1. \tag{2.37}$$

By [6, Theorem 4.4] and the fact that τ is deterministic, we have

$$(\phi^r, \tau^r) \xrightarrow{\mathcal{D}} \left(\frac{\sigma_X}{\sqrt{A}} \mathcal{W}, \tau\right), \tag{2.38}$$

and so since $\hat{\psi}^r = \phi^r + \tau^r$, we have

$$\hat{\psi}^r \xrightarrow{\mathcal{D}} \frac{\sigma_X}{\sqrt{A}}\mathcal{W} + \tau \tag{2.39}$$

by the continuous mapping theorem. It can be shown [52] that $m(\psi^r, \hat{\psi}^r) \xrightarrow{P} 0$, so that

$$\psi^r \xrightarrow{\mathcal{D}} \frac{\sigma_X}{\sqrt{A}}\mathcal{W} + \tau \tag{2.40}$$

by the converging together theorem. Since $v^r = f(\psi^r)$, we then have

$$v^r \xrightarrow{\mathcal{D}} f\left(\frac{\sigma_X}{\sqrt{A}}\mathcal{W} + \tau\right) \tag{2.41}$$

by the continuous mapping theorem.

2.3 Heavy Traffic Limit Theorems for Vacation Models under the Exhaustive Service Policy

We are now ready to establish heavy traffic results for vacation models. It is rather obvious intuitively that in heavy traffic, a vacation model under the exhaustive service policy behaves very similarly to a $GI/GI/1$ queue. This is true because as the traffic intensity becomes larger, the server has less time to take a vacation. It is also apparent from the fact that the stability condition for a vacation model under the exhaustive policy is equivalent to that of a $GI/GI/1$ system [21]. In this section, this idea is made rigorous.

2.3.1 The Probabilistic Setting

Let $\{E^r,\ r = 1, 2, \ldots\}$ be a sequence of vacation models under the exhaustive service policy, all defined on a common probability space $(\Omega, \mathcal{F}, P)$. For each $r = 1, 2, \ldots$, E^r is represented by three i.i.d. sequences of nonnegative random variables, namely, $\{A_n^r,\ n = 0, 1, \ldots\}$, $\{B_n^r,\ n = 0, 1, \ldots\}$, and $\{V_n^r,\ n = 0, 1, \ldots\}$. We assume that these three random sequences are mutually independent. For each $r = 1, 2, \ldots$, the first two sequences have the same interpretation as in the $GI/GI/1$ system G^r; the last sequence is the sequence of vacation lengths. The system E_r operates as follows: At the end of the nth service completion, the server immediately serves the $(n+1)$st customer if it is waiting in the queue (i.e., the service is on a FCFS basis); otherwise, the server starts a vacation. At the end of this vacation, the server immediately serves the $(n+1)$st customer if it has arrived by this time; otherwise, the server continues taking additional vacations until the $(n+1)$st customer is waiting in the queue when the server returns from a vacation. The lengths of the vacations are taken from the sequence $\{V_n^r,\ n=0,1,\ldots\}$, i.e., the length of the nth vacation is V_n^r. We shall assume that the sequences $\{A_{n-1}^r,\ n,r=1,2,\ldots\}$, $\{B_{n-1}^r,\ n,r=1,2,\ldots\}$ and $\{V_{n-1}^r,\ n,r=1,2,\ldots\}$ all satisfy Condition A and that conditions (C2.1) and (C2.2) are satisfied. We also assume that for each $r=1,2,\ldots,$ the first customer arrives at E^r at time $t=0$ (i.e., $A_0^r=0$) to an empty queue, and immediately receives service.

2.3.2 The Virtual Waiting Time Process

It is more convenient to consider the virtual waiting time process first. This process is equivalent to the workload process if we interpret a vacation as work, i.e., when the system becomes empty, the server, instead of starting a vacation of length V_n^r, adds V_n^r to the workload. Denote the workload at time t for E^r by $U^{E^r}(t)$; for its counterpart in G^r, use $U^r(t)$ as before. Let the sequence $\{v^{E^r},\ r=1,2,\ldots\}$ of $D[0,1]$-valued random functions be defined by

$$v^{E^r}(t)=\frac{1}{\sqrt{r}}U^{E^r}(rt), \qquad 0\le t\le 1;\ r=1,2,\ldots. \tag{2.42}$$

We want to show that $m(v^{E^r},v^r)\xrightarrow{P}0$ such that, by the converging together theorem and (2.41), we have

$$v^{E^r}\xrightarrow{\mathcal{D}} f\left(\frac{\sigma_X}{\sqrt{\bar{A}}}\mathcal{W}+\tau\right). \tag{2.43}$$

We shall make use of the following lemma which follows directly from [52, Lemma 3].

Lemma 2.1 *Let $\{Y_n^r,\ n,r=1,2,\ldots\}$ be a double sequence of nonnegative random variables satisfying Condition A. For $r=1,2,\ldots,$ let $\{\tilde{Y}^r(t),\ t\ge 0\}$ be the renewal* *counting process induced by the sequence* $\{Y_n^r,\ n = 1, 2, \ldots\}$, *i.e.,*

$$\tilde{Y}^r(t) = \begin{cases} \max\{k \geq 1 : Y_1^r + \cdots + Y_k^r \leq t\} & \textit{if } t \geq Y_1^r \\ 0 & \textit{otherwise.} \end{cases} \qquad t \geq 0. \tag{2.44}$$

Then, as $r \to \infty$, *we have*

$$\frac{1}{\sqrt{r}} \max_{1 \leq k \leq \tilde{Y}^r(r)} Y_k^r \xrightarrow{P} 0. \tag{2.45}$$

For $r = 1, 2, \ldots$ and $t \geq 0$, define $I^r(t)$ to be the total amount of time the server in G^r has spent idling (i.e., not serving a customer) in the interval $[0, t]$. For $r = 1, 2, \ldots$, define

$$n^r(t) = \min\left\{k \geq 1 : \sum_{i=0}^{k-1} V_i^r \geq I^r(t)\right\}, \qquad t \geq 0. \tag{2.46}$$

Then, $n^r(t)$ is the number vacations taken up to time t, including the one that might be in progress. By an argument similar to that of Doshi [21], we find that

$$U^{E^r}(t) = U^r(t) + \sum_{i=0}^{n^r(t)-1} V_i^r - I^r(t), \qquad t \geq 0. \tag{2.47}$$

Since

$$m(v^{E^r}, v^r) \leq \sup_{0 \leq t \leq 1} |v^{E^r}(t) - v^r(t)|, \qquad r = 1, 2, \ldots \tag{2.48}$$

[41, p. 156], we conclude from (2.32), (2.42), and (2.47) that

$$\begin{aligned} m(v^{E^r}, v^r) &\leq \frac{1}{\sqrt{r}} \sup_{0 \leq t \leq 1} \left[\sum_{i=0}^{n^r(rt)-1} V_i^r - I^r(rt)\right] \\ &\leq \frac{1}{\sqrt{r}} \max_{0 \leq i \leq n^r(r)-1} V_i^r, \qquad r = 1, 2, \ldots, \end{aligned} \tag{2.49}$$

where the second inequality follows from the definition (2.46) of $n^r(t)$. Define

$$\tilde{V}^r(t) = \begin{cases} \max\{k \geq 1 : V_0^r + \cdots + V_{k-1}^r \leq t\} & \text{if } t \geq V_0^r \\ 0 & \text{otherwise,} \end{cases} \qquad t \geq 0, \tag{2.50}$$

which can be interpreted as the number of vacations completed in E^r up to time t assuming that the server always takes vacations. Clearly, $n^r(r) \leq \tilde{V}^r(r)$ and therefore

$$m(v^{E^r}, v^r) \leq \frac{1}{\sqrt{r}} \max_{0 \leq i \leq \tilde{V}^r(r)-1} V_i^r. \tag{2.51}$$

By Lemma 2.1, the right hand side of the inequality above goes to zero in probability as $r \to \infty$, and so does $m(v^{E^r}, v^r)$. Therefore we have proved

Theorem 2.1 *Let $\{E^r,\ r = 1, 2, \ldots\}$ be a sequence of vacation models under the exhaustive service policy satisfying conditions (C2.1) and (C2.2). Assume that $\{A^r_{n-1},\ n, r = 1, 2, \ldots\}$, $\{B^r_{n-1},\ n, r = 1, 2, \ldots\}$, and $\{V^r_{n-1},\ n, r = 1, 2, \ldots\}$ all satisfy Condition A. Then*

$$v^{E^r} \xrightarrow{\mathcal{D}} f\left(\frac{\sigma_X}{\sqrt{A}}\mathcal{W} + \tau\right), \tag{2.52}$$

where v^{E^r} and β are defined in (2.42) and in (2.37), respectively.

2.3.3 The Actual Waiting Time Process

The waiting time of a customer is just the virtual waiting time evaluated at the customer's arrival epoch. More precisely, if we define

$$C^r_n = A^r_0 + \cdots + A^r_n, \qquad n = 0, 1, \ldots; r = 1, 2, \ldots, \tag{2.53}$$

then the waiting time of the nth customer in G^r can be expressed as

$$W^r_n = U^r(C^r_n), \qquad n = 0, 1, \ldots; r = 1, 2, \ldots. \tag{2.54}$$

Similarly, for E^r we have

$$W^{E^r}_n = U^{E^r}(C^r_n), \qquad n = 0, 1, \ldots; r = 1, 2, \ldots. \tag{2.55}$$

Using (2.47), we can relate $W^{E^r}_n$ to W^r_n by

$$W^{E^r}_n = W^r_n + \sum_{i=0}^{n^r(C^r_n)-1} V^r_i - I^r(C^r_n). \tag{2.56}$$

Define a sequence $\{\omega^{E^r},\ r = 1, 2, \ldots\}$ of $D[0, 1]$–valued random functions in the same manner as we define $\{\omega^r,\ r = 1, 2, \ldots\}$ in (2.1), i.e., for $r = 1, 2, \ldots$, we let

$$\omega^{E^r}(t) = \frac{1}{\sqrt{r}} W^{E^r}_{\lfloor rt \rfloor}, \qquad 0 \le t \le 1. \tag{2.57}$$

We want to show the following result.

Theorem 2.2 *Let $\{E^r,\ r = 1, 2, \ldots\}$ be a sequence of vacation models as in Theorem 2.1. Then*

$$\omega^{E^r} \xrightarrow{\mathcal{D}} f(\sigma_X \mathcal{W} + d). \tag{2.58}$$

Proof: It suffices to show that $m(\omega^{E^r}, \omega^r) \xrightarrow{P} 0$. Following the same arguments used for the virtual waiting time process and using equations (2.1), (2.57), and (2.56), we have

$$m(\omega^{E^r}, \omega^r) \leq \frac{1}{\sqrt{r}} \max_{0 \leq i \leq \tilde{V}^r(C^r_r)-1} V^r_i, \qquad r = 1, 2, \ldots . \tag{2.59}$$

So, $m(\omega^{E^r}, \omega^r) \xrightarrow{P} 0$ if the right hand side of (2.59) goes to 0 in probability. This will be shown in Lemma 2.3 which in turn relies on the following simple result.

Lemma 2.2 *Let $\{X^r_n,\ n, r = 1, 2, \ldots\}$ be a double sequence satisfying Condition A. Define $\{Z^r_n,\ n, r = 1, 2, \ldots\}$ by*

$$Z^r_n = X^r_1 + \cdots + X^r_n \qquad n, r = 1, 2, \ldots . \tag{2.60}$$

Then, $Z^r_r / r \rightarrow \bar{X}$ in probability.

Proof: In fact, we have a convergence in the L^2 sense, which of course implies convergence in probability. We have

$$\begin{aligned} E\left[(\frac{Z^r_r}{n} - \bar{X})^2\right] &= E\left[(\frac{Z^r_r}{r} - \bar{X}^r)^2\right] + (\bar{X}^r - \bar{X})^2 \\ &= \frac{1}{r}(\sigma^r_X)^2 + (\bar{X}^r - \bar{X})^2. \end{aligned} \tag{2.61}$$

Since the double sequence $\{X^r_n,\ n, r = 1, 2, \ldots\}$ satisfies Condition A, the mean $\bar{X}^r$ and the standard deviation σ^r_X each converges to a finite number, and so the right–hand side of (2.61) goes to zero. □

Lemma 2.3 *Let $\{Y^r_n,\ n, r = 1, 2, \ldots\}$ and $\{X^r_n,\ n, r = 1, 2, \ldots\}$ be two double sequences satisfying Condition A. Define $\{\tilde{Y}^r,\ r = 1, 2, \ldots\}$ as in Lemma 2.1 and $\{Z^r_{n-1},\ n, r = 1, 2, \ldots\}$ as in (2.60). Then*

$$\frac{1}{\sqrt{r}} \max\{Y^r_k, 1 \leq k \leq \tilde{Y}^r(Z^r_r)\} \xrightarrow{P} 0. \tag{2.62}$$

Proof: For each $r = 1, 2, \ldots$ and $\epsilon > 0$, we define the events $\mathcal{A}^\epsilon_r$ and $\mathcal{B}_r$, by

$$\mathcal{A}^\epsilon_r = \left\{ \frac{1}{\sqrt{r}} \max_{1 \leq k \leq \tilde{Y}^r(Z^r_r)} Y^r_k > \epsilon \right\} \tag{2.63}$$

and

$$\mathcal{B}_r = \left\{ \left| \frac{Z_r^r}{r} - \bar{X} \right| \leq |\bar{X}| \right\}. \tag{2.64}$$

Then,

$$P(\mathcal{A}_r^\epsilon) = P(\mathcal{A}_r^\epsilon \cap \mathcal{B}_r) + P(\mathcal{A}_r^\epsilon \cap \mathcal{B}_r^c), \qquad r = 1, 2, \ldots. \tag{2.65}$$

Observe that we have $Z_r^r \leq 2|\bar{X}|r$ on the set $\mathcal{B}_r$, and so

$$P(\mathcal{A}_r^\epsilon \cap \mathcal{B}_r) \leq P\left\{ \frac{1}{\sqrt{r}} \max_{1 \leq k \leq \tilde{Y}^r(2|\bar{X}|r)} Y_k^r > \epsilon \right\} \to 0 \qquad \text{as } r \to \infty \tag{2.66}$$

by Lemma 2.1. Furthermore,

$$P(\mathcal{A}_r^\epsilon \cap \mathcal{B}_r^c) \leq P(\mathcal{B}_r^c) \to 0 \qquad \text{as} \quad r \to \infty \tag{2.67}$$

by Lemma 2.2. So, $P(\mathcal{A}_r^\epsilon) \to 0$ as $r \to \infty$, and the proof is complete. □

2.3.4 The Queue Size Process

The queue size process will be analyzed with the help of the departure process. First we define $D^r(t)$ as the number of departures (service completions) up to and including time t in G^r and observe that the queue size at time t can be expressed as

$$Q^r(t) = \tilde{A}^r(t) - D^r(t), \qquad t \geq 0. \tag{2.68}$$

Similarly for E^r, define

$$Q^{E^r}(t) = \tilde{A}^r(t) - D^{E^r}(t), \qquad t \geq 0. \tag{2.69}$$

For $r = 1, 2, \ldots,$ define

$$\hat{\theta}^{E^r}(t) = U^{E^r}(t) - U^r(t) \qquad t \geq 0 \tag{2.70}$$

and

$$\theta^{E^r}(t) = \begin{cases} \hat{\theta}^{E^r}(t) & \text{if the server in } G^r \text{ is busy at time } t \\ \hat{\theta}^{E^r}(t') & \text{otherwise,} \end{cases} \qquad t \geq 0, \tag{2.71}$$

where t' is the last time before t the server in G^r is busy. The quantity $\theta^{E^r}(t)$ is piecewise constant with jumps at the start of busy period epochs and represents the difference between the time of occurrence of a departure in E^r and that of the corresponding departure in G^r. Hence

$$D^{E^r}(t) = D^r(t - \theta^{E^r}(t)), \qquad t \geq 0. \tag{2.72}$$

For $r = 1, 2, \ldots$, define

$$\chi^r(t) = \frac{1}{\sqrt{r}} Q^r(rt), \qquad 0 \leq t \leq 1 \tag{2.73}$$

and similarly

$$\chi^{E^r}(t) = \frac{1}{\sqrt{r}} Q^{E^r}(rt), \qquad 0 \leq t \leq 1. \tag{2.74}$$

Our goal is to show the following result.

Theorem 2.3 *Let $\{E^r,\ r = 1, 2, \ldots\}$ be a sequence of vacation models as in Theorem 2.1. Then*

$$\chi^{E^r} \xrightarrow{\mathcal{D}} f(\gamma \mathcal{W} + e). \tag{2.75}$$

Proof: We shall show that $m(\chi^{E^r}, \chi^r) \to 0$ in probability. Using equations (2.68), (2.69), and (2.72), we see that

$$\begin{aligned} m(\chi^{E^r}, \chi^r) &\leq \sup_{0 \leq t \leq 1} \left| \chi^{E^r}(t) - \chi^r(t) \right| \\ &= \frac{1}{\sqrt{r}} \sup_{0 \leq t \leq 1} [D^r(rt) - D^{E^r}(rt)] \\ &= \frac{1}{\sqrt{r}} \sup_{0 \leq t \leq 1} [D^r(rt) - D^r(rt - \theta^{E^r}(rt))]. \\ &\leq \frac{1}{\sqrt{r}} \sup_{0 \leq t \leq 1} \left[D^r(rt) - D^r\left(\left[rt - \sup_{t \leq s \leq 1} \theta^{E^r}(rs) \right]^+ \right) \right], \end{aligned} \tag{2.76}$$

where $[x]^+ = \max\{0, x\}$. So, to obtain the desired result, it suffices to show that both δ^r and $\hat{\delta}^r$ $(r = 1, 2, \ldots)$ defined by

$$\delta^r(t) = \frac{1}{\sqrt{r}} D^r(rt), \qquad 0 \leq t \leq 1 \tag{2.77}$$

and

$$\hat{\delta}^r(t) = \frac{1}{\sqrt{r}} D^r\left(\left[rt - \sup_{t \leq s \leq 1} \theta^{E^r}(rs) \right]^+ \right), \qquad 0 \leq t \leq 1 \tag{2.78}$$

converge weakly to the same random function. We shall do this with the help of the random time change theorem [41, Lemma 7.1], [6, p. 145]. See also Theorem A.3 in Appendix A.

For $r = 1, 2, \dots,$ define

$$\Phi^r(t) = \left[t - \sup_{t \le s \le 1} \frac{\theta^{E^r}(rs)}{r} \right]^+ \qquad 0 \le t \le 1. \tag{2.79}$$

Then for each r, Φ^r is an element of $D[0,1]$ which is nondecreasing and satisfies

$$0 \le \Phi^r(t) \le 1, \qquad 0 \le t \le 1. \tag{2.80}$$

Moreover, since

$$\begin{aligned} \sup_{0 \le t \le 1} |\Phi^r(t) - t| &\le \frac{1}{r} \sup_{0 \le t \le 1} \hat{\theta}^{E^r}(rt) \\ &= \frac{1}{\sqrt{r}} \sup_{0 \le t \le 1} \left[v^{E^r}(t) - v^r(t) \right] \xrightarrow{P} 0, \end{aligned} \tag{2.81}$$

(see our discussion on the virtual waiting time process), we then have

$$m(\Phi^r, I) \xrightarrow{P} 0, \tag{2.82}$$

where I is the identity mapping $[0,1] \to [0,1]$. Iglehart and Whitt [41, p. 163] show that δ^r converges weakly to some random function X which is continuous almost everywhere. So, by the random time change theorem and (2.82), we get

$$\hat{\delta}^r = \delta^r(\Phi^r) \xrightarrow{\mathcal{D}} X(I) = X, \tag{2.83}$$

and this completes the proof. □

2.4 Heavy Traffic Limit Theorems for Vacation Models under Bernoulli Service Policies

The heavy traffic results for vacation models under the exhaustive policy can be extended to vacation models under a Bernoulli service policy. A Bernoulli policy is a generalization of the exhaustive policy in that at a service completion, if the queue is not empty, the server does not necessarily serve the next customer; it does so with probability p, $0 \le p \le 1$, and takes a vacation with probability $1 - p$. The exact probabilistic model is stated in the next section. The virtual waiting time, the actual waiting time, and the queue size processes are then analyzed in that order.

2.4.1 The Probabilistic Setting

Let $\{B^r,\ r = 1, 2, \ldots\}$ be a sequence of vacation models defined on a common probability space $(\Omega, \mathcal{F}, P)$. For each $r = 1, 2, \ldots$, B^r is under the Bernoulli service policy with parameter p^r, $0 \le p^r \le 1$. The system B^r is described by the same three independent sequences governing E^r, namely $\{A^r_n,\ n = 0, 1, \ldots\}$, $\{B^r_n,\ n = 0, 1, \ldots\}$, and $\{V^r_n,\ n = 0, 1, \ldots\}$, together with an additional i.i.d. sequence of random variables $\{U^r_n,\ n = 0, 1, \ldots\}$ where

$$U^r_n = \begin{cases} 1 & \text{with probability } p^r \\ 0 & \text{with probability } 1 - p^r, \end{cases} \qquad n = 0, 1, \ldots;\ r = 1, 2, \ldots.$$

All four sequences are assumed to be mutually independent. The behavior of the server in B^r is the same as that in E^r except when the queue is not empty at a service completion. At the nth occurrence of this situation, the server in B^r looks at the value of U^r_n; it serves the next available customer if $U^r_n = 1$ and takes a vacation otherwise. Keilson and Servi [45] show that B^r is stable if and only if

$$\bar{B}^r + (1 - p^r)\bar{V}^r < \bar{A}^r.$$

For the purpose of our discussion, we shall assume that $\{A^r_{n-1},\ n, r = 1, 2, \ldots\}$, $\{B^r_{n-1},\ n, r = 1, 2, \ldots\}$, and $\{V^r_{n-1},\ n, r = 1, 2, \ldots\}$ all satisfy Condition A and that the following conditions are satisfied:

(C2.3) $(\bar{B}^r + (1 - p^r)\bar{V}^r - \bar{A}^r)\sqrt{r}$ approaches C_B with $-\infty < C_B < 0$ as $r \to \infty$;

(C2.4) $0 < \bar{A}^r \to \bar{A}$, $0 < \bar{B}^r \to \bar{B}$, $0 < \bar{V}^r \to \bar{V}$, and $p^r \to p$ with $0 < \bar{A}, \bar{B}, \bar{V} < \infty$ and $0 \le p^r, p \le 1$.

We also assume that for each $r = 1, 2, \ldots$, the first customer arrives at B^r at time $t = 0$ (i.e., $A^r_0 = 0$) to an empty queue and immediately receives service.

2.4.2 The Virtual and Actual Waiting Time Processes

It can be shown that, as far as waiting time processes are concerned, a vacation model under a Bernoulli policy with parameter p is equivalent to a vacation model under the exhaustive policy with modified service times. In this equivalent system, a service time is distributed as the original service time with probability p and as the original service time plus the vacation period with probability $1-p$. This result will be shown later in Chapter 5 (Lemma 5.4). Heuristically, the equivalence can be seen to hold from the following argument: At the service completion of the nth customer C_n, if customer C_{n+1} is present in the queue, the server serves C_{n+1} or takes a vacation and then serves C_{n+1}. The former occurs with probability p and the latter with probability $1-p$. As far as C_{n+1} is concerned, the vacation can be regarded as part of the service given to C_n. In the case where C_{n+1} has not arrived when C_n completes its service, the server takes a number of vacations until it finds C_{n+1} in the queue upon returning from a vacation. Whether the first of these vacations is part of the service given to C_n or not, does not change the waiting time of C_{n+1}.

Let $\{\hat{E}^r,\ r=1,2,\ldots\}$ be the sequence of the equivalent vacation models under the exhaustive policy. For each $r=1,2,\ldots$, $\hat{E}^r$ is described by the four sequences describing B^r, namely $\{A_n^r,\ n=0,1,\ldots\}$, $\{B_n^r,\ n=0,1,\ldots\}$, $\{V_n^r,\ n=0,1,\ldots\}$, and $\{U_n^r,\ n=0,1,\ldots\}$, and another sequence $\{\hat{V}_n^r,\ n=0,1,\ldots\}$. This sequence is independent of the other four and is identical in law to the sequence $\{V_n^r,\ n=0,1,\ldots\}$. Furthermore, $\{\hat{V}_n^r,\ n,r=0,1,\ldots\}$ is assumed to satisfy Condition A. The sequence $\{\hat{V}_n^r,\ n=0,1,\ldots\}$ is used in the modified service times in $\hat{E}^r$, i.e., the nth service time $\hat{B}_n^r$ in $\hat{E}^r$ is given by

$$\hat{B}_n^r = B_n^r + (1-U_n^r)\hat{V}_n^r, \qquad n=0,1,\ldots;\ r=1,2,\ldots. \tag{2.84}$$

Obviously, $\{\hat{B}_n^r,\ n=0,1,\ldots\}$ also satisfies Condition A. Using this fact and conditions (C2.3)–(C2.4), we can use the results established in the previous section to obtain the heavy traffic behavior of $\{\hat{E}^r,\ r=1,2,\ldots\}$, which is identical to that of $\{B^r,\ r=1,2,\ldots\}$. We then have the following results.

Theorem 2.4 *Let $\{B^r,\ r=1,2,\ldots\}$ be a sequence of vacation models under a Bernoulli service policy satisfying conditions (C2.3) and (C2.4). Assume that*

$\{A^r_{n-1},\ n, r = 1, 2, \ldots\}$, $\{B^r_{n-1},\ n, r = 1, 2, \ldots\}$, and $\{V^r_n,\ n, r = 1, 2, \ldots\}$ all satisfy Condition A. Then

$$v^{B^r} \xrightarrow{\mathcal{D}} f\left(\frac{\hat{\sigma}_X}{\sqrt{\bar{A}}}\mathcal{W} + \beta_B\right), \tag{2.85}$$

where

$$\hat{\sigma}_X^2 = \sigma_B^2 + (1-p)\sigma_V^2 + p(1-p)\bar{V}^2 + \sigma_A^2 \tag{2.86}$$

and $\beta_B(t) = (C_B/\bar{A})\,t$, $0 \le t \le 1$.

Theorem 2.5 *Let $\{B^r,\ r = 1, 2, \ldots\}$ be a sequence of vacation models as in Theorem 2.4. Then,*

$$\omega^{B^r} \xrightarrow{\mathcal{D}} f(\hat{\sigma}_X \mathcal{W} + d_B), \tag{2.87}$$

where $d_B(t) = C_B\,t$, $0 \le t \le 1$.

2.4.3 The Queue Size Process

Unlike the waiting time processes, the queue size process of a vacation model under a Bernoulli policy is slightly different from that of the equivalent vacation model under the exhaustive policy. This minor difference arises from the fact that for the equivalent system, a customer departs at the end of a modified service time which may include a vacation period while in the original system, the customer departs as soon as its real service is completed. However, this time difference cannot be larger than a vacation period and it will be shown below to be negligible in heavy traffic.

For $r = 1, 2, \ldots$ and $n = 0, 1, \ldots$, denote the nth departure epoch in B^r and in its equivalent system $\hat{E}^r$, by $T^{B^r}_n$ and $T^{\hat{E}^r}_n$, respectively. For $r = 1, 2, \ldots$ and $t \ge 0$, let $D^{B^r}(t)$ denote the number of service completions (departures) up to and including time t in B^r, and denote the same quantity in $\hat{E}^r$ by $D^{\hat{E}^r}(t)$. Defining

$$\theta^{B^r}(t) := \left[T^{B^r}_{D^{E^r}(t)} + (1 - U_{D^{E^r}(t)})\hat{V}^r_{D^{E^r}(t)} - t\right]^+, \qquad t \ge 0, \tag{2.88}$$

we then see that

$$D^{B^r}(t) = D^{\hat{B}^r}(t + \theta^{B^r}(t)), \qquad t \ge 0. \tag{2.89}$$

Since

$$\theta^{B^r}(t) \le \max_{0 \le i \le D^{E^r}(t)-1} \hat{V}_i^r, \qquad t \ge 0, \tag{2.90}$$

we can use the same argument as in Section 2.3.4 to show that in the heavy traffic limit, the queue size processes for both the original and the equivalent systems become identical. Hence,

Theorem 2.6 *Let $\{B^r,\ r = 1, 2, \ldots\}$ be a sequence of vacation models as in Theorem 2.4. Then,*

$$\chi^{B^r} \xrightarrow{\mathcal{D}} f(\gamma_B \mathcal{W} + e_B), \tag{2.91}$$

where

$$\gamma_B^2 = \frac{\sigma_A^2}{\bar{A}^3} + \frac{\sigma_{\hat{B}}^2}{\bar{\hat{B}}^3}, \tag{2.92}$$

$$\sigma_{\hat{B}}^2 = \sigma_B^2 + (1-p)\sigma_V^2 + p(1-p)\bar{V}^2, \tag{2.93}$$

$$\bar{\hat{B}} \doteq \bar{B} + (1-p)\bar{V}, \tag{2.94}$$

and $e_B(t) = C_B\, t$, $0 \le t \le 1$.

2.5 Heavy Traffic Limit Theorems for Vacation Models under Limited Service Policies

In this section, we establish heavy traffic limit theorems for the limited service policies. A limited service policy is characterized by an integer m. Under this service policy, the number of consecutive services that can be performed by the server between any two vacations is limited to m. Unlike for the Bernoulli policies, there is no known result for the limited policies which provides equivalent vacation models under the exhaustive policy (except for the case $m = 1$ which is equivalent to the Bernoulli policy with $p = 0$). Therefore, we cannot use the heavy traffic results for the exhaustive policy we established in Section 2.3 to study the heavy traffic behavior of the limited policies. We shall resort to two auxiliary systems which exhibit heavy traffic behavior very similar to that of a vacation model under a limited service policy.

2.5.1 The Probabilistic Setting

Let $\{L^r,\ r = 1,2,\ldots\}$ be a sequence of vacation models defined on a common probability space $(\Omega, \mathcal{F}, P)$. For each $r = 1,2,\ldots$, L^r is under the limited service policy with parameter m^r, $m^r = 1,2,\ldots$. The system L^r is represented by the same three random sequences describing E^r, namely $\{A_n^r,\ n = 0,1,\ldots\}$, $\{B_n^r,\ n = 0,1,\ldots\}$ and $\{V_n^r,\ n = 0,1,\ldots\}$. As usual we assume that these three random sequences are mutually independent. For each $r = 1,2,\ldots$, the operation of the system L^r is described as follows: At each service completion, the server starts a vacation if and only if either or both of these two conditions are satisfied: 1) the queue is empty 2) the server has performed m^r consecutive services since the last time it returned from a vacation. As in E^r, the server continues taking additional vacations until at least one customer is waiting in the queue when it returns from a vacation. We shall assume that customers are served on a FCFS basis. It can be shown (e.g., [51]) that for each $r = 1,2,\ldots$, the system L^r is stable if and only if

$$\bar{B}^r + \frac{\bar{V}^r}{m^r} < \bar{A}^r. \tag{2.95}$$

As usual, we assume that $\{A_{n-1}^r,\ n,r = 1,2,\ldots\}$, $\{B_{n-1}^r,\ n,r = 1,2,\ldots\}$ and $\{V_{n-1}^r,\ n,r = 1,2,\ldots\}$ each satisfies Condition A and that the additional conditions (C2.5) and (C2.6) are satisfied, where

(C2.5) $(\bar{B}^r + \bar{V}^r/m^r - \bar{A}^r)\sqrt{r}$ approaches C_L with $-\infty < C_L < 0$ as $r \to \infty$;

(C2.6) $0 < \bar{A}^r \to \bar{A}$, $0 < \bar{B}^r \to \bar{B}$, $0 < \bar{V}^r \to \bar{V}$ and $1 \le m^r \to m$, with $0 < \bar{A}, \bar{B}, \bar{V} < \infty$ and $1 \le m$.

We also assume that for each $r = 1,2,\ldots$, the 0th customer arrives to L^r at time $t = 0$ (i.e., $A_0^r = 0$) to an empty queue and immediately receives service.

2.5.2 The Auxiliary Systems

To obtain heavy traffic functional limit theorems for the sequence $\{L^r,\ r = 1,2,\ldots\}$, we shall use two sequences of auxiliary systems $\{\bar{L}^r,\ r = 1,2,\ldots\}$, and $\{\hat{L}^r,\ r = 1,2,\ldots\}$. The sequence $\{\bar{L}^r,\ r = 1,2,\ldots\}$ is defined in such a way that we can establish heavy traffic results for it using the functional central limit theorems for renewal processes. The sequence $\{\hat{L}^r,\ r = 1,2,\ldots\}$ serves the purpose of bridging between $\{\bar{L}^r,\ r = 1,2,\ldots\}$ and $\{L^r,\ r = 1,2,\ldots\}$.

We first describe the sequence $\{\hat{L}^r,\ r = 1,2,\ldots\}$. For each $r = 1,2,\ldots$, the server in $\hat{L}^r$ performs a series of m^r consecutive *potential services*, followed by one vacation, another series of m^r potential services, a vacation, and so on. The activity of the server is independent of the arrival process. The lengths of the potential services are taken from the sequence $\{B_n^r,\ n = 0,1,\ldots\}$, and the vacation lengths from the sequence $\{V_n^r,\ n = 0,1,\ldots\}$. The arrival process is governed by the sequence $\{A_n^r,\ n = 0,1,\ldots\}$. The potential services are so called because at each of their completions, a customer leaves the system (i.e., the potential service becomes a *real* service) if and only if the queue is not empty at that time. If the queue is empty, then the potential service is left unused. Note that it is possible for a customer to arrive in the middle of a potential service and then leave at the completion of the potential service. In this case, we can think of this customer as receiving a service of length equal to the remaining life of the potential service time.

We now describe the sequence $\{\bar{L}^r,\ r = 1,2,\ldots\}$. For each $r = 1,2,\ldots$, the system $\bar{L}^r$ is very similar to the system $\hat{L}^r$. The only difference is that in $\bar{L}^r$, we allow customers to depart only at vacation completion epochs. At each of these epochs, either a batch of m^r customers or all the available customers, whichever is fewer, leave the system simultaneously. In both $\hat{L}^r$ and $\bar{L}^r$, we adopt the FCFS service discipline.

2.5.3 The Queue Size Process

For each $r = 1,2,\ldots$, let $Q^{L^r}(t)$ denote the queue length of L^r at time t and define the $D[0,1]$–valued random function χ^{L^r} by

$$\chi^{L^r}(t) = \frac{1}{\sqrt{r}} Q^{L^r}(rt), \qquad 0 \le t \le 1. \tag{2.96}$$

Define $Q^{\hat{L}^r}(t)$, $Q^{\bar{L}^r}(t)$, $\chi^{\hat{L}^r}$ and $\chi^{\bar{L}^r}$ correspondingly.

For each $r = 1, 2, \ldots$, define the random sequence $\{R_n^r,\ n = 1, 2, \ldots\}$ by

$$R_n^r = B_{nm^r}^r + \cdots + B_{(n+1)m^r-1}^r + V_n^r, \qquad n = 0, 1, \ldots, \tag{2.97}$$

and let $\{\tilde{R}^r(t),\ t \geq 0\}$ be the renewal counting process induced by this sequence, i.e.,

$$\tilde{R}^r(t) = \begin{cases} \max\{k \geq 1 : R_0^r + \cdots + R_{k-1}^r \leq t\} & \text{if } t \geq R_0^r \\ 0 & \text{otherwise.} \end{cases} \qquad t \geq 0. \tag{2.98}$$

The following is the main result of this section.

Theorem 2.7 *We have*

$$\chi^{L^r} \xrightarrow{\mathcal{D}} f(\gamma^{\bar{L}}\mathcal{W} + e^{\bar{L}}) \tag{2.99}$$

as $r \to \infty$, *where*

$$\gamma^{\bar{L}} = \left(\frac{\sigma_A^2}{\bar{A}^3} + \frac{m^2\sigma_R^2}{\bar{R}^3}\right)^{\frac{1}{2}} \tag{2.100}$$

and

$$e^{\bar{L}}(t) = \frac{mC_L}{\bar{A}\bar{R}}t, \qquad 0 \leq t \leq 1. \tag{2.101}$$

This result directly follows from the following lemma.

Lemma 2.4 *As* $r \to \infty$, *we have*

$$\chi^{\bar{L}^r} \xrightarrow{\mathcal{D}} f(\gamma^{\bar{L}}\mathcal{W} + e^{\bar{L}}), \tag{2.102}$$

where $\gamma^{\bar{L}}$ *and* $e^{\bar{L}}$ *are defined in (2.100) and (2.101), respectively.*

Proof: For each $r = 1, 2, \ldots$, $Q^{\bar{L}^r}(t)$ can be expressed as

$$Q^{\bar{L}^r}(t) = \sup_{0 \leq s \leq t} \left[H^{\bar{L}^r}(t) - H^{\bar{L}^r}(s)\right], \qquad t \geq 0, \tag{2.103}$$

where

$$H^{\bar{L}^r}(t) = \tilde{A}^r(t) - m^r\tilde{R}^r(t), \qquad t \geq 0. \tag{2.104}$$

For each $r = 1, 2, \ldots$, define

$$\begin{aligned}
\alpha^r(t) &= \frac{1}{\sqrt{r}}\left[\tilde{A}^r(rt) - \frac{rt}{\bar{A}^r}\right], && (2.105)\\
\varrho^r(t) &= \frac{1}{\sqrt{r}}\left[\tilde{R}^r(rt) - \frac{rt}{\bar{R}^r}\right], && (2.106)\\
\eta^{\bar{L}^r}(t) &= \frac{1}{\sqrt{r}}\left[H^{\bar{L}^r}(rt) - rt(\frac{1}{\bar{A}^r} - \frac{m^r}{\bar{R}^r})\right], && (2.107)\\
&= \alpha^r(t) - \varrho^r(t), \qquad 0 \le t \le 1. && (2.108)
\end{aligned}$$

From (2.103), (2.107), and the definition of $\chi^{\bar{L}^r}$, we see that

$$\chi^{\bar{L}^r} = f(\eta^{\bar{L}^r} + e^{\bar{L}^r}), \tag{2.109}$$

where f is defined in (2.3) and $e^{\bar{L}^r}$ is a deterministic function in $D[0, 1]$ defined by

$$e^{\bar{L}^r}(t) = \sqrt{r}\left(\frac{1}{\bar{A}} - \frac{m^r}{\bar{R}^r}\right)t, \qquad 0 \le t \le 1. \tag{2.110}$$

From (C2.5), we see that

$$e^{\bar{L}^r}(t) \to e^{\bar{L}}(t), \qquad 0 \le t \le 1 \tag{2.111}$$

as $r \to \infty$. So, by the same argument as in Section 2.2.3, we obtain the desired result. □

Lemma 2.5 *As $r \to \infty$, we have*

$$m(\chi^{\bar{L}^r}, \chi^{\hat{L}^r}) \xrightarrow{P} 0. \tag{2.112}$$

Proof: As before, the inequalities

$$m(\chi^{\bar{L}^r}, \chi^{\hat{L}^r}) \le \frac{1}{\sqrt{r}} \sup_{0 \le t \le 1} |Q^{\bar{L}^r}(rt) - Q^{\hat{L}^r}(rt)|, \qquad r = 1, 2, \ldots \tag{2.113}$$

hold true. We shall show that for each $r = 1, 2, \ldots$,

$$|Q^{\bar{L}^r}(t) - Q^{\hat{L}^r}(t)| \le m^r, \qquad t \ge 0, \tag{2.114}$$

so that

$$m(\chi^{\bar{L}^r}, \chi^{\hat{L}^r}) \le \frac{1}{\sqrt{r}} m^r \to 0. \tag{2.115}$$

But,

$$|Q^{\bar{L}^r}(t) - Q^{\hat{L}^r}(t)| = |D^{\bar{L}^r}(t) - D^{\hat{L}^r}(t)|, \qquad t \geq 0, \tag{2.116}$$

where $D^{\bar{L}^r}(t)$ and $D^{\hat{L}^r}(t)$ are the number of departures up to time t from the system $\bar{L}^r$ and $\hat{L}^r$, respectively. Denote the nth vacation completion epoch (in both $\bar{L}^r$ and $\hat{L}^r$) by t_n^r. Then, it is easy to see that for each $n = 0, 1, \ldots,$ we have

$$D^{\hat{L}^r}(t_n^r) \leq D^{\bar{L}^r}(t_n^r), \tag{2.117}$$

$$D^{\hat{L}^r}(t) - D^{\hat{L}^r}(t_n^r) \leq m^r, \qquad t_n^r \leq t \leq t_{n+1}^r \tag{2.118}$$

and

$$D^{\bar{L}^r}(t) = D^{\bar{L}^r}(t_n^r), \qquad t_n^r \leq t \leq t_{n+1}^r. \tag{2.119}$$

Consequently, combining these three facts, we obtain

$$D^{\hat{L}^r}(t) - D^{\bar{L}^r}(t) \leq m^r, \qquad t \geq 0. \tag{2.120}$$

To show that

$$D^{\bar{L}^r}(t) - D^{\hat{L}^r}(t) \leq m^r, \qquad t \geq 0, \tag{2.121}$$

we use an argument by induction. Since $D^{\bar{L}^r}(t_0^r) \leq m^r$, we obviously have

$$D^{\bar{L}^r}(t_0^r) - D^{\hat{L}^r}(t_0^r) \leq m^r. \tag{2.122}$$

Now assume that for some $n = 0, 1, \ldots,$ we have

$$D^{\bar{L}^r}(t_n^r) - D^{\hat{L}^r}(t_n^r) \leq m^r. \tag{2.123}$$

In view of (2.119), we can conclude that

$$D^{\bar{L}^r}(t) - D^{\hat{L}^r}(t) \leq m^r, \qquad t_n^r \leq t \leq t_{n+1}^r. \tag{2.124}$$

We see that because of (2.123) and the fact that

$$Q^{\hat{L}^r}(t_n^r) \geq Q^{\hat{L}^r}(t_n^r) - Q^{\bar{L}^r}(t_n^r) = D^{\bar{L}^r}(t_n^r) - D^{\hat{L}^r}(t_n^r), \tag{2.125}$$

the number of service completions in $\hat{L}^r$ in the interval $(t_n^r, t_{n+1}^r]$ is at least $D^{\bar{L}^r}(t_n^r) - D^{\hat{L}^r}(t_n^r)$, i.e.,

$$D^{\hat{L}^r}(t_{n+1}^r) - D^{\hat{L}^r}(t_n^r) \geq D^{\bar{L}^r}(t_n^r) - D^{\hat{L}^r}(t_n^r), \tag{2.126}$$

or, equivalently,

$$D^{\hat{L}^r}(t^r_{n+1}) \geq D^{\bar{L}^r}(t^r_n). \tag{2.127}$$

Combining this with

$$D^{\bar{L}^r}(t^r_{n+1}) - D^{\bar{L}^r}(t^r_n) \leq m^r, \tag{2.128}$$

we get

$$D^{\bar{L}^r}(t^r_{n+1}) - D^{\hat{L}^r}(t^r_{n+1}) \leq m^r. \tag{2.129}$$

□

Lemma 2.6

$$m(\chi^{\hat{L}^r}, \chi^{L^r}) \xrightarrow{P} 0. \tag{2.130}$$

Proof: For each $r = 1, 2, \ldots$, we reconstruct each sample path of L^r from that of $\hat{L}^r$ such that

1. A bound for $m(\chi^{\hat{L}^r}, \chi^{L^r})$ which converges to 0 in probability as $r \to \infty$ can be easily identified;

2. The probabilistic structure of L^r is not altered.

The construction is carried out as follows: First, define for each $r = 1, 2, \ldots$ a random sequence $\{\check{V}^r_n,\ n = 0, 1, \ldots\}$ such that it is probabilistically identical to the sequence $\{V^r_n,\ n = 0, 1, \ldots\}$ and such that $\{\check{V}^r_{n+1},\ n, r = 1, 2, \ldots\}$ satisfies Condition A. For each $r = 1, 2, \ldots$, we call a customer in L^r an *initiating customer* if it arrives to an empty queue (e.g., the 0th customer). Denote the starting time of service given to the nth initiating customer by $s^{L^r}_n$ (e.g., $s^{L^r}_0 = 0$). Define the nth *busy period* to be the period that starts at time $s^{L^r}_n$ and ends at time $e^{L^r}_n$ when the queue empties for the first time since $s^{L^r}_n$. (Hence the nth initiating customer initiates the nth busy period.) The end of a busy period is necessarily a service completion which is immediately followed by a vacation, and any two busy periods are separated by at least one vacation. In $\hat{L}^r$, denote by $s^{\hat{L}^r}_n$ the first vacation completion epoch after time $s^{L^r}_n$, $n = 1, 2, \ldots$, and let $s^{\hat{L}^r}_0 = 0$.

We now construct the sample path of L^r from that of $\hat{L}^r$. In the 0th busy period, i.e., from time $t = 0$ $(= s_0^{L^r} = s_0^{\hat{L}^r})$ up to the first time the queue is empty, let the server in L^r follow exactly the server in $\hat{L}^r$, i.e., the server in L^r is serving a customer (resp. taking a vacation) whenever the server in $\hat{L}^r$ is. This means that during this period, the service and vacation times used in L^r are identical to those used in $\hat{L}^r$. At time $e_0^{L^r}$, the server in L^r starts a vacation and keeps taking additional vacations until it finds at least one customer waiting in the queue upon returning from a vacation. The lengths of these vacations are taken from the sequence $\{\check{V}_n^r,\ n = 0, 1, \ldots\}$. When the server in L^r returns from a vacation and finds at least one customer in the queue, it immediately starts serving. This vacation completion epoch is then the start of the next busy period. In this busy period (i.e., interval $[s_1^{L^r}, e_1^{L^r}))$, again we let the server in L^r follows the server in $\hat{L}^r$. This time we introduce a delay, i.e., the action of the server in L^r in the interval $[s_1^{L^r}, e_1^{L^r})$ is identical to that of the server in $\hat{L}^r$ not in the same interval but in the interval $[s_1^{\hat{L}^r}, s_1^{\hat{L}^r} + e_1^{L^r} - s_1^{L^r})$. In general, for each $n = 0, 1, \ldots,$ the action of the server in L^r in the interval $[s_n^{L^r}, e_n^{L^r})$ is identical to that of the server in $\hat{L}^r$ in the interval $[s_n^{\hat{L}^r}, s_n^{\hat{L}^r} + e_n^{L^r} - s_n^{L^r})$. And, the server in L^r takes vacations whose lengths are taken from the sequence $\{\check{V}_n^r,\ n = 0, 1, \ldots\}$ in the interval $[e_n^{L^r}, s_{n+1}^{L^r})$.

We can think of L^r as being constructed from the random sequence $\{A_n^r, {B'}_n^r, {V'}_n^r,\ n = 0, 1, \ldots\}$ with the usual interpretation. For each $n = 0, 1, \ldots,$ ${B'}_n^r = B_{k_n}^r$ where $k_n \geq n$ is the index of the service in $\hat{L}^r$ which is chosen to be the nth service in L^r. In particular, if the nth customer is an initiating customer in L^r, then k_n is the index of the service in $\hat{L}^r$ which starts at time $s_n^{\hat{L}^r}$ (e.g., $k_0 = 0$). The manner in which k_n, $n = 0, 1, \ldots,$ are determined guarantees that $\{{B'}_n^r,\ n = 0, 1, \ldots\}$ is i.i.d. with the same common distribution as that of $\{B_n^r,\ n = 0, 1, \ldots\}$, and so $\{{B'}_{n-1}^r,\ n, r = 1, 2, \ldots\}$ satisfies Condition A. For each $n = 1, 2, \ldots,$ ${V'}_n^r$ can either be taken from the sequence $\{\check{V}_n^r,\ n = 0, 1, \ldots\}$ (if the vacation is between two busy periods) or from the sequence $\{V_n^r,\ n = 0, 1, \ldots\}$ (if the vacation is within a busy period). Again, the random sequence $\{{V'}_n^r, n = 0, 1, \ldots\}$ is probabilistically identical to $\{V_n^r,\ n = 0, 1, \ldots\}$ and furthermore $\{{V'}_{n-1}^r,\ n, r = 1, 2, \ldots\}$ satisfies

Condition A.

We have just shown that the system L^r obtained by the elaborate construction described above possesses the desired probabilistic characteristics. It remains to show that this construction indeed facilitates comparison to $\hat{L}^r$ which leads to the proof of Lemma 2.6. To this end, we define for each $r = 1, 2, \ldots,$

$$\theta^{L^r}(t) = \sum_{n=0}^{\infty}(s_n^{\hat{L}^r} - s_n^{L^r})1\{s_n^{L^r} \le t < s_{n+1}^r\}, \qquad t \ge 0. \tag{2.131}$$

We shall see the significance of this quantity later. For now, we note from its definition that θ^{L^r} satisfies

$$0 \le \theta^{L^r}(t) \le R^r_{\tilde{R}^r(t)}, \qquad t \ge 0, \tag{2.132}$$

so that the following lemma holds.

Lemma 2.7

$$\frac{1}{\sqrt{r}} \sup_{0 \le t \le 1} \theta^{L^r}(rt) \xrightarrow{P} 0. \tag{2.133}$$

Proof: The result follows directly from Lemma 2.1 since

$$\frac{1}{\sqrt{r}} \sup_{0 \le t \le 1} \theta^{L^r}(rt) \le \frac{1}{\sqrt{r}} \max_{0 \le i \le \tilde{R}^r(r)} R_i^r \tag{2.134}$$

and $\{R^r_{n-1},\ n, r = 1, 2, \ldots\}$ satisfies Condition A. □

Proof of Lemma 2.6 (Continued): For each $r = 1, 2, \ldots,$ define $\tilde{S}^r(t)$ to be the number of potential service completions in $\hat{L}^r$ in the interval $[0, t]$ and

$$\varsigma^r(t) = \frac{1}{\sqrt{r}}\left[\tilde{S}^r(rt) - \frac{m^r rt}{\bar{R}^r}\right], \qquad 0 \le t \le 1. \tag{2.135}$$

It is plain that

$$\tilde{S}^r(t) - m^r \tilde{R}^r(t) \le m^r, \qquad t \ge 0, \tag{2.136}$$

such that

$$m(\varsigma^r, m^r \varrho^r) \le \frac{1}{\sqrt{r}} m^r, \qquad r = 1, 2, \ldots. \tag{2.137}$$

Therefore, as $r \longrightarrow \infty$, we get

$$m(\varsigma^r, m^r \varrho^r) \longrightarrow 0, \tag{2.138}$$

and so ς^r and $m^r\varrho^r$ both converge to the same $D[0,1]$-valued random function, namely some multiple of Wiener process defined on $[0,1]$.

To prove the lemma, it suffices to show that as $r \to \infty$,

$$\frac{1}{\sqrt{r}} \sup_{0 \le t \le 1} |Q^{\hat{L}^r}(rt) - Q^{L^r}(rt)| \xrightarrow{P} 0. \tag{2.139}$$

Define for each $r = 1, 2, \ldots,$

$$t^{L^r}(t) = \max\{s_n^{L^r}, n = 0, 1, \ldots : s_n^{L^r} \le t\}, \qquad t \ge 0, \tag{2.140}$$

i.e., the starting epoch of the busy period to which t belongs (if t is in the middle of a busy period) or of the last busy period before time t (if t is in between two busy periods). Also define

$$\tau^{L^r}(t) = \max\{s_n^{L^r}, n = 0, 1, \ldots : s_{n-1}^{L^r} \le t\}, \qquad t \ge 0. \tag{2.141}$$

We shall show that for each $r = 1, 2, \ldots,$

$$Q^{\hat{L}^r}(t) - Q^{L^r}(t) \le \tilde{S}^r(t + \theta^{L^r}(t)) - \tilde{S}^r(t), \qquad t \ge 0 \tag{2.142}$$

and

$$\begin{aligned} Q^{L^r}(t) - Q^{\hat{L}^r}(t) \;&\le\; \tilde{S}^r(t^r(t) + \theta^{L^r}(t^r(t))) - \tilde{S}^r(t^r(t)) \\ &\qquad + \max\{Q^{L^r}(t^r(t)), Q^{L^r}(\tau^r(t))\}, \qquad t \ge 0, \end{aligned} \tag{2.143}$$

such that

$$\begin{aligned} \frac{1}{\sqrt{r}} \sup_{0 \le t \le 1} |Q^{\hat{L}^r}(rt) - Q^{L^r}(rt)| \;&\le\; \frac{1}{\sqrt{r}} \sup_{0 \le t \le 1} [\tilde{S}^r(rt + \theta^{L^r}(rt)) - \tilde{S}^r(rt)] \\ &\qquad + \frac{1}{\sqrt{r}} \sup_{0 \le t \le 1} Q^{\hat{L}^r}(\tau^r(rt)). \end{aligned} \tag{2.144}$$

We showed earlier that ς^r converges weakly to a random function which is continuous a.s. Using this fact, together with Lemma 2.7, we can use the random time change argument used in Section 2.3.4 to show that the first term in (2.144) convergers to 0 in probability. The second term also goes to 0 in probability because $Q^{L^r}(\tau^r(t))$ is just the number of customers that arrive during the vacation which ends at time $\tau^r(t)$. Hence, the proof of the lemma is complete if we can show that the bounds (2.142) and (2.143) indeed hold.

To show (2.142), we need consider only the case $Q^{\hat{L}^r}(t) > 0$ since it trivially holds if $Q^{\hat{L}^r}(t) = 0$. Defining

$$t_1 = \sup\{0 \leq \tau < t : Q^{\hat{L}^r}(\tau) = 0\}, \tag{2.145}$$

we have

$$Q^{\hat{L}^r}(t) = \tilde{A}^r(t) - \tilde{A}^r(t_1) - [\tilde{S}^r(t) - \tilde{S}^r(t_1)]. \tag{2.146}$$

Here, we use the fact that since

$$Q^{\hat{L}^r}(\tau) > 0, \qquad t_1 \leq \tau \leq t, \tag{2.147}$$

each service completion in the interval $[t_1, t]$ leads to a customer departure. On the other hand, the number of departures from the system L^r in the same interval is bounded above by $\tilde{S}^r(t + \theta^{L^r}(t)) - \tilde{S}^r(t)$, and so

$$\begin{aligned} Q^{L^r}(t) &\geq Q^{L^r}(t_1) + \tilde{A}^r(t) - \tilde{A}^r(t_1) - [\tilde{S}^r(t + \theta^{L^r}(t)) - \tilde{S}^r(t)] \\ &\geq \tilde{A}^r(t) - \tilde{A}^r(t_1) - [\tilde{S}^r(t + \theta^{L^r}(t)) - \tilde{S}^r(t)]. \end{aligned} \tag{2.148}$$

Combining (2.146) and (2.148), we obtain (2.142).

To show (2.143), we again need only consider the case $Q^{L^r}(t) > 0$. If t is in the middle of a busy period, then we have

$$D^{L^r}(t) - D^{L^r}(t^r(t)) = \tilde{S}^r(t + \theta^{L^r}(t)) - \tilde{S}^r(t^r(t) + \theta^{L^r}(t^r(t))), \tag{2.149}$$

so that

$$Q^{L^r}(t) = Q^{L^r}(t^r(t)) + \tilde{A}^r(t) - \tilde{A}^r(t^r(t)) - [\tilde{S}^r(t + \theta^{L^r}(t)) - \tilde{S}^r(t^r(t) + \theta^{L^r}(t^r(t)))]. \tag{2.150}$$

But, since

$$D^{\hat{L}^r}(t) - D^{\hat{L}^r}(t^r(t)) \geq \tilde{S}^r(t) - \tilde{S}^r(t^r(t)), \tag{2.151}$$

we have

$$\begin{aligned} Q^{\hat{L}^r}(t) &\geq Q^{\hat{L}^r}(t^r(t)) + \tilde{A}^r(t) - \tilde{A}^r(t^r(t)) - [\tilde{S}^r(t) - \tilde{S}^r(t^r(t))] \\ &\geq \tilde{A}^r(t) - \tilde{A}^r(t^r(t)) - [\tilde{S}^r(t + \theta^{L^r}(t)) - \tilde{S}^r(t^r(t))], \end{aligned} \tag{2.152}$$

where the last inequality follows from the fact that $Q^{\hat{L}^r}(t^r(t)) \geq 0$ and $\tilde{S}^r(t + \theta^{L^r}(t)) \geq \tilde{S}^r(t)$. Combining (2.150) and (2.152), we obtain

$$Q^{L^r}(t) - Q^{\hat{L}^r}(t) \leq \tilde{S}^r(t^r(t) + \theta^{L^r}(t^r(t))) - \tilde{S}^r(t^r(t)) + Q^{L^r}(t^r(t)). \tag{2.153}$$

On the other hand, if t is between two busy period, then we have

$$Q^{L^r}(t) \leq Q^{L^r}(\tau^r(t)). \tag{2.154}$$

Combining this last inequality with (2.153), we obtain the desired result, and so the proof of Lemma 2.6 is now complete. □

Proof of Theorem 2.7: Combining Lemmas 2.5 and 2.6, we have

$$m(\chi^{L^r}, \chi^{\bar{L}^r}) \xrightarrow{P} 0, \tag{2.155}$$

and so by the converging together theorem the desired result follows from Lemma 2.4. □

CHAPTER 3

LIGHT TRAFFIC ANALYSIS AND INTERPOLATION APPROXIMATIONS FOR VACATION MODELS

3.1 Introduction

We have seen in the previous chapter that using heavy traffic analysis, valuable information can be obtained about the behavior of vacation models as the traffic rate approaches its critical value. The crucial fact there is that this asymptotic information can be obtained even though the behavior of the systems in moderate traffic is not known. In this chapter, we use the so-called light traffic analysis to study the behavior of vacation models as the traffic rate approaches zero. Like heavy traffic analysis, light traffic analysis provides valuable qualitative feel and worthwhile insights for those service policies for which exact analysis is not feasible. One use of heavy and light traffic results is to provide a rigorous basis for interpolation approximations. We study such approximations at the end of this chapter. In particular, we apply these approximations to the limited policies and study their performance by means of numerical validations.

Light traffic analysis first appeared in the literature in 1965 with the work of Beneš [5]. In 1983, Burman and Smith [12] proposed the interpolation based on light and heavy traffic information to study multi-server queues with Poisson arrivals. Recently, Reiman and Simon [67,68] developed a general method to obtain light traffic results for systems with Poisson (or Poisson driven) arrivals. They were the first authors to identify that light traffic analysis basically can be formulated as a problem of finding the derivatives of the performance measures of interest with respect to the arrival rate λ at $\lambda = 0$. They showed that if a performance measure satisfies some conditions (they call such a performance measure *admissible*), then a particular interchange of limits is justified and the method yields the correct values for the derivatives of that measure.

In the next section, we describe the Reiman–Simon method in more details. Then in Section 3.3, we apply this method to vacation models with Poisson arrivals. We shall concentrate throughout the chapter on the steady state waiting time as the performance measure. We should mention at this point the decomposition results established by Gelenbe and Iasnogorodski [36] and Doshi [21] which decompose the steady state waiting time in a vacation models under the exhaustive service policy into two components: the steady state waiting time of the corresponding single server queue without vacations (i.e., $M/GI/1$ queue) and the forward recurrence time of the vacation periods. Since the second component is independent of the arrival rate λ, only the first component will determine the derivatives with respect to λ, and so we can in fact use the light traffic results of the $M/GI/1$ queue. However, we shall derive in Section 3.3 the light traffic results for the exhaustive policy using the Reiman–Simon method. This exercise serves to illustrate how the Reiman–Simon method can be used to study vacation models in light traffic, in particular those under service policies for which decomposition result is not available, such as the limited policies. The light traffic results for the exhaustive policies can be readily extended to the Bernoulli policies by virtue of the equivalence result discussed in the previous chapter. Later in Section 3.3, we study the light traffic behavior of the limited policies. We exploit some properties of the Reiman–Simon method to show that again the light traffic results for the exhaustive policy can be extended to these service policies. In Section 3.4, we study the applications of interpolation approximations to the limited policies.

3.2 The Reiman–Simon Method

Consider an open queueing system fed by a Poisson arrival stream with rate λ. Let F be a (steady–state) performance measure of the system; in our case F is the steady–state waiting time. We are interested in determining the derivatives of

$\bar{F}(\lambda) = E_\lambda[F]$ evaluated at $\lambda = 0$, i.e.,

$$\bar{F}^{(n)}(0) := \lim_{\lambda \downarrow 0} \bar{F}^{(n)}(\lambda), \qquad n = 0, 1, \ldots, \tag{3.1}$$

with the notation $f^{(0)} = f$ and $f^{(n)}(\lambda) = \frac{d^n}{d\lambda^n} f(\lambda)$, $n = 1, 2, \ldots$.

The idea behind the Reiman–Simon method is that F can be interpreted as being evaluated at time zero assuming that the system has been running since time $-\infty$, i.e., the system has reached steady state. For instance if F is the steady state waiting time, then it can be interpreted as the waiting time of a 'tagged' customer which arrives at time zero. Let F_T denote the performance measure given that only arrivals in the interval $[-T, T]$ are taken into account, and let $\bar{F}_T(\lambda) = E_\lambda[F_T]$. Then, if F satisfies certain conditions—in which case we say F is admissible—then the interchange of limits

$$\lim_{\lambda \downarrow 0} \lim_{T \to \infty} \bar{F}_T^{(n)}(\lambda) = \lim_{T \to \infty} \lim_{\lambda \downarrow 0} \bar{F}_T^{(n)}(\lambda) \tag{3.2}$$

is shown to hold for $n = 0, 1, \ldots$. Furthermore, the left–hand side can be shown to be just the quantity we want, namely $\lim_{\lambda \downarrow 0} \bar{F}^{(n)}(\lambda)$. The significance of this interchange of limits is that the right–hand side of the equation can be computed.

For convenience, we state below the admissibility condition and the Reiman–Simon method.

Definition 3.1 (Admissibility) *The performance measure F is admissible if there exist constants K, N, a and θ with $K, N < \infty$, $1 < a < \infty$ and $\theta > 0$ such that for any $0 < T < S$ and $j, k = 0, 1, \ldots$, we have*

$$E\Big[|F_T - F_S| \;\Big|\; A_{(0,T]}(j), A_{(T,S]}(k)\Big] \leq K(j+k)^N a^{j+k} e^{-\theta T}, \tag{3.3}$$

where $A_{(a,b]}(l)$ is the event that there are l arrivals in the intervals $[-b, -a)$ and $(a, b]$.

Lemma 3.1 (Reiman–Simon) *If F is admissible, then*

$$\bar{F}(0) = E[F| \textit{ no other arrivals }] \tag{3.4}$$

and

$$\bar{F}^{(n)}(0) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \Psi(\{t_1, \ldots, t_n\})\, dt_n \cdots dt_1, \qquad n = 1, 2, \ldots, \tag{3.5}$$

where

$$\Psi(\{t_1, \ldots, t_n\}) = \sum_{j=0}^{n} (-1)^{n-j} \sum_{\{i_1,\ldots,i_j\} \subseteq \{1,\ldots,n\}} E[F | \{t_{i_1}, \ldots, t_{i_j}\}] \tag{3.6}$$

and

$$E[F | \{t_{i_1}, \ldots, t_{i_j}\}] = E[F |\ \textit{arrivals at time}\ t_{i_1}, \ldots, t_{i_j}]. \tag{3.7}$$

3.3 The Reiman–Simon Method Applied to Vacation Models

In this section, we use the Reiman-Simon method to study vacation models under the exhaustive policy in light traffic. We use the steady state waiting time to illustrate the method. Later in the section, the Bernoulli and limited policies are studied using the results for the exhaustive policy. We first describe the probabilistic setting in which the computation of the derivatives is carried out.

3.3.1 The Probabilistic Setting

In the setting, we have a vacation model which has been running since time $-\infty$. The system is governed by three mutually independent random sequences: $\{T_n,\ n \in \mathbb{Z}\}$, $\{B_n,\ n \in \mathbb{Z}\}$ and $\{V_n,\ n \in \mathbb{Z}\}$. The sequence $\{T_n,\ n \in \mathbb{Z}\}$ is a Poisson process constituting the arrivals into the system; we assume that $T_0 = 0$, i.e., there is a customer arrival at time $t = 0$. The sequences $\{B_n,\ n \in \mathbb{Z}\}$ and $\{V_n,\ n \in \mathbb{Z}\}$ are each constituted of i.i.d. random variables with distribution functions B and V, respectively, i.e.,

$$B(t) = P[B_n \le t] \quad \text{and} \quad V(t) = P[V_n \le t], \qquad n = 0, 1, \ldots. \tag{3.8}$$

Let R_V be a random variable which is distributed as the forward recurrence time of the sequence $\{V_n,\ n \in \mathbb{Z}\}$, i.e.,

$$R_V(t) = P[R_V \le t] = \frac{\int_0^t [1 - V(\tau)]\, d\tau}{E[V]}, \qquad t \ge 0. \tag{3.9}$$

The random variable B_n is the service time of the customer which arrives at time T_n. The sequence $\{V_n,\ n \in \mathbb{Z}\}$ and the random variable R_V are interpreted as follows: Suppose there exists $n^* \leq 0$ such that $T_n = -\infty$ for all $n < n^*$ and $-\infty < T_{n^*}$. Then, the customer which arrives at time T_{n^*} arrives to an empty queue and waits for the server to return from vacation. The length of this remaining vacation is given by the random variable R_V. The lengths of the subsequent vacations after time T_{n^*} are taken from the sequence $\{V_n,\ n \in \mathbb{Z}\}$. We do not specify the construction of the sample path for the case where $T_n > -\infty$ for all $n \in \mathbb{Z}$ since we do not need it for the computation of the derivatives using the Reiman–Simon method. It should be noted, however, that it is possible to construct the sample paths for this case. The reader familiar with Palm probabilities is referred to a monograph by Baccelli and Brémaud [4] for more details on construction of stationary queues.

3.3.2 The Exhaustive Policy

Let W_E denote the steady state waiting time in a vacation model under the exhaustive policy. In order to apply Lemma 3.1 for computing the derivatives of $\bar{W}_E(\lambda) = E_\lambda[W_E]$ at $\lambda = 0$, we have to show that W_E is admissible. Indeed, using the arguments used in [68], it can be shown that if there exists $\theta^* > 0$ such that $E[e^{\theta B}] < \infty$ for all $\theta \leq \theta^*$ where B is the generic service time, then W_E is admissible.

We shall now use the Reiman–Simon method to compute $\bar{W}_E(0)$ and $\bar{W}'_E(0)$. We assume that the condition for the admissibility of the steady state waiting time is satisfied. Denoting the waiting time of the customer which arrives at time $t = 0$ by W_0, we have by Lemma 3.1 that

$$\bar{W}_E(0) = E[W_0|\text{ no other arrivals }] = \bar{R}_V \tag{3.10}$$

and

$$\bar{W}'_E(0) = \int_{-\infty}^{\infty} \Big(E[W_0|\text{ exactly one arrival at time } t\ ] - \bar{W}_E(0) \Big)\, dt. \tag{3.11}$$

Notice that because of FCFS assumption, we have

$$E[W_0|\text{ exactly one arrival at time } t\,] = E[W_0|\text{ no other arrivals }], \qquad t > 0, \tag{3.12}$$

such that

$$\begin{aligned} \bar{W}_E'(0) &= \int_{-\infty}^{0} \Big(E[W_0|\text{ exactly one arrival at time } t\,] - \bar{W}_E(0)\Big)\,dt \\ &= \int_{-\infty}^{0} \Big(E[W_0|\text{ exactly one arrival at time } t\,] - \bar{R}_V\Big)\,dt. \end{aligned} \tag{3.13}$$

Consider the renewal process $\{V_n, n = 0, 1, \ldots\}$. Define

$$S_n = \sum_{i=0}^{n} V_i, \qquad n = 0, 1, \ldots, \tag{3.14}$$

and let $\{\tilde{V}(t),\ t \geq 0\}$ be the induced counting process, i.e.,

$$\tilde{V}(t) = \begin{cases} \max\{n \geq 1 : S_{n-1} \leq t\} & \text{if } V_0 \leq t \\ 0 & \text{otherwise,} \end{cases} \qquad t \geq 0. \tag{3.15}$$

The *residual life* at time t is defined by

$$\gamma_t = S_{\tilde{V}(t)} - t, \qquad t \geq 0. \tag{3.16}$$

Let $\{X, V_{n-1},\ n = 1, 2, \ldots\}$ be a *delayed* renewal process, i.e., the first renewal is given by the random variable X whose distribution might be different from that common to the rest of the random variables. Denote the residual life at time t of this delayed renewal process by γ_t^X.

Letting

$$g(r,t) := E[W_0|\text{ one arrival at time } t, R_V = r] \tag{3.17}$$

and noting that

$$\bar{R}_V = E[\gamma_t^{R_V}] = \int_0^\infty E[\gamma_t^r]\,dR_V(r), \qquad t \geq 0, \tag{3.18}$$

we then have

$$\bar{W}_E'(0) = \int_0^\infty \int_0^\infty \Big(g(r,-t) - E[\gamma_t^r]\Big)\,dR_V(r)\,dt. \tag{3.19}$$

But, it can be easily shown that

$$g(r,-t) = \begin{cases} E[\gamma^B_{t-r}] & \text{if } r < t \\ r - t + \bar{B} & \text{if } t \leq r, \end{cases} \qquad t \geq 0 \tag{3.20}$$

and

$$E[\gamma^r_t] = \begin{cases} E[\gamma_{t-r}] & \text{if } r < t \\ r - t & \text{if } t \leq r, \end{cases} \qquad t \geq 0, \tag{3.21}$$

such that

$$\begin{aligned} \bar{W}'_E(0) &= \int_0^\infty \int_0^t \left(E[\gamma^B_{t-r}] - E[\gamma_{t-r}]\right) dR_V(r)\, dt \\ &\qquad + \int_0^\infty \int_t^\infty \bar{B}\, dR_V(r)\, dt. \end{aligned} \tag{3.22}$$

Denoting the first and second terms on the right–hand side by T_1 and T_2, respectively, we then have

$$T_2 = \bar{B}\bar{R}_V \tag{3.23}$$

and

$$\begin{aligned} T_1 &= \int_0^\infty dR_V(r) \left[\int_r^\infty \left(E[\gamma^B_{t-r}] - E[\gamma_{t-r}]\right) dt\right]^{t-r=u} \\ &= \int_0^\infty \left(E[\gamma^B_u] - E[\gamma_u]\right) du \end{aligned} \tag{3.24}$$

Now, with

$$A_x(t) = P[\gamma_t > x] \qquad \text{and} \qquad A^B_x(t) = P[\gamma^B_t > x], \qquad x \geq 0, \tag{3.25}$$

we can write (3.24) as

$$T_1 = \int_0^\infty \int_0^\infty \left[A^B_x(t) - A_x(t)\right] dx\, dt. \tag{3.26}$$

By the usual renewal argument [44], we have

$$A_x(t) = 1 - V(t+x) + \int_0^t A_x(t-y)\, dV(y) \tag{3.27}$$

and

$$A^B_x(t) = 1 - B(t+x) + \int_0^t A_x(t-y)\, dB(y). \tag{3.28}$$

For simplicity, let us assume that $A_x(t)$ can be written as

$$A_x(t) = \int_0^t a_x(\tau)\,d\tau + A_x(0), \qquad t \geq 0. \tag{3.29}$$

So, we can write (3.27) and (3.28) as

$$A_x(t) = 1 - V(t+x) + \int_0^t V(t-\tau)a_x(\tau)\,d\tau + A_x(0)V(t) \tag{3.30}$$

and

$$A_x^B(t) = 1 - B(t+x) + \int_0^t B(t-\tau)a_x(\tau)\,d\tau + A_x(0)B(t), \tag{3.31}$$

respectively. In that case, we obtain

$$\begin{aligned} T_1 &= \int_0^\infty \int_0^\infty \Big(1 - B(t+x)\Big)\,dx\,dt - \int_0^\infty \int_0^\infty \Big(1 - V(t+x)\Big)\,dx\,dt \\ &\quad + \int_0^\infty \int_0^\infty \int_0^t \Big(B(t-\tau) - V(t-\tau)\Big)a_x(\tau)\,d\tau\,dx\,dt \\ &\quad + \int_0^\infty \int_0^\infty A_x(0)\Big(B(t) - V(t)\Big)\,dx\,dt. \end{aligned} \tag{3.32}$$

The first and second terms on the right–hand side are easily computed to yield $\overline{B^2}/2$ and $\overline{V^2}/2$, respectively, while the last term can be shown to be $(\bar{V} - \bar{B})\bar{V}$. The third term, denoted by T_{13}, is given by

$$\begin{aligned} T_{13} &= \int_0^\infty \int_0^\infty \int_0^t \Big(B(t-\tau) - V(t-\tau)\Big)a_x(\tau)\,d\tau\,dt\,dx \\ &= \int_0^\infty \int_0^\infty \int_\tau^\infty \Big(B(t-\tau) - V(t-\tau)\Big)\,dt\,a_x(\tau)\,d\tau\,dx \\ &= \int_0^\infty \int_0^\infty (\bar{V} - \bar{B})a_x(\tau)\,d\tau\,dx \\ &= (\bar{V} - \bar{B})\int_0^\infty \Big(A_x(\infty) - A_x(0)\Big)\,dx \\ &= (\bar{V} - \bar{B})(\bar{R}_V - \bar{V}). \end{aligned} \tag{3.33}$$

Now we can substitute all the terms into (3.32), and then (3.23) and (3.32) into (3.22). Noting that $\bar{R}_V = \overline{V^2}/(2\bar{V})$, we finally obtain

$$\bar{W}_E'(0) = \frac{\overline{B^2}}{2}. \tag{3.34}$$

This result is in agreement with that obtained using the decomposition result for vacation models and the Pollaczek–Khintchine formula.

3.3.3 The Bernoulli Policies

As mentioned in the previous chapter, there exists for any vacation model under a Bernoulli service policy with parameter p, $0 \leq p \leq 1$, an equivalent vacation model under the exhaustive policy. The equivalence is in terms of waiting times and the equivalent vacation model differs from the original one only in the distribution of the service times. The service times in the equivalent model are distributed as a service time in the original model with probability p and as a service time plus a vacation duration with probability $1-p$. In the previous chapter, we use the transient version of this result; here we use the steady–state version of the result which was established by Keilson and Servi in [45].

Let $W_{B(p)}$ denote the steady state waiting time for a vacation model under the Bernoulli policy with parameter p. So, if $W_{B(p)}$ is an admissible performance measure, then we see from (3.34) that

$$\bar{W}'_{B(p)}(0) = \frac{1}{2}(\overline{B^2} + (1-p)\overline{V^2} + 2(1-p)\bar{B}\bar{V}). \tag{3.35}$$

The admissibility condition in this case can be shown to be satisfied if both the service time and vacation length distributions have exponential tail, i.e., there exists $\theta_1 > 0$ such that $E[e^{\theta B}] < \infty$ and $E[e^{\theta V}] < \infty$ for all $\theta \leq \theta_1$. Higher order derivatives $\bar{W}^{(n)}_{B(p)}(0)$ can be readily obtained using the Pollaczek–Khintchine formula.

3.3.4 The Limited Policies

Unlike a vacation model under the Bernoulli policy, there is no known equivalent exhaustive system for a vacation model under the limited policy. However, we shall see that the very definition of the limited policies and the fact that the computation of the nth derivative using the Reiman–Simon method involves $(n+1)$ or fewer customers enable us to use the light traffic results for the exhaustive policy.

We first discuss the admissibility of the steady–state waiting time for a vacation model under a limited policy. We know that waiting time is admissible for the Bernoulli policy with any parameter p, $0 \leq p \leq 1$, as long as both the service time and vacation length distributions have exponential tail. Following the argument in Reiman and Simon [68], to show the admissibility of waiting time for a limited policy, it suffices to show a Bernoulli policy which bounds the limited policy in terms of waiting time. But, it will be shown in the next chapter that indeed the Bernoulli policy with parameter $p = 0$ (which actually is identical to the limited policy with parameter $m = 1$) bounds all limited policies (in fact, it bounds all policies) in terms of waiting time. So, for the limited policies, the admissability condition is again satisfied if both the service time and vacation duration distribution have exponential tail.

We now see how we can obtain light traffic results for the limited policy with parameter m, where $m \geq 2$. Let $W_{L(m)}$ denote the steady–state waiting time in a vacation model under the limited policy with limit m. We see from (3.5) that the computation of the nth derivative at $\lambda = 0$ involves at most $(n + 1)$ customers (n arrivals plus the customer at time $t = 0$). Thus as long as $n + 1 \leq m$, there can never be more than m customers in the system for all the sample paths considered in the computation of the nth derivative. For these sample paths, we never use the fact that the service policy used is the limited policy with limit m, and so we might as well think that the exhaustive service policy is used. This implies that

$$\bar{W}_{L(m)}^{(n)}(0) = \bar{W}_{E}^{(n)}(0), \qquad n = 0, 1, \ldots, m - 1. \tag{3.36}$$

The derivatives for the exhaustive policy can be obtained using either the Pollaczek–Khintchine formula or the Reiman–Simon method as discussed in the Section 3.3.2.

3.4 Interpolation Approximations for Vacation Models

Light traffic analysis provides us with the derivatives of a performance measure $\bar{F}$ with respect to the arrival rate λ at $\lambda = 0$. These results naturally lead to a (truncated) Taylor series expansion at $\lambda = 0$ as an approximation for $\bar{F}(\lambda)$. However, most performance measures have a *critical rate* $\lambda_c < \infty$ which as $\lambda \longrightarrow \lambda_c$ the value of the performance measure goes unbounded. So, a truncated Taylor expansion which is a polynomial in λ is not a very suitable approximation for these performance measures. Fortunately, if we weight (or "normalize") the performance measure by some function $c(\lambda)$ which goes to 0 as $\lambda \to \lambda_c$, there is a chance that $c(\lambda)\bar{F}(\lambda)$ might be bounded on the interval $[0, \lambda_c)$ and so it can be approximated by a polynomial. Indeed, heavy traffic analysis shows that we can take $c(\lambda)$ to be $(\lambda_c - \lambda)$ for most performance measures. Moreover, heavy traffic analysis also provides us with the limit

$$\lim_{\lambda \uparrow \lambda_c} (\lambda_c - \lambda)\bar{F}(\lambda) \tag{3.37}$$

which can be combined with the light traffic results to form an interpolation approximations.

Interpolation approximations based on light and heavy traffic results were first proposed by Burman and Smith [13] to study single–server systems with bursty arrivals. This approach was later extended by Reiman and Simon [67] to a larger class of systems with Poisson or Poisson driven arrivals. To describe the approximation more precisely, let

$$G(\lambda) := (\lambda_c - \lambda)\bar{F}(\lambda), \qquad 0 \le \lambda < \lambda_c \tag{3.38}$$

and let $g(\lambda)$ denote the approximation to $G(\lambda)$. Assuming we have up to the $(n-1)$st derivative of $\bar{F}(\lambda)$ at $\lambda = 0$ and the heavy traffic limit (3.37), we construct an nth degree polynomial $g(\lambda)$ which satisfies the conditions

$$\begin{aligned} g^{(i)}(0) &= G^{(i)}(0) \\ &= \lambda_c \bar{F}^{(i)}(0) - i\bar{F}^{(i-1)}(0), \qquad i = 0, 1, \ldots, n-1 \end{aligned} \tag{3.39}$$

and

$$g(\lambda_c) = G(\lambda_c). \tag{3.40}$$

Let $f(\lambda)$ be the approximation to $\bar{F}(\lambda)$. To obtain $f(\lambda)$, we 'unnormalize' $g(\lambda)$, i.e., we let

$$f(\lambda) = \frac{g(\lambda)}{\lambda_c - \lambda}, \qquad 0 \le \lambda < \lambda_c. \tag{3.41}$$

In the next subsection, we obtain heavy traffic limits of the form (3.37) using the heavy traffic results established in Chapter 2. We then obtain the interpolation approximations for vacation models and compare their performance with exact (simulation) results. Numerical examples indicate that the agreement is extremely good; in fact, the interpolation can be easily shown to be exact for the exhaustive and Bernoulli policies.

3.4.1 Heavy Traffic Limits

In Chapter 2, we showed that the transient version of various processes converges to reflected Brownian motion with negative drift. In some cases, it is possible to obtain from these results the heavy traffic behavior of the steady–state distribution of the corresponding processes. As an example, consider the actual waiting time process in the $GI/GI/1$ queue. As in Chapter 2, we consider a sequence $\{G^r,\ r = 1, 2, \ldots\}$ of $GI/GI/1/$ systems. For each $r = 1, 2, \ldots$, let $\{W_n^r,\ n = 0, 1, \ldots\}$ be the waiting time process of the system G^r and denote its limiting random variable by W_∞^r. We showed in Chapter 2 that the sequence $\{\omega^r,\ r = 1, 2, \ldots\}$ of $D[0,1]$–valued random processes defined by (2.1) converges weakly to a reflected Brownian motion $f(\sigma_X \mathcal{W} + d)$. Whitt [90,92] extended this result to the space $D_1[0,\infty)$ which is the set of right continuous functions $x\colon [0,\infty) \to I\!R$ which have left limits with $x(t) \to -\infty$ as $t \to \infty$, endowed with the so–called Whitt's metric. Using a special case of the argument used by Harrison in [37], we shall see below that the weak convergence

$$\frac{1}{\sqrt{r}} W_\infty^r \xrightarrow{\mathcal{D}} E \quad \text{as} \quad r \to \infty, \tag{3.42}$$

where E is exponentially distributed, and indeed be obtained from the weak convergence of $\{\omega^r,\ r = 1, 2, \ldots\}$.

Let $=_{\text{st}}$ denote equivalence in probability distribution. A familiar result by Lindley [59] states that $(r = 1, 2, \ldots)$

$$W_n^r =_{\text{st}} \max_{0 \le k \le n} Z_k^r, \qquad n = 0, 1, \ldots, \tag{3.43}$$

where $\{Z_k^r,\ k = 0, 1, \ldots\}$ are defined in (2.4). This implies that

$$\omega^r(t) =_{\text{st}} \sup_{0 \le s \le t} [\zeta^r(s) + d^r(s)], \qquad t \ge 0,\ r = 1, 2, \ldots, \tag{3.44}$$

where ζ^r and d^r are defined in (2.6) and (2.7), respectively. In the steady state, we thus have

$$\frac{1}{\sqrt{r}} W_\infty^r =_{\text{st}} \sup_{s \ge 0} [\zeta^r(s) + d^r(s)]. \tag{3.45}$$

Whitt [90] showed that the functional $\pi\colon D_1[0,\infty) \to I\!R$ defined by

$$\pi(x) = \sup_{t \ge 0} x(t), \qquad x \in D_1[0,\infty) \tag{3.46}$$

is continuous, and so by the continuous mapping theorem and the fact that $\zeta^r \xrightarrow{\mathcal{D}} \sigma_X \mathcal{W}$ and $d^r(t) \to d(t) := Ct$, $t \ge 0$, we have

$$\frac{1}{\sqrt{r}} W_\infty^r \xrightarrow{\mathcal{D}} \pi(\sigma_X \mathcal{W} + d). \tag{3.47}$$

By [44, Corollary 5.1, p. 361], the right–hand side is a random variable which is exponentially distributed with rate $2|C|/\sigma_X^2$.

To obtain the heavy traffic limit (3.37) for the mean waiting time, we fix the mean service time for all the $GI/GI/1$ queues, i.e., $\bar{B}^r = \bar{B}$ for all $r = 1, 2, \ldots$ and we let $\lambda_c = 1/\bar{B}$. For each $r = 1, 2, \ldots$, we let $\lambda^r = 1/\bar{A}_r$. We note from (C2.1) that

$$(\lambda_c - \lambda)\sqrt{r} \to -C\lambda_c^2. \tag{3.48}$$

Taking the expectation of the left–hand side of (3.47), multiplying by $(\lambda_c - \lambda)\sqrt{r}$ and then letting $r \to \infty$, we obtain

$$\lim_{r \to \infty} (\lambda_c - \lambda^r) \bar{W}_\infty^r = \frac{\lambda_c^2 \sigma_X^2}{2} = \frac{\lambda_c^2(\sigma_A^2 + \sigma_B^2)}{2}. \tag{3.49}$$

For vacation models, the waiting times do not have a nice representation such as (3.43), and so the argument used above for the $GI/GI/1$ queues do not extend to vacation models. Notice that what we just showed above for the $GI/GI/1$ queues is some form of limit interchange. Indeed, it can be shown [39, p. 14] using a time reversal argument similar to that used to show (3.43) that

$$\sup_{0 \le s \le t} [\sigma_X \mathcal{W}(s) + d(s)] =_{\text{st}} f(\sigma_X \mathcal{W} + d)(t). \tag{3.50}$$

As a result, (3.47) can be interpreted as (with an abuse of notation)

$$\lim_{r \to \infty} \lim_{t \to \infty} \omega^r(t) =_{\text{st}} \lim_{t \to \infty} \lim_{r \to \infty} \omega^r(t). \tag{3.51}$$

We shall conjecture that such interchange of limits also holds for vacation models under the exhaustive, Bernoulli and limited policies.

In order to obtain heavy traffic limits in the form (3.37), we consider the sequences $\{E^r,\ r=1,2,\ldots\}$, $\{B^r,\ r=1,2,\ldots\}$ and $\{L^r,\ r=1,2,\ldots\}$ of vacation models under the exhaustive, Bernoulli and limited service policies, respectively. We shall assume in each of these sequences that $V_n^r =_{\text{st}} V$ and $B_n^r =_{\text{st}} B$ for all $n=0,1,\ldots$ and $r=1,2,\ldots$. In addition, we set $p^r=p$, $r=1,2,\ldots$, for the sequence $\{B^r,\ r=1,2,\ldots\}$ and $m^r=m$, $r=1,2,\ldots$, for the sequence $\{L^r,\ r=1,2,\ldots\}$. In other words, in each sequence we let only the arrival process vary with r. The critical arrival rates are then given by

$$\lambda_c^E = \frac{1}{\bar{B}}; \tag{3.52}$$

$$\lambda_c^B = \frac{1}{\bar{B}+(1-p)\bar{V}}; \tag{3.53}$$

$$\lambda_c^L = \frac{1}{\bar{B}+\bar{V}/m}. \tag{3.54}$$

Let us write $\bar{W}_E(\lambda^r)$ for $\bar{W}_\infty^{E^r}$, $\bar{W}_{B(p)}(\lambda^r)$ for $\bar{W}_\infty^{B^r}$, $\bar{Q}_{L(m)}(\lambda^r)$ for $\bar{Q}_\infty^{Q^r}$ and $\bar{W}_{L(m)}(\lambda^r)$ for $\bar{W}_\infty^{Q^r}$. From the heavy traffic results established in the previous chapter and from the various convergence conditions stated in (C2.1)–(C2.6), we then obtain

$$\lim_{r\to\infty}(\lambda_c^E-\lambda^r)\bar{W}_E(\lambda^r) = \frac{(\lambda_c^E)^2(\sigma_A^2+\sigma_B^2)}{2}, \tag{3.55}$$

$$\lim_{r\to\infty}(\lambda_c^B-\lambda^r)\bar{W}_{B(p)}(\lambda^r) = \frac{(\lambda_c^B)^2(\sigma_A^2+\sigma_B^2+(1-p)\sigma_V^2+p(1-p)\bar{V}^2)}{2} \tag{3.56}$$

and

$$\lim_{r\to\infty}(\lambda_c^L-\lambda^r)\bar{Q}_{L(m)}(\lambda^r) = \frac{(\lambda_c^L)^3(\sigma_A^2+\sigma_B^2+\sigma_V^2/m)}{2}. \tag{3.57}$$

By Little's result, the last limit is equivalent to

$$\lim_{r\to\infty}(\lambda_c^L-\lambda^r)\bar{W}_{L(m)}(\lambda^r) = \frac{(\lambda_c^L)^2(\sigma_A^2+\sigma_B^2+\sigma_V^2/m)}{2}. \tag{3.58}$$

3.4.2 Interpolation Approximations for Limited Policies

In this subsection, we apply the interpolation approximations based on heavy and light traffic information to the limited service policies. In particular, we shall obtain some closed–form approximate formulae for the mean waiting time which incorporate the heavy and traffic results for the limited policies established in the previous sections. The accuracy of these formulae are then studied by comparing them against simulation results.

Before we do this, we first note that for the exhaustive and Bernoulli policies, interpolating between the heavy and light traffic limits only (without the derivatives) yields

$$\bar{W}_E(\lambda) = \frac{\lambda_c^E \overline{B^2} \lambda}{2(\lambda_c^E - \lambda)} + \bar{R}_V \tag{3.59}$$

and

$$\bar{W}_B(\lambda) = \frac{\lambda_c^B (\overline{B^2} + (1-p)\overline{V^2} + 2(1-p)\bar{B}\bar{V})\lambda}{2(\lambda_c^B - \lambda)} + \bar{R}_V, \tag{3.60}$$

respectively. We see from the decomposition result for vacation models [21] and the Pollaczek–Khintchine formula that these formulae are indeed the exact mean waiting time for the respective policies. For the limited policies we do not obtain exact formulae, but we can expect that the approximate formulae to be very close to the exact ones. Indeed, numerical examples indicate that this is the case.

In the previous section, we have established that for the limited service policy with limit m, the ith derivative of the mean waiting time with respect to λ at $\lambda = 0$, where i ranges from 0 to $m-1$, is given by that of the exhaustive service policy. This derivative then can be computed either using the Reiman–Simon method as discussed in the previous section or simply by taking the derivative of the Pollaczek–Khintchine formula for the corresponding $M/GI/1$ system. So, together with the heavy traffic limit (3.58), we have $m+1$ pieces of information which can be used to construct an mth degree polynomial interpolation. As can be seen from some numerical examples given below, the accuracy of the second degree polynomial interpolation in fact is already quite acceptable for most practical purposes.

A second order interpolation approximation

From (3.10), (3.34), (3.36) and (3.58), we obtain a second order interpolation approximation for the limited policy using the method described at the outset of this section. Denoting the approximate mean waiting time by $\bar{w}_L$, we obtain

$$\bar{w}_L(\lambda) = \frac{\left((\frac{1}{\lambda_c^L})^2 - \bar{B}^2 + \frac{1}{m}\sigma_V^2\right)\lambda^2 + \lambda_c^L \overline{B^2}\lambda}{2(\lambda_c^L - \lambda)} + \bar{R}_V. \tag{3.61}$$

Numerical Examples

We now compare (3.61) against simulation results. We consider three types of distribution for the service times and vacation lengths, namely exponential, Erlangian and hyperexponential. For each type of distribution we consider two values of m—2 and 5—so that we have a total of six systems. We use $\bar{B} = 1.0$ and $\bar{V} = 0.5$ for all the systems, and for each system we vary λ/λ_c from 0.1 to 0.9. The variances of the service times and vacation lengths are respectively 0.5 and 0.125 for the Erlangian case, and 1.04 and 0.29 for the hyperexponential case.

The simulated results and the approximate values computed using (3.61) are shown in Tables 3.1, 3.2 and 3.3 below. The simulated results and the approximate values are listed under the headings "$\bar{W}_L(\lambda)$" and "$\bar{w}_L(\lambda)$", respectively. The "% Error" is computed as

$$\% \text{ Error} = \frac{\bar{w}_L(\lambda) - \bar{W}_L(\lambda)}{\bar{W}_L(\lambda)} \times 100. \tag{3.62}$$

For the computation of the confidence interval of all the simulated points, we employ the "batch means" method discussed in Law and Kelton [53, Sec. 8.6.1], and we use 90% confidence level.

	$m = 2$			$m = 5$		
λ/λ_c	$\bar{W}_L(\lambda)$	$\bar{w}_L(\lambda)$	% Error	$\bar{W}_L(\lambda)$	$\bar{w}_L(\lambda)$	% Error
0.1	0.544 ± 0.013	0.592	+8.8	0.555 ± 0.010	0.602	+8.5
0.2	0.700 ± 0.008	0.714	+2.0	0.703 ± 0.006	0.733	+4.3
0.3	0.885 ± 0.010	0.878	−0.8	0.864 ± 0.008	0.905	+4.7
0.4	1.143 ± 0.017	1.107	−3.1	1.097 ± 0.012	1.138	+3.7
0.5	1.482 ± 0.019	1.438	−3.0	1.429 ± 0.027	1.468	+2.7
0.6	2.016 ± 0.028	1.948	−3.4	1.916 ± 0.033	1.970	+2.8
0.7	2.937 ± 0.080	2.816	−4.1	2.793 ± 0.085	2.814	+0.8
0.8	4.627 ± 0.175	4.580	−1.0	4.489 ± 0.160	4.515	−0.6
0.9	10.038 ± 0.368	9.928	−1.1	9.622 ± 0.210	9.639	+0.2

Table 3.1: Exponential service and vacation times

	$m = 2$			$m = 5$		
λ/λ_c	$\bar{W}_L(\lambda)$	$\bar{w}_L(\lambda)$	% Error	$\bar{W}_L(\lambda)$	$\bar{w}_L(\lambda)$	% Error
0.1	0.404 ± 0.013	0.444	+9.9	0.415 ± 0.011	0.452	+8.9
0.2	0.520 ± 0.008	0.538	+3.5	0.521 ± 0.006	0.551	+5.8
0.3	0.665 ± 0.006	0.664	−0.2	0.647 ± 0.007	0.681	+5.3
0.4	0.864 ± 0.012	0.842	−2.5	0.818 ± 0.007	0.858	+4.9
0.5	1.142 ± 0.015	1.100	−3.7	1.068 ± 0.012	1.110	+3.9
0.6	1.556 ± 0.029	1.500	−3.6	1.431 ± 0.023	1.494	+4.4
0.7	2.264 ± 0.039	2.183	−3.6	2.070 ± 0.043	2.140	+3.4
0.8	3.662 ± 0.112	3.575	−2.4	3.324 ± 0.137	3.444	+3.6
0.9	7.960 ± 0.286	7.800	−2.0	7.387 ± 0.202	7.377	−0.1

Table 3.2: Erlangian service and vacation times

	$m = 2$			$m = 5$		
λ/λ_c	$\bar{W}_L(\lambda)$	$\bar{w}_L(\lambda)$	% Error	$\bar{W}_L(\lambda)$	$\bar{w}_L(\lambda)$	% Error
0.1	0.629 ± 0.016	0.634	+0.8	0.642 ± 0.011	0.644	+0.3
0.2	0.789 ± 0.008	0.758	−3.9	0.793 ± 0.008	0.778	−1.9
0.3	0.986 ± 0.008	0.926	−6.1	0.969 ± 0.013	0.953	−1.7
0.4	1.253 ± 0.011	1.159	−7.5	1.214 ± 0.016	1.191	−1.9
0.5	1.610 ± 0.016	1.498	−7.0	1.554 ± 0.016	1.528	−1.7
0.6	2.186 ± 0.042	2.019	−7.6	2.065 ± 0.053	2.041	−1.7
0.7	3.062 ± 0.085	2.906	−5.0	2.972 ± 0.079	2.903	−2.3
0.8	4.897 ± 0.182	4.710	−3.8	4.773 ± 0.196	4.639	−2.8
0.9	10.658 ± 0.377	10.176	−4.5	10.162 ± 0.500	9.872	−2.8

Table 3.3: Hyper–exponential service and vacation times

CHAPTER 4

STOCHASTIC COMPARISON AND MONOTONICITY RESULTS IN VACATION MODELS

4.1 Introduction

In vacation models considered in this thesis, the server can start a vacation either at a service completion or at the end of a vacation, and only at these epochs. A vacation is always taken if the queue is empty at either a service or vacation completion. If the queue is not empty at a vacation completion, the server has no choice but to resume its duty. If the queue is not empty at a service completion, however, the server has the choice of either serving the next available customer or starting a vacation. Whether or not the server takes a vacation in this case is determined by the *service policy* used.

The purpose of this chapter is to identify some conditions under which two service policies can be compared. Two service policies are compared by comparing some quantities of interest in a vacation model under one service policy to the corresponding quantities under the other service policy. In this chapter, we use an ordering called *stochastic ordering* on random processes. This ordering is very strong; indeed, two processes are said to be stochastically ordered if we can construct, on a common probability space, two processes which are probabilistically identical to the original processes and such that each sample path of one process lies below that of the other. We show that two service policies can be compared in this manner under some fairly general conditions. The setting in which the comparisons are made is also very general, as we make little assumptions on the probabilistic structure of the interarrival, service and vacation processes of the vacation models considered. Furthermore, for some quantities of interest, the comparisons can be made independently of the order in which the customers are served. Comparisons between service policies are useful for obtaining stochastic monotonicity results, bounds, and approximations for policies which are too difficult to analyze exactly. Work along this line has just been started for polling systems by Levy et al. [57].

This chapter is organized as follows. In Section 2, we present a precise description of vacation models while at the same time introducing the notation used throughout the chapter. In section 3, some basic relations and facts about vacation models are discussed. There, we also introduce the notion of stochastic ordering on random processes. A general framework for stochastic comparisons between two service policies is developed in Section 4. We then show in Section 5 that various service policies in the literature can be stochastically compared in this framework. In Section 6, we establish stochastic monotonicity results for various classes of parametrizable service policies.

4.2 The Model and Notation

A vacation model is governed by the sequence of random variables $\{A_n, B_n, V_n, U_n, n = 0, 1, \ldots\}$ with the following interpretation ($n = 0, 1, \ldots$):

A_{n+1} = the time between the nth and the $(n+1)$st arriving customer (with the convention that A_0 is the arrival time of the 0th customer);

B_n = the length of the nth service;

V_n = the length of the nth vacation period;

U_n = the server's decision at the end of the nth service, with $U_n = 1$ (resp. $U_n = 0$) if the server decides to serve the next customer (resp. to take a vacation).

The random variables $\{U_n,\ n = 0, 1, \ldots\}$ constitute a service policy which we shall denote by U. From the random variables above, we define the following quantities ($n = 0, 1, \ldots$):

T_n = the arrival time of the nth arriving customer ($= \sum_{j=0}^{n} A_j$);

D_n = the departure (i.e., service completion) time of the nth departing customer;

R_n = the index of the last vacation completed before time D_n;

$Q_n =$ the number of customers left behind by the nth departing customer;

$S_n =$ the number of customers (including the nth departing customer) that have been served up to time D_n since the end of the last (i.e., the R_nth) vacation;

$C_n =$ the number of customers in the queue at the end of the R_nth vacation;

$W_n =$ the waiting time (i.e., the time from the arrival epoch to the start of service) of the nth arriving customer;

$N(t) =$ the number of customers in the system at time $t \geq 0$.

Notice that we make the distinction between the nth arriving and the nth departing customers as we do *not* limit ourselves to the first–come–first–serve (FCFS) discipline. We make the following assumptions (A1)-(A3), where

(A1) The 0th customer arrives at time $t = 0$ (i.e., $A_0 = 0$) to an empty system and immediately receives service.

(A2) Once a customer enters the system, it does not leave until its service is completed;

(A3) Once a service is started, it is carried out to completion, i.e., there is no service preemption;

Using the notation introduced above, a vacation model can be described as follows: At time D_n the server completes a service of length B_n. If $U_n = 1$, a new service of length B_{n+1} begins. However, if $U_n = 0$, the server starts a vacation of length V_{R_n+1}. At the end of this vacation (i.e., at time $D_n + V_{R_n+1}$), the server starts a service of length B_{n+1} if the queue is not empty. Otherwise, it takes additional vacations until at least one customer is present when it returns from a vacation.

In this chapter, we consider only *simple* service policies which we define below.

Definition 4.1 *A service policy U is said to be simple if the conditions*

$$P[U_n = 1 | Z, U_j, \ j = 0, \ldots, n-1] = P[U_n = 1 | Q_n, S_n, C_n], \qquad n = 0, 1, \ldots \quad (4.1)$$

are satisfied, where Z denotes the random variables $\{A_n, B_n, V_n, \ n = 0, 1, \ldots\}$.

Let $\mathcal{S}$ be the set of 3–tuples of integers (q, s, c) where $q = 0, 1, \ldots$; $s, c = 1, 2, \ldots$; and $c - s \leq q$. With any simple service policy U, we can associate a sequence

$\{f_n,\ n = 0, 1, \ldots\}$ of mappings $f_n : \mathcal{S} \to [0,1]$, $n = 0, 1, \ldots$, defined by

$$f_n(q, s, c) = P[U_n = 1 | Q_n = q, S_n = s, C_n = c], \qquad (q, s, c) \in \mathcal{S}. \tag{4.2}$$

Notice that f_n is not defined for $q < c - s$ because in general we have

$$\begin{aligned} Q_n &= C_n - S_n + X_n, \\ &\geq C_n - S_n, \qquad\qquad n = 0, 1, \ldots, \end{aligned} \tag{4.3}$$

where X_n denotes the number of arrivals that occur during the $(n - S_n + 1)$st up to the nth services, which obviously is nonnegative.

4.3 Preliminaries

In this section, we note some basic relations which are crucial to the subsequent development of this chapter. We then introduce the notion of stochastic order between two random processes and recall some well–known facts which we shall use in proving the main results of this chapter in the next section.

First, we observe that at any given time, the server is either serving a customer or taking a vacation. From this fact, the conservation principle

$$D_n = \sum_{j=0}^{n} B_j + \sum_{k=0}^{R_n} V_k, \qquad n = 0, 1, \ldots \tag{4.4}$$

readily follows.

Next, we see from the description of the system that the sequence $\{R_n,\ n = 0, 1, \ldots\}$ evolves according to the recursion

$$R_{n+1} = \begin{cases} R_n + (1 - U_n) & \text{if } Q_n > 0 \\ \min\{l > R_n : \sum_{j=0}^{n} B_j + \sum_{k=0}^{l} V_k \geq T_{n+1}\} & \text{if } Q_n = 0, \end{cases} \qquad n = 0, 1, \ldots, \tag{4.5}$$

with $R_0 = 0$ by Assumption (A1). For the case $Q_n = 0$ in (4.5), we use the fact that the server immediately starts a vacation and continues taking additional vacations until the next customer arrives. Because of Conditions (A2) and (A3), the next customer is necessarily the $(n + 1)$st arriving customer.

We now introduce the notion of stochastic ordering for random processes. First, we define a stochastic order for $\mathbb{R}^k$–valued random variables.

Definition 4.2 *An $\mathbb{R}^k$–valued random variable X^1 is stochastically smaller than another $\mathbb{R}^k$–valued random variable X^2, denoted $X^1 \leq_{\mathrm{st}} X^2$, if*

$$E[f(X^1)] \leq E[f(X^2)] \tag{4.6}$$

for every monotone nondecreasing function $f: \mathbb{R}^k \to \mathbb{R}$ for which the expectations are well defined. Here a monotone nondecreasing function $f: \mathbb{R}^k \to \mathbb{R}$ is understood as a function with the property that $f(x_1, \ldots, x_k) \leq f(y_1, \ldots, y_k)$ whenever $x^i \leq y^i, 1 \leq i \leq k$.

We define a stochastic ordering for random processes as follows.

Definition 4.3 *Let $X^i = \{X^i(t),\ t \in T\}$, $i = 1, 2$, be two $\mathbb{R}$–valued random processes with $T \subseteq \mathbb{R}$. We write $X^1 \leq_{\mathrm{st}} X^2$ if*

$$(X^1(t_1), \ldots, X^1(t_n)) \leq_{\mathrm{st}} (X^2(t_1), \ldots, X^2(t_n)) \tag{4.7}$$

for each $n = 1, 2, \ldots$ and for any $t_1, \ldots, t_n$ in T.

The following result provides an equivalent definition of the stochastic ordering for random processes. It is a special case of Proposition 1.10.4 in Stoyan [77, p. 28] and was originally proved by Kamae et al. [43].

Lemma 4.1 *Let $X^i = \{X^i(t),\ t \in T\}$, $i = 1, 2$, be either two random sequences with $T = \{0, 1, \ldots\}$, or two stochastic processes with sample paths which are right continuous with left limits with $T = [0, \infty)$. Then $X^1 \leq_{\mathrm{st}} X^2$ if and only if there exist two stochastic processes $\{\hat{X}^i(t),\ t \in T\}$, $i = 1, 2$, defined on a common probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})$ such that*

$$\hat{X}^1(t) \leq \hat{X}^2(t), \qquad t \in T \tag{4.8}$$

$\hat{P}$-almost surely (a.s.) and

$$\{\hat{X}^i(t),\ t \in T\} =_{\mathrm{st}} \{X^i(t),\ t \in T\}, \qquad i = 1, 2 \tag{4.9}$$

where $=_{\rm st}$ denotes equivalence in probability law. Furthermore, $X^1 =_{\rm st} X^2$ if and only if we have equality in (4.8).

We shall make use of the following result obtained by O'Brien [63] which allows us to establish a stochastic ordering between two discrete–time random processes by comparing their transition probabilities.

Lemma 4.2 *Let $X^i = \{X_n^i,\ n = 0,1,\ldots\}$, $i = 1,2$, be two sequences of $\mathbb{R}$–valued random variables. If*

$$X_0^1 \leq_{\rm st} X_0^2 \tag{4.10}$$

and if for each $n = 0,1,\ldots,$ the inequalities

$$x_j^1 \leq x_j^2, \qquad j = 0,\ldots,n \tag{4.11}$$

imply

$$P[X_{n+1}^1 \leq x | X_j^1 = x_j^1, j = 0,\ldots,n] \geq P[X_{n+1}^2 \leq x | X_j^2 = x_j^2, j = 0,\ldots,n], \quad x \in \mathbb{R}, \tag{4.12}$$

then we have

$$X^1 \leq_{\rm st} X^2. \tag{4.13}$$

4.4 The Main Comparison Results

In this section, we derive the main result of the chapter. We first describe the probabilistic setting as follows: On a probability space $(\Omega, \mathcal{F}, P)$, define the interarrival, service and vacation processes $\{A_{n+1}, B_n, V_n,\ n = 0,1,\ldots\}$ and two simple service policies U^1 and U^2, where $U^i = \{U_n^i,\ n = 0,1,\ldots\}$, $i = 1,2$. For each $i = 1,2$, construct a vacation model from the sequence of random variables $\{A_{n+1}, B_n, V_n, U_n^i,\ n = 0,1,\ldots\}$ in the fashion described in Section 2. Notice that the two vacation models share the same interarrival, service and vacation processes. In the sequel, we shall use superscript $i = 1,2$ for any quantity associated with the service policy U^i. Our goal is to identify some general conditions on U^1 and U^2 that lead to a comparison between the sequences $\{R_n^1,\ n = 0,1,\ldots\}$ and $\{R_n^2,\ n = 0,1,\ldots\}$.

We shall see that this is a key comparison in that comparison results for other quantities of interest can be readily derived from it.

We first prove the following simple lemma which we shall use in the proof of the main result. This lemma, which is a direct consequence of (4.4) and (4.5), holds for a vacation model under *any* pair of service policies.

Lemma 4.3 *For* $n = 0, 1, \ldots,$

$$R_n^1 < R_n^2 \qquad \textit{implies} \qquad R_{n+1}^1 \le R_{n+1}^2. \tag{4.14}$$

Proof: The proof relies solely on equations (4.4) and (4.5). We consider the following three cases.

Case (i) — $\mathbf{Q_n^1 > 0}$: In this case, we see from (4.5) that $R_{n+1}^1 \le R_n^1 + 1$, and so $R_n^1 < R_n^2$ implies $R_{n+1}^1 \le R_n^2$. Since $R_n^2 \le R_{n+1}^2$, we thus have $R_{n+1}^1 \le R_{n+1}^2$.

Case (ii) — $\mathbf{Q_n^1 = Q_n^2 = 0}$: Since $Q_n^1 = 0$, we see from (4.5) that R_{n+1}^1 is given by

$$R_{n+1}^1 = \min\{l > R_n^1 : \sum_{j=0}^{n} B_j + \sum_{k=0}^{l} V_k \ge T_{n+1}\}. \tag{4.15}$$

But, $Q_n^2 = 0$ implies

$$D_n^2 = \sum_{j=0}^{n} B_j + \sum_{k=0}^{R_n^2} V_k < T_{n+1}, \tag{4.16}$$

and so, since $R_n^2 > R_n^1$, the minimization in (4.15) can be restricted to $l > R_n^2$, i.e.,

$$\begin{aligned} R_{n+1}^1 &= \min\{l > R_n^2 : \sum_{j=0}^{n} B_j + \sum_{k=0}^{l} V_k \ge T_{n+1}\} \\ &= R_{n+1}^2. \end{aligned} \tag{4.17}$$

Case (iii) — $\mathbf{Q_n^1 = 0}$ and $\mathbf{Q_n^2 > 0}$: Since $Q_n^2 > 0$, we have

$$D_n^2 = \sum_{j=1}^{n} B_j + \sum_{k=1}^{R_n^2} V_k \ge T_{n+1} \tag{4.18}$$

and so R_n^2 satisfies the condition of the minimization in (4.15). Hence, $R_{n+1}^1 \le R_n^2 \le R_{n+1}^2$. □

As it will become apparent later, the probabilistic structure of the random sequence $\{A_n, B_n, V_n,\ n = 0, 1, \ldots\}$ does not play any role in the proof of the comparison results. Consequently, we may assume $\{A_n, B_n, V_n,\ n = 0, 1, \ldots\}$ to be *any* deterministic sequence. In other words, in the case where the random variables $\{A_n, B_n, V_n,\ n = 0, 1, \ldots\}$ are indeed random, we can obtain stronger comparison results, i.e., the comparisons can be shown to hold on any given sample path of $\{A_n, B_n, V_n,\ n = 0, 1, \ldots\}$.

We shall assume in the sequel that the sequence $\{A_n, B_n, V_n,\ n = 0, 1, \ldots\}$ is a deterministic sequence. To remind us of this, we shall denote the sequence by $\{a_n, b_n, v_n,\ n = 0, 1, \ldots\}$. In this setting, simple service policies U^1 and U^2 each satisfies $(i = 1, 2)$

$$P[U_n^i = 1 | U_j^i,\ j = 0, \ldots, n-1] = P[U_n^i = 1 | Q_n^i, S_n^i, C_n^i], \qquad n = 0, 1, \ldots. \tag{4.19}$$

We also note that the sets of random variables $\{U_j^i,\ j = 0, \ldots, n-1\}$ and $\{R_j^i,\ j = 0, \ldots, n\}$ both contain the same amount of information, i.e., either set can be constructed from the other. This fact, together with (4.19), implies

$$P[U_n^i = 1 | R_j^i,\ j = 0, \ldots, n] = P[U_n^i = 1 | Q_n^i, S_n^i, C_n^i], \qquad n = 0, 1, \ldots. \tag{4.20}$$

We are now in the position to prove the following theorem.

Theorem 4.1 *Let U^1 and U^2 be two simple policies, and suppose for each $n = 0, 1, \ldots$ that the mapping f_n^1, $f_n^2 \colon \mathcal{S} \to [0,1]$ satisfy*

$$f_n^1(q, s, c) \geq f_n^2(q, s', c'), \qquad \textit{for each } (q, s, c),\ (q, s', c') \in \mathcal{S}$$
$$\textit{such that } s \leq s',\ c - s \geq c' - s'. \tag{4.21}$$

Then, the stochastic comparison

$$\{R_n^1,\ n = 0, 1, \ldots\} \leq_{\text{st}} \{R_n^2,\ n = 0, 1, \ldots\} \tag{4.22}$$

holds.

Proof: By Lemma 4.2, it suffices to show that

$$R_0^1 \leq_{\mathrm{st}} R_0^2 \tag{4.23}$$

and that for each $n = 0, 1, \ldots$, the inequalities

$$r_j^1 \leq r_j^2, \qquad j = 0, \ldots, n \tag{4.24}$$

imply

$$P[R_{n+1}^1 \leq r | R_j^1 = r_j^1,\ j = 0, \ldots, n] \geq P[R_{n+1}^2 \leq r | R_j^2 = r_j^2,\ j = 0, \ldots, n],$$
$$r = 0, 1, \ldots \tag{4.25}$$

Inequality (4.23) is trivial since $R_0^1 = R_0^2 = 0$. Now assuming that (4.24) holds for some $n = 1, 2, \ldots$, we show that (4.25) holds. To that end, we distinguish the following three cases based only on the choice of $\{r_j^i,\ j = 0, \ldots, n\}$, $i = 1, 2$.

Case (i) — $\mathbf{r_n^1 < r_n^2}$: In this case, (4.25) follows directly by Lemma 4.3 since $R_n^1 = r_n^1 < r_n^2 = R_n^2$ implies $R_{n+1}^1 \leq R_{n+1}^2$.

Case (ii) — $\mathbf{r_n^1 = r_n^2}$ and $\mathbf{Q_n^2 = 0}$: Letting $R_n^1 = r_n^1 = r_n^2 = R_n^2$, we have from (4.4) that $D_n^1 = D_n^2$. Since the same arrival process is used under both policies, we necessarily have $Q_n^1 = Q_n^2 = 0$. The server under both policies starts a vacation at time $D_n^1 = D_n^2$ and continues to take additional vacations until the $(n+1)$st customer arrives. Again, since the same arrival process is used, we have $R_{n+1}^1 = R_{n+1}^2$, and so (4.25) holds.

Case (iii) — $\mathbf{r_n^1 = r_n^2}$ and $\mathbf{Q_n^2 > 0}$: As in Case (ii), we let $R_n^1 = r_n^1 = r_n^2 = R_n^2$, and so we have $Q_n^1 = Q_n^2 > 0$. By (4.5), we have

$$R_{n+1}^i = R_n^i + (1 - U_n^i), \qquad i = 1, 2, \tag{4.26}$$

and so, to obtain (4.25), it suffices to show that

$$P[U_n^1 = 1 | R_j^1 = r_j^1,\ j = 1, \ldots, n] \geq P[U_n^2 = 1 | R_j^2 = r_j^2,\ j = 1, \ldots, n]. \tag{4.27}$$

But, this immediately follows from (4.20) and (4.21) if we can show that in this case,

$$S_n^1 \leq S_n^2 \tag{4.28}$$

and

$$C_n^1 - S_n^1 \geq C_n^2 - S_n^2. \tag{4.29}$$

Inequality (4.29) follows from (4.28). Indeed,

$$C_n^i - S_n^i = Q_n^i - X_n^i, \qquad i = 1, 2, \tag{4.30}$$

where X_n^i is the number of arrivals that occur during the $(n - S_n^i + 1)$st up to the nth services in the system under U^i. Since $S_n^1 \leq S_n^2$, $D_n^1 = D_n^2$ and a common arrival process is used, we obviously have $X_n^1 \leq X_n^2$. This and the fact that $Q_n^1 = Q_n^2$, readily imply (4.29) via (4.30).

To prove (4.28), assume that $S_n^1 > S_n^2$. From the definition of S_n^1 and S_n^2, we see that, under the policy U^i, $i = 1, 2$, the $(n - S_n^i + 1)$st up to the nth services are not interrupted by any vacation, whereas the $(n - S_n^i)$th and the $(n - S_n^i + 1)$st services are separated by at least one vacation. Thus, we have

$$R_{n-S_n^i}^i < R_{n-S_n^i+1}^i = R_{n-S_n^i+2} = \cdots = R_n^i, \qquad i = 1, 2. \tag{4.31}$$

Since $S_n^1 > S_n^2$, we have $n - S_n^1 + 1 \leq n - S_n^2 \leq n - 1$, and thus from (4.31) (with $i = 1$) we see that

$$R_{n-S_n^2}^1 = R_n^1. \tag{4.32}$$

But, (4.31) also implies $R_{n-S_n^2}^2 < R_n^2$, and so, since $R_n^1 = R_n^2$, we have

$$R_{n-S_n^2}^2 < R_{n-S_n^2}^1. \tag{4.33}$$

This contradicts (4.24), and therefore we must have $S_n^1 \leq S_n^2$.

The proof of Theorem 4.1 is now complete. □

Remark 4.1: We have proved Theorem 4.1 without using any knowledge of the order in which the customers are served.

Remark 4.2: Condition (4.21) states that given the same number of customers in the queue, the same (or smaller) number of customers that have been served since the last vacation and the same (or larger) number of unserved customers which were in the queue when the server returned from the last vacation, the server is less likely to go on vacation under the policy U^1 than under the policy U^2.

Using Theorem 4.1, comparisons of other processes in the vacation model can be readily made. By Lemma 4.1, there exists a common probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})$ where the random variables $\{\hat{R}_n^1,\ n = 0, 1, \ldots\}$ and $\{\hat{R}_n^2,\ n = 0, 1, \ldots\}$ are defined. For each $i = 1, 2$, the probabilistic structure of $\{\hat{R}_n^i,\ n = 0, 1, \ldots\}$ under $\hat{P}$ is identical to that of $\{R_n^i,\ n = 0, 1, \ldots\}$ under P and, furthermore, the comparison

$$\hat{R}_n^1 \leq \hat{R}_n^2, \qquad n = 0, 1, \ldots \tag{4.34}$$

holds $\hat{P}$–a.s.

The comparison between the departure processes $\{D_n^1,\ n = 0, 1, \ldots\}$, $i = 1, 2$, can be made in the following manner. On $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})$, define

$$\hat{D}_n^i = \sum_{j=1}^{n} b_j + \sum_{k=1}^{\hat{R}_n^i} v_k, \qquad n = 0, 1, \ldots;\ i = 1, 2. \tag{4.35}$$

Obviously, for each $i = 1, 2$, the probabilistic structure of $\{\hat{D}_n^i,\ n = 0, 1, \ldots\}$ under $\hat{P}$ is identical to that of $\{D_n^i,\ n = 0, 1, \ldots\}$ under P and

$$\hat{D}_n^1 \leq \hat{D}_n^2, \qquad n = 0, 1, \ldots \tag{4.36}$$

holds $\hat{P}$–a.s. Thus, by Lemma 4.1, the stochastic comparison $\{D_n^1,\ n = 0, 1, \ldots\} \leq_{\text{st}} \{D_n^2,\ n = 0, 1, \ldots\}$ holds.

The comparison of other processes such as the queue size process $\{N^i(t),\ t \geq 0\}$ and the waiting time process $\{W_n^i,\ n = 0, 1, \ldots\}$ (if the service discipline is FCFS) can be made in a similar fashion. Indeed, for $i = 1, 2$ and $t \geq 0$, $N^i(t) = A(t) - D^i(t)$ where $A(t) = \max\{k : T_k \leq t\}$ and $D^i(t) = \max\{k : D_k^i \leq t\}$. If the service is FCFS, we have $W_n^i = D_n^i - B_n - A_n$ for $n = 0, 1, \ldots$ and $i = 1, 2$.

Corollary 4.1 *For the vacation model and the service policies in Theorem 4.1, the stochastic comparisons*

$$\{D_n^1,\ n = 0, 1, \ldots\} \leq_{\text{st}} \{D_n^2,\ n = 0, 1, \ldots\} \tag{4.37}$$

and

$$\{N^1(t),\ t \geq 0\} \leq_{\text{st}} \{N^2(t),\ t \geq 0\} \tag{4.38}$$

hold. Furthermore, if the service discipline is FCFS, then the comparison

$$\{W_n^1,\ n=0,1,\ldots\} \leq_{\mathrm{st}} \{W_n^2,\ n=0,1,\ldots\} \tag{4.39}$$

also holds.

It should also be mentioned that stochastic ordering between two random processes readily extends to stochastic ordering between their limiting random variables, whenever they exist [77, Prop. 1.2.3].

4.5 The Comparison of Various Service Policies

In this section, we use the results established in the previous section to compare various service policies. The most extensively analyzed policies in the literature are the *exhaustive* and *gated* policies. Under the exhaustive policy, the server goes on vacation if and only if the queue is empty. Under the gated policy, the server, if it finds some customers in the queue upon returning from a vacation, serves these customers continuously before taking a vacation; those customers that arrive during the service of these customers are served when the server returns from the next vacation.

Among the policies introduced later are the *limited* and *Bernoulli* policies. Under the limited policy, there is a fixed integer m such that the number of consecutive services that the server can perform is limited to m. The limited policy comes in two variants: the *limited–exhaustive* and *limited–gated.* Under the limited–gated policy, the server, if it finds n customers in the queue upon returning from a vacation, serves continuously $\min(n,m)$ customers before taking a vacation. Under the limited–exhaustive policy, those customers that arrive during the service of the original n customers can still be served as long as the limit m has not been reached. Under the Bernoulli policy, there is a fixed probability p, $0 \leq p \leq 1$, such that at every service completion, the server serves the next available customer (if any) with probability p and goes on vacation with probability $1-p$. The Bernoulli policy also comes in two variants: the *Bernoulli–exhaustive* and the *Bernoulli–gated.*

Note that all the above-mentioned service policies are simple. Using our notation, we can describe these policies more precisely as follows ($n = 0, 1, \ldots$):

- The exhaustive policy (denoted by E):

$$f_n^E(q,s,c) = \begin{cases} 1 & \text{if } q \geq 1 \\ 0 & \text{otherwise;} \end{cases} \tag{4.40}$$

- The gated policy (denoted by G):

$$f_n^G(q,s,c) = \begin{cases} 1 & \text{if } c - s \geq 1 \\ 0 & \text{otherwise;} \end{cases} \tag{4.41}$$

- The limited-exhaustive schedule with parameter m (denoted by $LE(m)$), $m = 1, 2, \ldots$:

$$f_n^{LE(m)}(q,s,c) = \begin{cases} 1 & \text{if } q \geq 1 \text{ and } s < m \\ 0 & \text{otherwise;} \end{cases} \tag{4.42}$$

- The limited-gated schedule with parameter m (denoted by $LG(m)$), $m = 1, 2, \ldots$:

$$f_n^{LG(m)}(q,s,c) = \begin{cases} 1 & \text{if } c - s \geq 1 \text{ and } s < m \\ 0 & \text{otherwise;} \end{cases} \tag{4.43}$$

- The Bernoulli-exhaustive schedule with parameter p (denoted by $BE(p)$), $0 \leq p \leq 1$:

$$f_n^{BE(p)}(q,s,c) = \begin{cases} p & \text{if } q \geq 1 \\ 0 & \text{otherwise;} \end{cases} \tag{4.44}$$

- The Bernoulli-gated schedule with parameter p (denoted by $BG(p)$), $0 \leq p \leq 1$:

$$f_n^{BG(p)}(q,s,c) = \begin{cases} p & \text{if } c - s \geq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{4.45}$$

4.5.1 The Policies E and $LE(1)$ as Extreme Policies

We show in the following theorem that the policies E and $LE(1)$ bound any simple policy from below and from above, respectively, in terms of the quantities of interest discussed in Section 4.

Theorem 4.2 *For any simple service policy U, the stochastic comparisons*

$$\{R_n^E,\ n=0,1,\ldots\} \le_{\rm st} \{R_n^U,\ n=0,1,\ldots\} \le_{\rm st} \{R_n^{LE(1)},\ n=0,1,\ldots\}, \qquad (4.46)$$

$$\{D_n^E,\ n=0,1,\ldots\} \le_{\rm st} \{D_n^U,\ n=0,1,\ldots\} \le_{\rm st} \{D_n^{LE(1)},\ n=0,1,\ldots\}, \qquad (4.47)$$

and

$$\{N(t)^E,\ t\ge 0\} \le_{\rm st} \{N(t)^U,\ t\ge 0\} \le_{\rm st} \{N(t)^{LE(1)},\ t\ge 0\} \qquad (4.48)$$

hold. Furthermore, if the service discipline is FCFS, the comparison

$$\{W_n^E,\ n=0,1,\ldots\} \le_{\rm st} \{W_n^U,\ n=0,1,\ldots\} \le_{\rm st} \{W_n^{LE(1)},\ n=0,1,\ldots\} \qquad (4.49)$$

also holds.

Proof: For any service policy U, we have

$$f_n^U(q,s,c) = 0 \qquad \text{if} \qquad q=0, \qquad (4.50)$$

and so we see from (4.40) that

$$f_n^E(q,s,c) \ge f_n^U(q,s',c'), \qquad \text{for each } (q,s,c),\ (q,s',c') \in \mathcal{S} \text{ such that } s \le s',\ c-s \ge c'-s'. \qquad (4.51)$$

On the other hand, for the policy $LE(1)$ (which is identical to the policies $LG(1)$, $BE(0)$ and $BG(0)$), we have

$$f_n^{LE(1)}(q,s,c) = 0, \qquad (q,s,c) \in \mathcal{S}, \qquad (4.52)$$

and so

$$f_n^U(q,s,c) \ge f_n^{LE(1)}(q,s',c'), \qquad \text{for each } (q,s,c),\ (q,s',c') \in \mathcal{S} \text{ such that } s \le s',\ c-s \ge c'-s'. \qquad (4.53)$$

By Theorem 4.1 and Corollary 4.1, inequalities (4.51) and (4.53) readily yield the desired result. □

4.5.2 Gated–Type vs. Nongated–Type Policies

Let us define a service policy as of *gated–type* if it satisfies

$$f_n^U(q,s,c)=0, \quad \text{whenever} \quad c-s\le 0, \qquad n=0,1,\ldots. \tag{4.54}$$

We have the following result.

Theorem 4.3 *Let U^1 and U^2 be two simple service policies each satisfying (i=1,2)*

$$f_n^i(q,s,c)\ge f_n^i(q,s',c'), \qquad \textit{for each } (q,s,c),\ (q,s',c')\in\mathcal{S}$$
$$\textit{such that } s\le s',\ c-s\ge c'-s'. \tag{4.55}$$

If U^2 is of gated–type and

$$f_n^1(q,s,c)=f_n^2(q,s,c), \qquad q,s,c=1,2,\ldots;1\le c-s\le q, \tag{4.56}$$

then the stochastic comparisons (4.22), (4.37), (4.38) and (4.39) (if the service is FCFS) hold.

Proof: It can be easily shown from (4.54)–(4.56) that the condition (4.21) is satisfied, and so, by Theorem 4.1 and Corollary 4.1, the desired result readily follows. □

Remark 4.3: Condition (4.55) basically states that under the policy U^i, the server is more likely to go on vacation the more customers it has served and the fewer the customers which were in the queue when the server returned from the last vacation are still unserved. Most (if not all) service policies in the literature satisfy this condition.

4.6 Stochastic Monotonicity Properties

The limited and Bernoulli service policies introduced in the previous section share a common characteristic in that each is a family of service policies parameterized by a single parameter, i.e., the limit m for the limited policy and the probability

p for the Bernoulli policy. In this section, we show some stochastic monotonicity properties of these service policies with respect to their respective parameter.

Consider the policies $LE(m_1)$ and $LE(m_2)$ with $m_2 \leq m_1$. From (4.42), we observe that

$$f_n^{LE(m_1)}(q,s,c) \geq f_n^{LE(m_2)}(q,s',c'), \qquad \text{for each } (q,s,c),\,(q,s',c') \in \mathcal{S} \text{ such that } s \leq s',\; c-s \geq c'-s', \tag{4.57}$$

and so by Theorem 4.1, the policy $LE(m_2)$ dominates the policy $LE(m_1)$ in terms of the various processes we have been considering. In other words, these processes under the service policy $LE(m)$ are stochastically monotone decreasing with respect to the parameter m. Such a stochastic monotonicity property can also be shown for the policies $LG(m)$, $BE(p)$ and $BG(p)$.

Theorem 4.4 *Under the service policies $LE(m)$, $LG(m)$, $BE(p)$ and $BG(p)$, the random processes*

$$\begin{gathered}\{R_n,\ n=0,1,\ldots\},\ \{D_n,\ n=0,1,\ldots\},\ \{N(t),\ t\geq 0\} \\ \textit{and } \{W_n,\ n=0,1,\ldots\} \textit{ (if the service is FCFS)}\end{gathered} \tag{4.58}$$

are stochastically monotone decreasing with respect to the respective service policy parameter.

4.7 Conclusions

In this chapter, we have developed a general framework for stochastic comparisons between two service policies in multiple–vacation models. In this framework, we are able to establish comparison results between some well–known service policies. We also obtained stochastic monotonicity properties of some parametrizable service policies.

In an independent study, Levy et al. [57] have obtained comparison results for various service policies in polling systems. The process they consider is the total amount of unfinished work found in the system. The difference between their approach and ours is in the characterization of a service policy. Unlike theirs, our characterization facilitates comparison between exhaustive– and gated–type policies.

Unfortunately, we cannot compare the limited and Bernoulli policies in the general framework established in this chapter. The comparison between these two policies is discussed in the next chapter. The comparison is in the convex increasing ordering, which is weaker than the stochastic order discussed in this chapter.

CHAPTER 5

COMPARING BERNOULLI AND LIMITED POLICIES

5.1 Introduction

In this chapter, we compare the limited and Bernoulli policies. Fuhrmann [32] obtained an upper bound for the steady–state mean waiting time of a symmetric polling system under a limited service policy. Servi and Yao [73] subsequently showed that, in the context of vacation models, this bound was exactly the mean waiting time under a Bernoulli policy (with suitably chosen parameters). Servi and Yao also obtained other comparison results which suggest that the steady–state waiting times for the limited and Bernoulli policies can be compared in the increasing convex ordering. In this chapter, we show that such a comparison indeed holds when the vacation periods are identical and deterministic. In fact, the comparison holds not only in the steady–state but also in the transient regime.

In the next section, we introduce the notion of increasing convex ordering. There, we also note some preliminary results crucial to the subsequent development. In Section 3 and 4, we derive the comparison results for the transient and steady–state cases, respectively.

5.2 Preliminaries

Since we only consider the exhaustive variant of both the limited (with parameter m) and Bernoulli (with paramter p) policies, we shall denote these policies by $L(m)$ and $B(p)$, respectively. In the sequel, we shall use superscript $B(p)$ (resp. $L(m)$) for any quantity associated with the service policy $B(p)$ (resp. $L(m)$). Let us briefly recall the definitions of these two policies. We shall assume in this chapter that service is given on a FCFS basis. A vacation model under the policy $L(m)$ is described as follows, where we use the notation of Chapter 4. At time D_n, $n = 0, 1, \ldots$, the server finishes serving the nth customer (denoted by C_n). If $Q_n > 0$, the server immediately serves C_{n+1} if $1 \leq S_n < m$; otherwise if $S_n = m$, then the server takes a vacation (of length V_{R_n+1}) and then serves C_{n+1}. Under the policy $B(p)$, the server's action in the case $Q_n > 0$ is determined by the random variable X_n: If $X_n = 0$, the server serves C_{n+1}; if $X_n = 1$, it takes one vacation and then serves C_{n+1}. The sequence $\{X_n,\ n = 0, 1, \ldots\}$ is composed of i.i.d. $\{0,1\}$–valued random variables with

$$P[X_n = 1] = 1 - p \qquad n = 0, 1, \ldots \tag{5.1}$$

and is assumed to be independent of the random variables $\{A_n, B_n, V_n,\ n = 0, 1, \ldots\}$. Under either policy, if $Q_n = 0$, the server takes a vacation and keeps on taking additional vacations until at least one customer is present when it returns from a vacation.

We note below several simple relations which are crucial to the discussion in the next sections. First, from the conservation principle (4.4), we have

$$D_n^i = \sum_{j=0}^{n} B_j + \sum_{k=0}^{R_n^i} V_k, \quad n = 0, 1, \ldots \tag{5.2}$$

for $i = B(p)$, $L(m)$.

From the description of the policies $B(p)$ and $L(m)$ given above, we obtain the following recursive equations describing the evolution of R_n under each policy. Under the service policy $B(p)$, we have

$$R_{n+1}^{B(p)} = \begin{cases} R_n^{B(p)} + X_n & \text{if } Q_n^{B(p)} > 0 \\ \min\{l > R_n^{B(p)} \colon \sum_{j=0}^{n} B_j + \sum_{k=0}^{l} V_k \geq T_{n+1}\} & \text{if } Q_n^{B(p)} = 0, \end{cases} \qquad n = 0, 1, \ldots, \tag{5.3}$$

whereas under the service policy $L(m)$, we have

$$R_{n+1}^{L(m)} = \begin{cases} R_n^{L(m)} & \text{if } Q_n^{L(m)} > 0 \text{ and } S_n^{L(m)} < m \\ R_n^{L(m)} + 1 & \text{if } Q_n^{L(m)} > 0 \text{ and } S_n^{L(m)} = m \\ \min\{l > R_n^{L(m)} \colon \sum_{j=0}^{n} B_j + \sum_{k=0}^{l} V_k \geq T_{n+1}\} & \text{if } Q_n^{L(m)} = 0. \end{cases}$$

$$n = 0, 1, \ldots . \tag{5.4}$$

Notice that under both the limited and Bernoulli policies, we have

$$R_{n+1}^i = \min\{l > R_n^i : \sum_{j=0}^{n} B_j + \sum_{k=0}^{l} V_k \geq T_{n+1}\} \text{ if } Q_n^i = 0, \qquad n = 0, 1, \ldots. \tag{5.5}$$

with $i = B(p), L(m)$.

We now introduce the notion of convex increasing ordering [69, p. 270].

Definition 5.1 *We say that a real-valued random variable X is smaller than another real-valued random variable Y in the convex increasing ordering, denoted by $X \leq_{\text{icx}} Y$, if*

$$E[f(X)] \leq E[f(Y)] \tag{5.6}$$

for every monotone non-decreasing and convex function $f: I\!R \to I\!R$ for which the expectations are well defined.

We shall make use of the following facts whose proofs can be found in Ross [69, Chap. 8] and Stoyan [77].

(F1) If X and Y are nonnegative random variables, then $X \leq_{\text{icx}} Y$ if and only if

$$E[(X-a)^+] \leq E[(Y-a)^+], \qquad a \geq 0, \tag{5.7}$$

with the notation $(x)^+ = \max\{0, x\}$;

(F2) If $E[X]$ is finite, then it follows from Jensen's inequality that

$$E[X] \leq_{\text{icx}} X; \tag{5.8}$$

(F3) Let Z be independent of X and Y. If $X \leq_{\text{icx}} Y$, then

$$X + Z \leq_{\text{icx}} Y + Z. \tag{5.9}$$

5.3 The Comparison in The Transient Regime

In this section, we prove that for suitably chosen parameters p and m, the waiting time of the nth customer under the policy $L(m)$ is smaller than that under the policy $B(p)$ in the ordering $\leq_{\text{icx}}$. In the next section, this comparison is shown to extend to the steady–state waiting time distributions, whenever they exist. We first prove the following lemmas.

Lemma 5.1 *Consider any two service policies U_1 and U_2 each satisfying (5.5). Then, for all $n = 0, 1, \ldots,$ we have*

$$R^2_{n+1} \leq R^1_{n+1} \qquad \textit{whenever} \qquad Q^2_n = 0. \tag{5.10}$$

Proof: Since $Q^2_n = 0$, we have $D^2_n < T_{n+1}$ which by (5.2) (with $i = 2$) implies

$$\sum_{j=0}^{n} B_j + \sum_{k=0}^{R^2_n} V_k < T_{n+1}. \tag{5.11}$$

We consider the following three cases which exhaust all the possibilities.

Case (i) — $\mathbf{R^1_n \leq R^2_n}$:

By (5.2), we have $D^1_n \leq D^2_n < T_{n+1}$ and so necessarily $Q^1_n = 0$. Since U_1 satisfies (5.5), we then have

$$R^1_{n+1} = \min\{l > R^1_n : \sum_{j=0}^{n} B_j + \sum_{k=0}^{l} V_k \geq T_{n+1}\}. \tag{5.12}$$

But, because of (5.11), we can restrict the minimization in (5.12) to $l > R^2_n$, i.e.,

$$R^1_{n+1} = \min\{l > R^2_n : \sum_{j=0}^{n} B_j + \sum_{k=0}^{l} V_k \geq T_{n+1}\}, \tag{5.13}$$

and therefore $R^1_{n+1} = R^2_{n+1}$ by making use of (5.5).

Case (ii) — $\mathbf{R^1_n > R^2_n}$ and $\mathbf{Q^1_n = 0}$:

This case is similar to Case (i) since both Q^1_n and Q^2_n are zero, and the same argument holds with U_1 and U_2 interchanged to yield $R^1_{n+1} = R^2_{n+1}$.

Case (iii) — $\mathbf{R^1_n > R^2_n}$ and $\mathbf{Q^1_n > 0}$:

Since $Q^1_n > 0$, we have

$$D^1_n = \sum_{j=0}^{n} B_j + \sum_{k=0}^{R^1_n} V_k > T_{n+1}. \tag{5.14}$$

From the definition of R^2_{n+1} in (5.5), it then follows that

$$R^2_{n+1} \leq R^1_n \tag{5.15}$$

and this last fact trivially implies $R^2_{n+1} \leq R^1_{n+1}$.

This completes the proof of Lemma 5.1. □

Lemma 5.2 *Consider a vacation model under the policies $B(p)$ and $L(m)$ where p and m are related by*

$$1 - p = 1/m. \tag{5.16}$$

Then, conditioned on any realization of the random sequence $\{A_n, B_n, V_n,\ n = 0, 1, \ldots\}$, the comparison

$$R_n^{L(m)} \leq_{\mathrm{icx}} R_n^{B(p)} \qquad n = 0, 1, \ldots \tag{5.17}$$

holds.

Proof: The proof proceeds by induction. By assumption (A1) in the previous chapter, we have $R_0^{L(m)} = R_0^{B(p)} = 0$, and so (5.17) is automatically satisfied for $n = 0$. Assuming that (5.17) holds for $n = 0, 1, \ldots, k$, we are to show that

$$R_{k+1}^{L(m)} \leq_{\mathrm{icx}} R_{k+1}^{B(p)} \tag{5.18}$$

holds.

We identify three mutually exclusive cases: (i) $Q_k^{L(m)} > 0$, $1 \leq S_k^{L(m)} < m$; (ii) $Q_k^{L(m)} > 0$, $S_k^{L(m)} = m$; and (iii) $Q_k^{L(m)} = 0$ based on the realization of the system under $L(m)$. Since the policy $L(m)$ contains no randomness, given a sample path of $\{A_n, B_n, V_n,\ n = 0, 1, \ldots\}$, the realization of the system under $L(m)$ is deterministic, i.e. there is only one such realization. So, only one of the three cases mentioned above can occur, and this case occurs with probability one. So, "restricted" to this case, the asserted induction hypothesis still holds true.

We first observe from (5.3) that for a Bernoulli policy,

$$R_n^{B(p)} + X_n \leq R_{n+1}^{B(p)} \qquad n = 0, 1, \ldots. \tag{5.19}$$

Here we use the fact that if $Q_n^{B(p)} = 0$ then the server must take at least one vacation, i.e., $R_n^{B(p)} + 1 \leq R_{n+1}^{B(p)}$ and otherwise $R_{n+1}^{B(p)} = R_n^{B(p)} + X_n$.

Case (i) — $\mathbf{Q_k^{L(m)} > 0,\ 1 \leq S_k^{L(m)} < m}$:

In this case we have $R_{k+1}^{L(m)} = R_k^{L(m)}$ from (5.4). It then follows from the induction hypothesis and (5.19) that

$$R_{k+1}^{L(m)} = R_k^{L(m)} \leq_{\mathrm{icx}} R_k^{B(p)} \leq R_{k+1}^{B(p)}. \tag{5.20}$$

Case (ii) — $\mathrm{Q}_{\mathrm{k}}^{\mathrm{L(m)}} > 0$, $\mathrm{S}_{\mathrm{k}}^{\mathrm{L(m)}} = \mathrm{m}$:

Since necessarily $k \geq m-1$, we can apply (5.19) m times to obtain

$$R_{k+1-m}^{B(p)} + \sum_{j=k+1-m}^{k} X_j \leq R_{k+1}^{B(p)}. \tag{5.21}$$

The enforced relation (5.16) implies that $E[\sum_{j=k+1-m}^{k} X_j] = 1$, and so by (F2) we have

$$1 \leq_{\text{icx}} \sum_{j=k+1-m}^{k} X_j. \tag{5.22}$$

Since $R_{k+1-m}^{B(p)}$ is independent of $\{X_j,\ j = k+1-m, \ldots, k\}$, we conclude from (F3), (5.21) and the induction hypothesis that

$$R_{k+1-m}^{L(m)} + 1 \leq_{\text{icx}} R_{k+1}^{B(p)}. \tag{5.23}$$

The desired inequality (5.18) then follows since we have $R_{k+1-m}^{L(m)} + 1 = R_{k+1}^{L(m)}$ from (5.4).

Case (iii) — $\mathrm{Q}_{\mathrm{k}}^{\mathrm{L(m)}} = 0$:

In this case $R_{k+1}^{L(m)} \leq R_{k+1}^{B(p)}$ by Lemma 5.1 which trivially implies (5.18). □

We are now in a position to prove the main result of this section.

Theorem 5.1 *Consider a vacation model with deterministic vacation periods, i.e.,*

$$V_n = V, \qquad n = 0, 1, \ldots \tag{5.24}$$

for some constant V. For policies $B(p)$ and $L(m)$ satisfying (5.16), the comparison

$$W_n^{L(m)} \leq_{\text{icx}} W_n^{B(p)} \qquad n = 0, 1, \ldots \tag{5.25}$$

holds.

Proof: From (5.2) and (5.24), we conclude that for $i = B(p), L(m)$

$$D_n^i = \sum_{j=0}^{n} B_j + VR_n^i, \qquad n = 0, 1, \ldots. \tag{5.26}$$

Let us consider a sample path of $\{A_n, B_n,\ n = 0, 1, \ldots\}$. We see from Lemma 5.2 and (F3) that conditioned on this sample path, we have $D_n^{L(m)} \leq_{\text{icx}} D_n^{B(p)}$. Since the waiting time of C_n can be expressed as

$$W_n^i = D_n^i - B_n - T_n \qquad i = B(p), L(m);\ n = 0, 1, \ldots, \tag{5.27}$$

we again obtain from (F3) that $W_n^{L(m)} \leq_{\text{icx}} W_n^{B(p)}$. Unconditioning with respect to $\{A_n, B_n,\ n = 0, 1, \ldots\}$ completes the proof. □

5.4 The Comparison in The Steady–State Regime

In this section, the comparison result derived in the previous section is extended to the steady–state regime. We assume that under both $B(p)$ and $L(m)$ (with p and m related by (5.16)) the vacation model is stable, i.e.,

$$W_n^{B(p)} \xrightarrow{\mathcal{D}} W^{B(p)} \quad \text{and} \quad W_n^{L(m)} \xrightarrow{\mathcal{D}} W^{L(m)}, \tag{5.28}$$

where $W^{B(p)}$ and $W^{L(m)}$ are almost surely finite. It is well known [45,51] that the necessary and sufficient condition for this stability is given by

$$\bar{B} + (1-p)V = \bar{B} + \frac{1}{m}V < \bar{A}, \tag{5.29}$$

where $\bar{B} = E[B_n]$ and $\bar{A} = E[A_n]$.

Theorem 5.2 *For the vacation model of Theorem 5.1 under the stability condition (5.28), the comparison*

$$W^{L(m)} \leq_{\text{icx}} W^{B(p)} \tag{5.30}$$

holds.

Before we prove the theorem above, we first establish some general results stated in the following lemmas.

Lemma 5.3 *Let a modified vacation model $\{\hat{A}_n, \hat{B}_n,\ n = 0, 1, \ldots\}$ be constructed from the random variables $\{A_n, B_n, X_n, n = 0, 1, \ldots\}$ where $\hat{A}_n$ and $\hat{B}_n$ are given by*

$$\hat{A}_n = A_n \qquad \hat{B}_n = B_n + X_n V, \qquad n = 0, 1, \ldots \tag{5.31}$$

and the length of each vacation period is given by V. Then, the waiting time $\hat{W}_n^E$ of the nth customer in this vacation model under the exhaustive service policy is identical to that of the original vacation model under the Bernoulli policy which corresponds to the sequence $\{X_n,\ n = 0, 1, \ldots\}$, i.e.,

$$\hat{W}_n^E = W_n^{B(p)}, \qquad n = 0, 1, \ldots. \tag{5.32}$$

Proof: The proof proceeds by induction. By Assumption (A3), we have

$$\hat{W}_0^E = W_0^{B(p)} = 0 \tag{5.33}$$

and so (5.32) trivially holds for $n = 0$. We take as induction hypothesis that

$$\hat{W}_n^E = W_n^{B(p)} =: W_n, \qquad n = 0, \ldots, k \tag{5.34}$$

and show that

$$\hat{W}_{k+1}^E = W_{k+1}^{B(p)}. \tag{5.35}$$

With the notation

$$j_k^{B(p)} = \min\{j \geq 0 : W_k + B_k - A_{k+1} + jV \geq 0\}, \tag{5.36}$$

we see that the waiting time $W_{k+1}^{B(p)}$ of the $(k+1)$st customer is given by

$$W_{k+1}^{B(p)} = W_k + B_k - A_{k+1} + X_k 1\{W_k + B_k - A_{k+1} \geq 0\}V + j_k^{B(p)}V. \tag{5.37}$$

From the definition of the exhaustive policy and (5.31), we see that $\hat{W}_{k+1}^E$ is given by

$$\hat{W}_{k+1}^E = W_k + B_k + X_k V - A_{k+1} + j_k^E V, \tag{5.38}$$

where

$$j_k^E = \min\{j \geq 0 : W_k + B_k + X_k V - A_{k+1} + jV \geq 0\}. \tag{5.39}$$

Thus to obtain (5.35), we are to show that

$$j_k^E + X_k = j_k^B + X_k 1\{W_k + B_k - A_{k+1} \geq 0\}. \tag{5.40}$$

Let us consider the case $W_k + B_k - A_{k+1} \geq 0$ first. In this case, we also have $W_k + B_k + X_k V - A_{k+1} \geq 0$ since $X_k \geq 0$, and so from (5.36) and (5.39) we have

$j_k^B = j_k^E = 0$ and therefore (5.40) immediately follows. For the case $W_k + B_k - A_{k+1} < 0$, we have to show that

$$X_k + j_k^E = j_k^B. \tag{5.41}$$

But, this is apparent from (5.36) and (5.39) upon examining the cases $X_k = 0$ and $X_k = 1$ separately. □

Remark: Lemma 5.3 above can be generalized to the case where the vacation periods are not necessarily deterministic. This result, which actually is not needed in this chapter, was used in Section 2.4.2 and is stated in Lemma 5.4 below.

Lemma 5.4 *Consider the vacation model under the Bernoulli policy and the modified vacation model under the exhaustive policy in Lemma 5.3, except that the vacation periods* $\{V_n,\ n = 0, 1, \ldots\}$ *are not necessarily deterministic. Then, the stochastic equalities*

$$\hat{W}_n^E =_{\text{st}} W_n^{B(p)}, \qquad n = 0, 1, \ldots \tag{5.42}$$

hold.

Proof: Consider a random sequence $\{V_n',\ n = 0, 1, \ldots\}$ which is probabilistically equivalent to $\{V_n,\ n = 0, 1, \ldots\}$. We redefine the vacation models as follows. For the Bernoulli model, the lengths of vacations which are started when the queue is not empty are chosen from the sequence $\{V_n,\ n = 0, 1, \ldots\}$, while the vacations which are started when the queue is empty are chosen from $\{V_n',\ n = 0, 1, \ldots\}$. Similarly for the modified exhaustive model, the lengths of vacations used in the modified service times (cf. (5.31)) are chosen from $\{V_n,\ n = 0, 1, \ldots\}$ while others are chosen from $\{V_n',\ n = 0, 1, \ldots\}$. Obviously, redefining the vacation models this way does not change their probabilistic structure. Moreover, we can follow the proof of Lemma 5.3 with the vacation lengths being chosen from either $\{V_n,\ n = 0, 1, \ldots\}$ or $\{V_n',\ n = 0, 1, \ldots\}$ as appropriate to show that (5.32) indeed holds for the redefined vacation models. This fact readily yields the desired result (5.42). □

Lemma 5.5 *Let $\{Z_n,\ n = 0, 1, \ldots\}$ and Z be nonnegative random variables and suppose that $Z_n \xrightarrow{\mathcal{D}} Z$. If there exists an integrable nonnegative random variable Y such that*

$$Z_n \leq_{\text{icx}} Y, \qquad n = 0, 1, \ldots, \tag{5.43}$$

then Z is also integrable and moreover

$$\lim_{n\to\infty} E[Z_n] = E[Z]. \tag{5.44}$$

Proof: Taking $f(x) = x$ in the definition of the ordering $\leq_{\text{icx}}$, we obtain from (5.43) that

$$E[Z_n] \leq E[Y], \qquad n = 0, 1, \ldots. \tag{5.45}$$

Since $Z_n \xrightarrow{\mathcal{D}} Z$, it follows by [6, Theorem 5.3] that

$$E[Z] \leq \liminf_{n\to\infty} E[Z_n]. \tag{5.46}$$

Therefore, combining these inequalities yields

$$E[Z] \leq E[Y] < \infty, \tag{5.47}$$

and so Z is integrable.

For each $a \geq 0$, define a mapping $f_a : I\!R_+ \to I\!R_+$ by

$$f_a(x) = \begin{cases} x & \text{if } 0 \leq x < a \\ a & \text{if } x \geq a. \end{cases} \tag{5.48}$$

Since f_a is bounded and continuous, the definition of weak convergence then asserts that

$$\lim_{n\to\infty} E[f_a(Z_n)] = E[f_a(Z)]. \tag{5.49}$$

Combining this fact with

$$E[Z_n] = E[f_a(Z_n)] + E[(Z_n - a)^+] \tag{5.50}$$

and

$$E[Z] = E[f_a(Z)] + E[(Z - a)^+], \tag{5.51}$$

we obtain

$$\limsup_{n\to\infty}\left|E[Z_n] - E[Z]\right| \leq \sup_n\left\{E[(Z_n - a)^+] + E[(Z - a)^+]\right\}. \tag{5.52}$$

But, by (F1), we have from (5.43) that

$$E[(Z_n - a)^+] \leq E[(Y - a)^+], \qquad n = 0, 1, \ldots, \tag{5.53}$$

and so

$$\limsup_{n\to\infty}\left|E[Z_n] - E[Z]\right| \leq E[(Y - a)^+] + E[(Z - a)^+]. \tag{5.54}$$

The proof of (5.44) is now completed by letting $a \to \infty$ in (5.54) and using the fact that both Z and Y are integrable. □

Proof of Theorem 5.2: Only the case where $E[W^{B(p)}] < \infty$ needs to be considered since $E[W^{B(p)}] = \infty$ trivially implies (5.30). By Proposition 1.3.2 of [77], (5.30) directly follows if we can show that the limits

$$\lim_{n\to\infty} E[W_n^{B(p)})] = E[W^{B(p)}] \quad \text{and} \quad \lim_{n\to\infty} E[W_n^{L(m)})] = E[W^{L(m)}] \tag{5.55}$$

hold true and finite. To that end, we shall make use of Lemmas 5.3 and 5.5.

Doshi [21] established a sample path comparison between a vacation model under the exhaustive policy and its corresponding $GI/GI/1$ queue (the same system with no vacations). Using this result, $\hat{W}_n^E$ can be decomposed into the waiting time $\hat{W}_n$ of the nth customer in the corresponding $GI/GI/1$ queue and a term Y_n that corresponds to the vacations. The term Y_n is always smaller than the last vacation period taken, and so

$$\hat{W}_n^E \leq \hat{W}_n + V, \qquad n = 0, 1, \ldots. \tag{5.56}$$

Note that under the stability assumption, the corresponding $GI/GI/1$ queue is also stable, i.e., $\hat{W}_n \xrightarrow{\mathcal{D}} \hat{W}$ for some nondefective $\hat{W}$. Moreover, since $E[W^{B(p)}]$ is finite, so is $E[\hat{W}]$. Theorem 5.1.1 of [77] thus yields

$$\hat{W}_n \leq_{\text{icx}} \hat{W}, \qquad n = 0, 1, \ldots. \tag{5.57}$$

This inequality, together with (5.32) and (5.56) implies that

$$W_n^{B(p)} \leq_{\text{icx}} \hat{W} + V, \qquad n = 0, 1, \ldots. \tag{5.58}$$

Subsequently, a use of Theorem 5.1 yields

$$W_n^{L(m)} \leq_{\text{icx}} \hat{W} + V, \qquad n = 0, 1, \ldots . \tag{5.59}$$

By Lemma 5.5, the icx–ordering bounds (5.58) and (5.59), together with (5.28) readily imply (5.55), and so the proof of Theorem 5.2 is complete. □

PART II

POLLING SYSTEMS

CHAPTER 6

POLLING SYSTEMS UNDER THE BERNOULLI POLICIES

6.1 Introduction

In this chapter and the next, we study polling systems under the Bernoulli and limited policies, respectively. As mentioned earlier in the thesis, we shall see that results obtained for vacation models often can be either extended to or used in approximating polling systems. In this chapter, we establish exact results for polling systems under the Bernoulli policy by first studying a queue in isolation; in the next chapter, we use the interpolation approximations for limited vacation models to devise an approximation scheme for polling systems under the limited policy.

In this chapter, we establish the pseudo–conservation law for the Bernoulli policy with the help of a work decomposition result recently established by Boxma and Groenendijk [8,9,10]. The pseudo–conservation law basically equates the weighted sum of mean waiting times associated with the individual queues to a simple expression which depends only on the first and second moments of the interarrival, service and switchover times distributions. This law extends to polling systems the classical conservation law for $M/GI/1$ queues established by Kleinrock [49],

The decomposition result of Boxma and Groenendijk states that the mean amount of work in a general cyclic–service queueing system can be decomposed into two components: one component is independent of the service policy while the other is the sum of the mean amounts of work the server leaves behind at various queues when it switches from one queue to the next. The second component depends on the service policy and once it is determined, the pseudo–conservation law for that service policy is established. To obtain this component for the Bernoulli policy, we establish other exact results for the Bernoulli policy which are of interest in their own right.

Pseudo–conservation laws have been established for other service policies. Ferguson and Aminetzah [30] and Watson [88] independently established this law for the exhaustive and gated policies; in [88], Watson also presented the law for the limited–to–one policy. In [27] and [28], Everitt found explicit forms of the pseudo–conservation law for variants of the limited policies. For these policies, however, the pseudo–conservation law still contains some unknowns which were identified as the second factorial moments of the numbers of customers served in one service period at various queues. In [27], an approximation to these unknowns is investigated. Fuhrmann and Wang [34,35] established bounds for the pseudo–conservation law for the limited policies which they then used as the basis for an approximation.

This chapter is organized as follows. The model is described in Section 2. In Section 3, we derive the Laplace–Stieltjes transform (LST) of the limiting waiting time at a particular queue in terms of the limiting probability generating function (PGF) of the number of customers in that queue at the beginning of a service period. In this derivation, we make use of an expression for the average number of customers served from a queue in a service period. The derivation of this expression is deferred until Section 5. The result of Section 3 is used in Section 4 to establish the pseudo–conservation law for the Bernoulli policy. In Section 6, we analyze a homogeneous cyclic–service queue under a Bernoulli policy. In this case, the pseudo–conservation law readily provides a closed–form formula for the mean waiting time. In Section 7, we conclude the chapter by studying an approximation for the nonhomogeneous queue which is based on the pseudo–conservation law.

Throughout the chapter we use the following convention. For any $\mathbb{R}_+$–valued random variable X, we use X^* to denote the LST of its distribution. If X is integer–valued, we use $\tilde{X}$ to denote its PGF. In both cases, we use $\bar{X}$, $\overline{X^2}$, and σ_X^2 to denote the mean, second moment, and variance of X, respectively.

6.2 Model and Notation

We consider an $M/GI/1$ cyclic–service system consisting of N infinite capacity queues which are denoted by $Q_1, \ldots, Q_N$. A Bernoulli service policy, which is parametrized by a vector of probabilities $(p_1, \ldots, p_N)$ where $0 \leq p_j \leq 1$, $j = 1, \ldots, N$, is described as follows. At the beginning of each visit to a queue, the server always serves a customer if the queue is not empty. At the completion of every service given to a customer at Q_j, if the queue is not empty, the server *flips a biased coin.* The outcome of the flip is '1' with probability p_j or '0' with probability $1 - p_j$. If the server flips '1', the next available customer in the queue is served. Otherwise, the server goes to the next queue down the line, i.e., $Q_{j+1\,(\mathrm{mod}\,N)}$. If Q_j is empty at the service completion, the server goes to $Q_{j+1\,(\mathrm{mod}\,N)}$ with probability one. Within a queue, the service discipline is FCFS. The server takes a random amount of time, the so–called *switchover time*, to go from Q_j to $Q_{j+1\,(\mathrm{mod}\,N)}$.

The random variables governing the system are listed below, with $j = 1, \ldots, N$ and $n = 0, 1, \ldots$

A_j^{n+1} = the interarrival time between the nth and the $(n+1)$st customer in Q_j, with the convention that the customer with index 0 arrives at time $t = 0$. The process $\{A_j^n,\ n = 0, 1, \ldots\}$ is assumed Poisson with parameter λ_j.

B_j^n = the service time of the nth customer in Q_j. The random variables $\{B_j^n,\ n = 0, 1, \ldots\}$ are i.i.d. with a general common distribution. Throughout, B_j denotes a generic service time at Q_j.

U_j^n = the nth coin flip in Q_j. The random variables $\{U_j^n,\ n = 0, 1, \ldots\}$ are i.i.d., $\{0, 1\}$–valued with $P[U_j^n = 1] = 1 - P[U_j^n = 0] = p_j$ for all $n = 0, 1, \ldots$.

V_j^n = the nth switchover time from Q_j to $Q_{j+1\,(\mathrm{mod}\,N)}$. The random variables $\{V_j^n,\ n = 0, 1, \ldots\}$ are i.i.d. with a general common distribution. Throughout, V_j denotes a generic switchover time from Q_j.

We assume that all the processes described above within a queue as well as among all the queues in the network are mutually independent. Also define $\rho_j = \lambda_j \bar{B}_j$, $j = 1, \ldots, N$, $\rho_T = \sum_{j=1}^N \rho_j$, and set $V_T = \sum_{j=1}^N V_j$.

Throughout this chapter, we refer to an instant the server arrives at Q_j from Q_{j-1} as a *polling instant* of Q_j and the period starting from a polling instant for Q_j and ending with the server's departure as a *server's visit period* to Q_j or simply a *service period* of Q_j.

6.3 The Waiting Time at a Queue in Isolation

In this section, we focus our attention to one particular queue in a cyclic–service system under a Bernoulli policy. We suppress the index of the queue for notational simplicity and define

K^m = the number of customers at the beginning of the mth service period, $m = 0, 1, \ldots$;

Z^m = the number of customers served during the mth service period, $m = 0, 1, \ldots$;

W^r = the waiting time of the rth customer, $r = 0, 1, \ldots$.

We assume that K^m, Z^m, and W^r converge in distribution to some non–defective random variables K, Z, and W, respectively. In fact, it can be shown that each queue is stable if

$$\lambda_j < \frac{1 - \rho_j + \rho_T}{(1 - p_j)\bar{V}_T + \bar{B}_j}, \qquad j = 1, \ldots, N. \tag{6.1}$$

We have the following result which expresses the LST of the limiting waiting time in terms of the PGF of the limiting number of customers at the polling instant.

Theorem 6.1 *Consider a particular queue in a stable $M/GI/1$ cyclic–service queueing system under a Bernoulli policy. Let p, $0 \le p \le 1$, be the Bernoulli probability assigned to that queue. Then, the limiting LST of the waiting time of an arbitrary customer in that queue is given by*

$$W^*(s) = \frac{\lambda(1-p)}{1 - \tilde{K}(\Phi(0,p))} \cdot \frac{\tilde{K}(1 - s/\lambda) - \tilde{K}(\Phi(0,p))}{\lambda - s - p\lambda B^*(s)}, \qquad |\lambda - s| < \lambda, \tag{6.2}$$

where $\tilde{K}$ is the limiting PGF of the number of customers at the beginning of a service period, and Φ satisfies the relation

$$\Phi(s,z) = zB^*(s + \lambda - \lambda\Phi(s,z)), \quad \Re(s) \ge 0, |z| \le 1. \tag{6.3}$$

The value of $W^*(s)$ for $p = 1$ is understood to be the limit of the right-hand side of (6.2) as $p \to 1$.

Proof. In what follows, an epoch is either a polling instant or a service completion. Describe the queue by a sequence of pairs of random variables $\{(X^n, J^n),\ n = 0, 1, \ldots\}$ with the following interpretation: X^n denotes the number of customers in the queue at the nth epoch, while $J^n = 1$ (resp. $J^n = 0$) if the epoch marks the end of a service (resp. a polling instant). Let $r(n)$ denote the index of the customer whose service is completed at or immediately before the nth epoch. Similarly, define $c(n)$ as the index of the polling instant which occurs at or immediately before the nth epoch. Both $r(n)$ and $c(n)$ are random variables that go to infinity with n. Also, let R^r be the number of arrivals that occur during the service of the rth customer. The evolution of the sequence $\{(X^n, J^n),\ n = 0, 1, \ldots\}$ is then given by

$$(X^{n+1}, J^{n+1}) = \begin{cases} (X^n - 1 + R^{r(n+1)}, 1) & \text{if } X^n > 0, J^n = 0 \\ & \text{or } X^n > 0,\ J^n = 1,\ U^{r(n)} = 1 \\ (K^{c(n+1)}, 0) & \text{otherwise.} \end{cases} \tag{6.4}$$

We are interested in Q^r, the number of customers in the queue immediately following the departure of the rth customer, and its limiting random variable Q. For $n = 0, 1, \ldots,$ define

$$\Psi^n(z) = E[z^{X^n} | J^n = 1] \quad \text{and} \quad \Xi^n(z) = E[z^{X^n} | J^n = 0], \qquad |z| \leq 1 \tag{6.5}$$

so that $\tilde{Q}^{r(n)}(z) = \Psi^n(z)$, $\tilde{K}^{c(n)}(z) = \Xi^n(z)$, and

$$\lim_{n\to\infty} \Psi^n(z) = \tilde{Q}(z), \qquad \text{and} \qquad \lim_{n\to\infty} \Xi^n(z) = \tilde{K}(z), \qquad |z| \leq 1. \tag{6.6}$$

From (6.4), we see that

$$\begin{aligned} \Psi^{n+1}(z) &= E[z^{X^{n+1}} | J^{n+1} = 1] \\ &= E[z^{X^n - 1 + R^{r(n+1)}} | J^{n+1} = 1] \\ &= z^{-1} B^*(\lambda - \lambda z)\, E[z^{X^n} | J^{n+1} = 1]. \end{aligned} \tag{6.7}$$

We also observe from (6.4) that

$$[J^{n+1} = 1] = \mathcal{A}^n \cup \mathcal{B}^n, \qquad n = 0, 1, \ldots, \tag{6.8}$$

where the sets $\mathcal{A}^n$ and $\mathcal{B}^n$ are defined by

$$\mathcal{A}^n = [X^n > 0, J^n = 0] \qquad \text{and} \qquad \mathcal{B}^n = [X^n > 0, J^n = 1, U^{r(n)} = 1], \tag{6.9}$$

respectively. Since $\mathcal{A}^n$ and $\mathcal{B}^n$ are disjoint, we have from (6.8) that

$$E[z^{X^n} | J^{n+1} = 1] = \frac{E[z^{X^n} I_{\mathcal{A}^n}] + E[z^{X^n} I_{\mathcal{B}^n}]}{P[J^{n+1} = 1]}, \tag{6.10}$$

where I_A is the indicator function of set A.

From the definition of Ξ^n and Ψ^n, we have

$$E[z^{X^n} I_{\mathcal{A}^n}] = P[J^n = 0]\big(\Xi^n(z) - \Xi^n(0)\big) \tag{6.11}$$

and

$$E[z^{X^n} I_{\mathcal{B}^n}] = p\,P[J^n = 1]\big(\Psi^n(z) - \Psi^n(0)\big) \tag{6.12}$$

which we can substitute into (6.10) and then into (6.7) to obtain

$$\Psi^{n+1}(z) = \frac{z^{-1} B^*(\lambda - \lambda z)}{P[J^{n+1} = 1]} \Big[P[J^n = 0]\big(\Xi^n(z) - \Xi^n(0)\big) + p\,P[J^n = 1]\big(\Psi^n(z) - \Psi^n(0)\big)\Big]. \tag{6.13}$$

Taking the limit of the above as $n \to \infty$ and rearranging terms, we finally obtain

$$\tilde{Q}(z) = \frac{P[J = 0]}{P[J = 1]} \, \frac{\tilde{K}(z) - \tilde{K}(0) - p\frac{P[J=1]}{P[J=0]}\tilde{Q}(0)}{z - pB^*(\lambda - \lambda z)} B^*(\lambda - \lambda z). \tag{6.14}$$

Note that $\frac{P[J=1]}{P[J=0]}$ is nothing but the average number of customers served in one service period $\bar{Z}$, so that

$$\tilde{Q}(z) = \frac{1}{\bar{Z}} \, \frac{\tilde{K}(z) - \tilde{K}(0) - p\tilde{Q}(0)\bar{Z}}{z - pB^*(\lambda - \lambda z)} B^*(\lambda - \lambda z). \tag{6.15}$$

Letting $z = 1$ in (6.15) and solving for $\tilde{Q}(0)$, we obtain

$$\tilde{Q}(0) = \frac{1}{p\bar{Z}} \big[1 - (1 - p)\bar{Z} - \tilde{K}(0)\big], \tag{6.16}$$

which can be substituted back into (6.15) to get

$$\tilde{Q}(z) = \frac{1}{\bar{Z}} \frac{\tilde{K}(z) - 1 + (1-p)\bar{Z}}{z - pB^*(\lambda - \lambda z)} B^*(\lambda - \lambda z). \tag{6.17}$$

Since the service is given on a FCFS basis, the customers present when an arbitrary customer leaves the queue are those that arrive during its waiting and service time, whence

$$\tilde{Q}(z) = W^*(\lambda - \lambda z)\, B^*(\lambda - \lambda z). \tag{6.18}$$

Solving for W^* from (6.17) and (6.18) and letting $s = \lambda - \lambda z$, we have

$$W^*(s) = \frac{\lambda}{\bar{Z}} \frac{\tilde{K}(1 - s/\lambda) - 1 + (1-p)\bar{Z}}{\lambda - s - p\lambda B^*(s)}, \quad |\lambda - s| < \lambda. \tag{6.19}$$

We show in Section 5 that $\bar{Z}$ can be expressed in terms of $\tilde{K}$ as

$$\bar{Z} = \frac{1 - \tilde{K}(\Phi(0,p))}{1-p}. \tag{6.20}$$

Hence, substituting (6.20) into (6.19) yields (6.2), and the proof is complete. □

Note that by letting $p = 0$ in (6.2), we obtain the waiting time for the limited-to-one policy as

$$W^*(s) = \frac{\lambda}{1 - \tilde{K}(0)} \frac{\tilde{K}(1 - s/\lambda) - \tilde{K}(0)}{\lambda - s}, \tag{6.21}$$

which is in agreement with previously established results [88, p. 526]. Note also that Φ as defined in (6.3) is the joint LST–PGF of the length of a busy period and the number of customers served during that busy period in an $M/GI/1$ system. It can be easily shown that $\frac{d}{dp}\Phi(0,p)$ evaluated at $p = 1$ is equal to $1/(1-\rho)$. Using this fact, we can let $p = 1$ in (6.2) and perform l'Hôspital's rule on the first factor of the right–hand side to obtain

$$W^*(s) = \frac{\lambda(1-\rho)}{\bar{K}} \frac{\tilde{K}(1 - s/\lambda) - 1}{\lambda(1 - B^*(s)) - s}, \tag{6.22}$$

which is the waiting time LST for the exhaustive policy.

We readily obtain the mean waiting time in terms of the mean number of customers at the beginning of the busy period by taking the first derivative of both sides of (6.2) with respect to s and letting $s = 0$.

Corollary 6.1 *The mean waiting time of an arbitrary customer in the queue in Theorem 6.1 is given by*

$$\bar{W} = \frac{\bar{K}(1-p) - (1-\tilde{K}(\Phi(0,p)))(1-p\rho)}{\lambda(1-\tilde{K}(\Phi(0,p)))(1-p)}, \qquad (6.23)$$

where the value evaluated at $p = 1$ is understood to be the limit as $p \to 1$.

Again, (6.23) evaluated at the extreme values of p agrees with the formulae available in the literature [88, p. 527]. Note that for $p = 1$, we have to use l'Hôspital's rule twice.

6.4 Pseudo–conservation law for the Bernoulli policy

In this section, we shall derive the pseudo–conservation law for the Bernoulli policy using the result in the previous section and the work decomposition result established by Boxma and Groenendijk [9,10]. The work decomposition result states that for an $M/GI/1$ cyclic–service under any service policy, we have

$$\sum_{j=1}^{N} \rho_j \bar{W}_j = C + \sum_{j=1}^{N} \bar{L}_j \bar{B}_j, \qquad (6.24)$$

where $\bar{L}_j$ is the mean number of remaining customers in Q_j when the server leaves that queue, i.e., the number of customers at the end of a service period for Q_j. In general, $\bar{L}_j$ is dependent on the service policy. On the other hand, C is independent of the service policy and is given by

$$C = \frac{\rho_T}{2(1-\rho_T)} \sum_{j=1}^{N} \lambda_j \overline{B_j^2} + \frac{\rho_T \sigma_{V_T}^2}{2\bar{V}_T} + \frac{\bar{V}_T}{2(1-\rho_T)}\Big(\rho_T - \sum_{j=1}^{N} \rho_j^2\Big). \qquad (6.25)$$

To obtain the pseudo–conservation law, we are going to express $\bar{L}_j$ for $j = 1, \ldots, N$ as a function of $\bar{W}_j$. To this end, we first note that

$$\bar{L}_j = \bar{K}_j + \bar{Z}_j(\rho_j - 1), \qquad j = 1, \ldots, N \qquad (6.26)$$

which simply says that the mean number of customers at the end of a service period in Q_j is the mean number at the beginning of the service period plus the mean number of arrivals minus the mean number of departures in the same service period.

For an $M/GI/1$ cyclic–service queue, $\bar{Z}_j$ is independent of the service policy and is given by

$$\bar{Z}_j = \frac{\lambda_j \bar{V}_T}{1-\hat{\rho}}, \qquad j = 1, \ldots, N. \tag{6.27}$$

Using (6.20), we can rewrite (6.23) as

$$\bar{W}_j = \frac{1}{1-p_j}\left(\frac{\bar{K}_j}{\lambda_j \bar{Z}_j} - \frac{1-p_j\rho_j}{\lambda_j}\right), \qquad j = 1, \ldots, N. \tag{6.28}$$

Combining (6.24)–(6.28) and collecting the terms $\bar{W}_j$ to the left–hand side, the pseudo–conservation law is obtained in the following form.

Theorem 6.2 *For a stable $M/GI/1$ cyclic–service queueing system under a Bernoulli service policy with parameter $(p_1, \ldots, p_N)$, $0 \le p_j \le 1$, $j = 1, \ldots, N$, the relation*

$$\sum_{j=1}^{N} \rho_j \left[1 - \frac{\lambda_j \bar{V}_T}{1-\rho_T}(1-p_j)\right] \bar{W}_j = C + \frac{\bar{V}_T}{(1-\rho_T)} \sum_{j=1}^{N} \rho_j^2 (1-p_j) \tag{6.29}$$

holds.

Remark 6.1: Fuhrmann and Wang [35] showed that for a polling system under the limited policy with parameters $(m_1, \ldots, m_N)$, the mean waiting times $\bar{W}_j^L$, $j = 1, \ldots, N$, satisfy the inequality

$$\sum_{j=1}^{N} \rho_j \left[1 - \frac{\lambda_j \bar{V}_T}{m_j(1-\rho_T)}\right] \bar{W}_j^L \le C + \frac{\bar{V}_T}{(1-\rho_T)} \sum_{j=1}^{N} \frac{\rho_j^2}{m_j}. \tag{6.30}$$

If we choose

$$m_j = \frac{1}{1-p_j}, \qquad j = 1, \ldots, N, \tag{6.31}$$

then both sides of equality (6.29) are identical to their respective side of inequality (6.30), and so we have

$$\sum_{j=1}^{N} \rho_j \left[1 - \frac{\lambda_j \bar{V}_T}{m_j(1-\rho_T)}\right] \bar{W}_j^L \le \sum_{j=1}^{N} \rho_j \left[1 - \frac{\lambda_j \bar{V}_T}{1-\rho_T}(1-p_j)\right] \bar{W}_j^B, \tag{6.32}$$

where $\bar{W}_j^B$'s denote the mean waiting times under the Bernoulli policy.

6.5 The Service Period

Equation (6.20) relates the mean number of customers served in a service period with the number of customers at the beginning of the service period. As promised, we now present the derivation of this relation. As it will become apparent, it is more natural to include the length of the service period in the analysis as well.

For $m = 0, 1, \ldots,$ we define S^m as the length of the mth service period and denote the joint LST–PGF of S^m and Z^m by

$$F^m(s,z) = E[e^{-sS^m} z^{Z^m}], \quad \Re(s) \geq 0, \ |z| \leq 1. \tag{6.33}$$

The following theorem provides an expression for this quantity.

Theorem 6.3 *For the queue in Theorem 6.1, the joint LST–PGF of S^m and Z^m is given by*

$$F^m(s,z) = (1 - \Gamma(s,z))\tilde{K}^m(\Phi(s,pz)) + \Gamma(s,z), \tag{6.34}$$

where

$$\Gamma(s,z) = \frac{(1-p)zB^*(s)}{1-pzB^*(s)}, \qquad \Re(s) \geq 0, |z| \leq 1 \tag{6.35}$$

and $\Phi(s,z)$ satisfies (6.3).

Proof. Without loss of generality, we shall prove the theorem for $m = 0$. First, we define ξ and τ as the length of the first service period and the number of customers served in the service period, respectively, when the queue is saturated, i.e.

$$\tau = \min\{k \geq 1 : U^{k-1} = 0\} \qquad \text{and} \qquad \xi = B^0 + \cdots + B^{\tau-1}. \tag{6.36}$$

We can easily verify that $E[e^{-s\xi}z^\tau] = \Gamma(s,z)$ as defined in (6.35). Next, define β and η as the length of the first service period and the number of customers served in the service period, respectively, when K^0 customers are present at the start of the service period, and the service policy is exhaustive. If $K^0 = 1$, then β and η are just the length of a busy period and the number of customers served in the busy period for a regular $M/GI/1$ system. It is a well–known fact that if $\rho < 1$, then $\Phi(s,z) := E[e^{-s\beta}z^\eta | K^0 = 1]$ satisfies (6.3). Moreover, using Takács' argument and the fact that K^0 is independent of the subsequent interarrival and service times, we can also show that

$$E[e^{-s\beta}z^{\eta}] = \tilde{K}^0(\Phi(s,z)). \tag{6.37}$$

Note that the number of customers served in a service period is determined either by τ, if the server flips 0 before the queue is empty, or by η, if the queue is empty before the server flips 0, and so

$$Z^0 = \min(\tau,\eta) \qquad \text{and} \qquad S^0 = B^0 + \cdots + B^{Z^0-1}. \tag{6.38}$$

Defining the set $A = [\eta < \tau]$, we can write

$$F^0(s,z) = E[e^{-sS^0}z^{Z^0}I_A] + E[e^{-sS^0}z^{Z^0}I_{A^c}]. \tag{6.39}$$

On the set A, $Z^0 = \eta$ and $S^0 = \beta$, so that

$$\begin{aligned} E[e^{-sS^0}z^{Z^0}I_A] &= E[e^{-s\beta}z^{\eta}I_A] \\ &= E[e^{-s\beta}z^{\eta}E[I_A|\beta,\eta]]. \end{aligned} \tag{6.40}$$

But, under the enforced independence assumptions, we have

$$E[I_A|\beta,\eta] = P[A|\beta,\eta] = p^{\eta} \tag{6.41}$$

and it therefore follows from (6.37) that

$$\begin{aligned} E[e^{-sS^0}z^{Z^0}I_A] &= E[e^{-s\beta}z^{\eta}p^{\eta}] \\ &= \tilde{K}^0(\Phi(s,pz)). \end{aligned} \tag{6.42}$$

On the set A^c, we have $Z^0 = \tau$ and $S^0 = \xi$, so that

$$\begin{aligned} E[e^{-sS^0}z^{Z^0}I_{A^c}] &= E[e^{-s\xi}z^{\tau}I_{A^c}] \\ &= E[e^{-s\xi}z^{\tau}] - E[e^{-s\xi}z^{\tau}I_A]. \end{aligned} \tag{6.43}$$

Let us now define τ' and ξ' as

$$\tau' = \min\{k \geq 1 : U^{\eta+k-1} = 0\} \qquad \text{and} \qquad \xi' = B^{\eta} + \cdots + B^{\eta+\tau'-1}. \tag{6.44}$$

The random variable η is a stopping time with respect to the i.i.d. sequence $\{(A^{n+1}, B^n, U^n),\ n = 0,1,\ldots\}$. We have $\eta < \infty$ a.s. since $\rho < 1$. Furthermore, it can be shown that $\{(A^{n+1}, B^n, U^n),\ n = 0,1,\ldots,\eta\}$ is independent of $\{(A^{\eta+n+1}, B^{\eta+n}, U^{\eta+n}),\ n = 0,1,\ldots\}$ which itself is i.i.d. In particular,

$\{(A^{\eta+n+1}, B^{\eta+n}, U^{\eta+n}),\ n=0,1,\ldots\}$ is independent of η and β. Notice also that I_A is determined once we know the value of η and U^n, $n=0,1,\ldots,\eta$, i.e., I_A is a function of $\{(A^{n+1}, B^n, U^n),\ n=0,1,\ldots,\eta\}$. So, $\{(A^{\eta+n+1}, B^{\eta+n}, U^{\eta+n}),\ n=0,1,\ldots\}$ is independent of I_A as well. The random variables τ' and ξ' defined above are functions of $\{(A^{\eta+n+1}, B^{\eta+n}, U^{\eta+n}),\ n=0,1,\ldots\}$ and so they are independent of η, β and I_A. Moreover, by the i.i.d. nature of $\{(A^{\eta+n+1}, B^{\eta+n}, U^{\eta+n}),\ n=0,1,\ldots\}$, we have

$$E[e^{-s\xi'}z^{\tau'}] = \Gamma(s,z). \tag{6.45}$$

On the set A, we have $\tau = \eta + \tau'$ and $\xi = \beta + \xi'$, and consequently, by the previous remarks, we find that

$$\begin{aligned} E[e^{-s\xi}z^{\tau}I_A] &= E[e^{-s\beta}e^{-s\xi'}z^{\eta}z^{\tau'}I_A] \\ &= E[e^{-s\beta}z^{\eta}I_A]\,E[e^{-s\xi'}z^{\tau'}] \\ &= \tilde{K}^0(\Phi(s,pz))\Gamma(s,z). \end{aligned} \tag{6.46}$$

Substituting (6.46) into (6.43), and then (6.42)-(6.43) into (6.39) completes the proof.

□

If the system is stable, there exist F and $\tilde{K}$ such that as $m \to \infty$, $F^m(s,z)$ goes to $F(s,z)$ and $\tilde{K}^m(z)$ goes to $\tilde{K}(z)$ for all $\Re(s) \geq 0$ and $|z| \leq 1$. As a result, we easily obtain the following corollary.

Corollary 6.2 *For the queue in Theorem 6.1, the relation*

$$F(s,z) = (1-\Gamma(s,z))\tilde{K}(\Phi(s,pz)) + \Gamma(s,z), \qquad \Re(s) \geq 0, |z| \leq 1 \tag{6.47}$$

holds with Γ *and* Φ *defined by (6.35) and (6.3), respectively.*

By letting $s=0$ in (6.47), we obtain the number of customers served in a service period.

Corollary 6.3 *If the system is stable with* Z^m *going to* Z *in distribution, then*

$$\tilde{Z}(z) = (1-\Gamma(0,z))\tilde{K}(\Phi(0,pz)) + \Gamma(0,z). \tag{6.48}$$

Taking the first derivative of both sides of (6.48) and letting $z = 1$ yield

$$\bar{Z} = \frac{1 - \tilde{K}(\Phi(0, p))}{1 - p}, \tag{6.49}$$

which is the result we wanted to show in this section.

We note that a result that expresses the joint LST–PGF of the length of a service period and the number of customers present at the end of that service period in terms of the PGF of the number of customers at the beginning of the service period can also be obtained using the method used in this section. This result has been derived by Ramaswamy and Servi in [66] using a different method.

6.6 The Exact Solution for the Homogeneous Case

If the system is *homogeneous* in the sense that the processes governing each queue (including the switchover times) are stochastically identical, then $\bar{W}_1 = \cdots = \bar{W}_N = \bar{W}$. Using (6.29) we can directly solve for $\bar{W}$ since $\bar{W}$ can be pulled out of the summation. In fact, the average waiting time taken over all the queues in the system can be solved under a much less stringent condition. If the system is such that $\lambda_1 = \cdots = \lambda_N = \lambda$, $\bar{B}_1 = \cdots = \bar{B}_N = \bar{B}$, and $p_1 = \cdots = p_N = p$, then the coefficients in front of $\bar{W}_j$'s are all identical and can be pulled out of the summation. If we define

$$\bar{W} = \frac{1}{N} \sum_{j=1}^{N} \bar{W}_j, \tag{6.50}$$

then again we can solve for $\bar{W}$ using (6.29). Notice that, under this condition, it is not clear that the individual $\bar{W}_j$'s will be identical. The following theorem states the result discussed above.

Theorem 6.4 *For a stable $M/GI/1$ cyclic-service queueing system under Bernoulli service policy where $\lambda_1 = \cdots = \lambda_N = \lambda$, $\bar{B}_1 = \cdots = \bar{B}_N = \bar{B}$, and $p_1 = \cdots = p_N = p$, $0 \leq p \leq 1$, the average waiting time defined in (6.50) is given by*

$$\bar{W} = \frac{1}{2[1 - \rho_T - \lambda \bar{V}_T (1 - p)]} \left[\lambda \sum_{j=1}^{N} \overline{B_j^2} + \frac{(1 - \rho_T)\sigma_{\bar{V}_T}^2}{\bar{V}_T} + \bar{V}_T (1 + \rho - 2\rho p) \right], \tag{6.51}$$

where $\rho = \lambda \bar{B}$, and $\hat{\rho} = N\rho$.

6.7 Approximation for the Nonhomogeneous Case

Following [11,26,28,34,35], we can use the pseudo–conservation law to approximate the individual waiting times for the nonhomogeneous case. Generally the approximate mean waiting time takes the form

$$\bar{W}_j \approx \gamma_j x, \quad j = 1, \ldots, N, \tag{6.52}$$

where γ_j is a known function of the system parameters, and x is an unknown which is solved using equation (6.29).

The application of the pseudo–conservation law to approximation methods was first studied by Everitt [26] for the gated and exhaustive policies and by Boxma and Meister [11] for the limited–to–one policy. Recently, Fuhrmann and Wang [34,35] and Everitt [28] extended this study to the general limited policies.

Fuhrmann and Wang [35] heuristically derived the approximations

$$\bar{W}_j \approx \frac{1 - \rho_j + (\rho_j/m_j)[1 + 1/(1-\rho_T)]}{1 - \frac{\lambda_j V_T}{(1-\rho_T)m_j}} x, \quad j = 1, \ldots, N, \tag{6.53}$$

for the limited service policy. For the Bernoulli policy, we can use the same heuristic arguments to arrive to the same approximations; the only difference is that $1/m_j$ is now replaced by $(1 - p_j)$. Hence,

$$\bar{W}_j \approx \frac{1 - \rho_j + (1-p_j)\rho_j[1 + 1/(1-\rho_T)]}{1 - \frac{\lambda_j V_T}{1-\rho_T}(1-p_j)} x, \quad j = 1, \ldots, N. \tag{6.54}$$

Numerical Results

In the following, we compare the mean waiting times obtained using the approximations above with simulation results. We consider four examples, the first two with small N ($N = 3$) and the last two with large N ($N = 10$). From these examples, we observe that the approximation tends to perform worse as the system becomes less and less homogeneous. This trend is expected since by its construction the approximation becomes exact when the system is homogeneous.

Example 1. Small N ($N = 3$); medium load ($\rho_T = 0.5$); homogeneous service times

(p_1, p_2, p_3)	Measure	Simulation	Approx.
$(\frac{2}{3}, 0, 0)$	$\bar{W}_1$	1.406	1.487
	$\bar{W}_2$	2.859	2.826
	$\bar{W}_3$	2.945	2.826
$(\frac{5}{6}, \frac{1}{2}, 0)$	$\bar{W}_1$	1.399	1.474
	$\bar{W}_2$	1.912	1.957
	$\bar{W}_3$	3.084	3.025
$(\frac{5}{6}, 0, \frac{2}{3})$	$\bar{W}_1$	1.474	1.508
	$\bar{W}_2$	3.283	3.096
	$\bar{W}_3$	1.667	1.720
$(0, \frac{2}{3}, \frac{2}{3})$	$\bar{W}_1$	2.481	2.342
	$\bar{W}_2$	1.825	1.843
	$\bar{W}_3$	1.865	1.843

Table 6.1: Example 1

and switchover times ($\bar{B}_i = 1.0$ exponential, $V_i = 0.25$ deterministic, $i = 1, 2, 3$); nonhomogeneous arrival rates ($\lambda_1 = 0.1$, $\lambda_2 = \lambda_3 = 0.2$).

Example 2. Small N ($N = 3$); medium load ($\rho_T = 0.6$); homogeneous switchover times ($V_i = 0.25$, deterministic, $i = 1, 2, 3$); nonhomogeneous service times and arrival rates ($\bar{B}_1 = 2.0$, $\bar{B}_2 = \bar{B}_3 = 1.0$ all exponential, $\lambda_1 = 0.1$, $\lambda_2 = \lambda_3 = 0.2$).

Example 3. Large N ($N = 10$); low load ($\rho_T = 0.4$); homogeneous service times, switchover times and arrival rates ($\bar{B}_i = 1.0$ exponential, $V_i = 0.25$ deterministic, $\lambda_i = 0.04$, $i = 1, \ldots, 10$).

Example 4. Large N ($N = 10$); medium load ($\rho_T = 0.594$); homogeneous switchover times ($V_i = 0.25$ deterministic, $i = 1, \ldots, 10$); nonhomogeneous service times and arrival rates ($\bar{B}_i = 1.0$, $i = 1, 2, 3$, $\bar{B}_i = 0.6$, $i = 4, \ldots, 10$, all exponential; $\lambda_i = 0.1$, $i = 1, 2, 3$, $\lambda_i = 0.07$, $i = 4, \ldots, 10$).

(p_1, p_2, p_3)	Measure	Simulation	Approx.
$(\frac{2}{3}, 0, 0)$	$\bar{W}_1$	2.017	2.494
	$\bar{W}_2$	2.985	5.430
	$\bar{W}_3$	8.979	5.430
$(\frac{5}{6}, \frac{1}{2}, 0)$	$\bar{W}_1$	1.884	2.355
	$\bar{W}_2$	2.424	3.522
	$\bar{W}_3$	9.641	5.972
$(\frac{5}{6}, 0, \frac{2}{3})$	$\bar{W}_1$	2.309	2.417
	$\bar{W}_2$	4.368	6.130
	$\bar{W}_3$	3.673	3.017
$(0, \frac{2}{3}, \frac{2}{3})$	$\bar{W}_1$	3.088	3.427
	$\bar{W}_2$	2.729	3.189
	$\bar{W}_3$	3.659	3.189

Table 6.2: Example 2

$(p_1, \ldots, p_{10})$	Measure	Simulation	Approx.
$(\frac{2}{3}, 0, 0, 0, 0, 0, 0, 0, 0, 0)$	$\bar{W}_1$	2.791	2.808
	$\bar{W}_{2-10}$	3.398	3.409
$(\frac{5}{6}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, 0, 0, 0, 0)$	$\bar{W}_1$	2.660	2.704
	$\bar{W}_{2-5}$	2.918	2.972
	$\bar{W}_{6-10}$	3.456	3.441
$(\frac{4}{5}, 0, 0, 0, 0, \frac{2}{3}, \frac{2}{3}, \frac{2}{3}, \frac{2}{3}, \frac{2}{3})$	$\bar{W}_1$	2.653	2.744
	$\bar{W}_{2-5}$	3.443	3.460
	$\bar{W}_{6-10}$	2.822	2.850

Table 6.3: Example 3

$(p_1, \ldots, p_{10})$	Measure	Simulation	Approx.
$(\frac{2}{3}, 0, 0, 0, 0, 0, 0, 0, 0, 0)$	$\bar{W}_1$	4.662	4.948
	$\bar{W}_{2-3}$	13.431	12.563
	$\bar{W}_{4-10}$	7.566	7.511
$(\frac{5}{6}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, 0, 0, 0, 0)$	$\bar{W}_1$	4.105	4.299
	$\bar{W}_{2-3}$	6.419	6.246
	$\bar{W}_{4-5}$	5.198	5.292
	$\bar{W}_{6-10}$	8.296	7.812
$(\frac{4}{5}, 0, 0, 0, 0, \frac{2}{3}, \frac{2}{3}, \frac{2}{3}, \frac{2}{3}, \frac{2}{3})$	$\bar{W}_1$	4.203	4.850
	$\bar{W}_{2-3}$	14.391	16.154
	$\bar{W}_{4-5}$	7.818	8.950
	$\bar{W}_{6-10}$	4.472	5.161

Table 6.4: Example 4

CHAPTER 7

POLLING SYSTEMS UNDER THE LIMITED POLICIES

7.1 Introduction

As mentioned earlier, the limited policies do not lend themselves to analysis by "traditional" methods used very successfully for the exhaustive and gated policies. Nor can they be analyzed using the method used in the preceding chapter since the limited policies lack the memoryless property enjoyed by the Bernoulli policies. This property was crucial in the derivation of Theorem 6.1 which leads to the relationship (6.28) used in establishing the pseudo–conservation law. In this chapter, we propose and study an approximation algorithm for polling systems with Poisson arrivals under the limited policies. This algorithm is based on the heavy and light traffic interpolation approximation for the vacation models developed in Chapter 3.

The proposed algorithm takes advantage of the nonexhaustive nature of the limited policy and the fact that the interpolation approximation for vacation models under the limited policy is extremely accurate. We shall see that the nonexhaustiveness of the limited policy enables us to approximate the mean waiting time at a particular queue using two polling systems, each with one less queue than the original system. This reduction of dimensionality is repeated until we end up with vacation models, at which point we can use the interpolation approximation for the vacation models.

In the next section, we present the model and notation used throughout the chapter. For ease of exposition, we shall first consider polling systems with two queues. In Section 3, we propose an algorithm for these systems. In Section 4, we extend the algorithm to polling systems with an arbitrary number of queues. Finally in Section 5, we report some numerical results on the performance of the algorithm.

7.2 Model and Notation

We consider a polling system consisting of N queues, $Q_1, \ldots, Q_N$, under a limited policy. The policy is parametrized by a vector $(m_1, \ldots, m_N)$, where m_j is the maximum number of customers that can be served during one server's visit to Q_j. In this chapter we more or less adopt the notation of the preceding chapter. We denote the random variables constituting a polling system by $\{A_j^{n+1}, B_j^n, V_j^n,\ n = 0, 1, \ldots; j = 1, 2, \ldots, N\}$, with the same interpretations as those given in the preceding chapter. We also make the same assumptions concerning the probabilistic structure of these random variables.

7.3 An Approximation for $N = 2$

We consider in this section a polling system consisting of two queues. We first look closely at the stability conditions for this system and then make some key observations which lead to the development of an approximation method to compute the mean waiting times at the individual queues.

7.3.1 Stability Conditions

We say that Q_j, $j = 1, 2$, is stable if the waiting time W_j^n of the nth type–j customer converges to an almost surely finite random variable W_j. We say that the system is stable if both Q_1 and Q_2 are stable. Let $\bar{C}$ denote the (steady state) mean cycle time, i.e., the average amount of time it takes the server to make one full trip around the system. If both queues are stable, then for $j = 1, 2$, $\lambda_j \bar{C}$, is the mean number of type–j customers served in one cycle, and so

$$\bar{C} = \sum_{j=1}^{2} \lambda_j \bar{C} \bar{B}_j + \bar{V}_T, \tag{7.1}$$

where $\bar{V}_T = \bar{V}_1 + \bar{V}_2$. Rearranging the terms yields

$$\bar{C} = \frac{\bar{V}_T}{1 - \rho_T}, \tag{7.2}$$

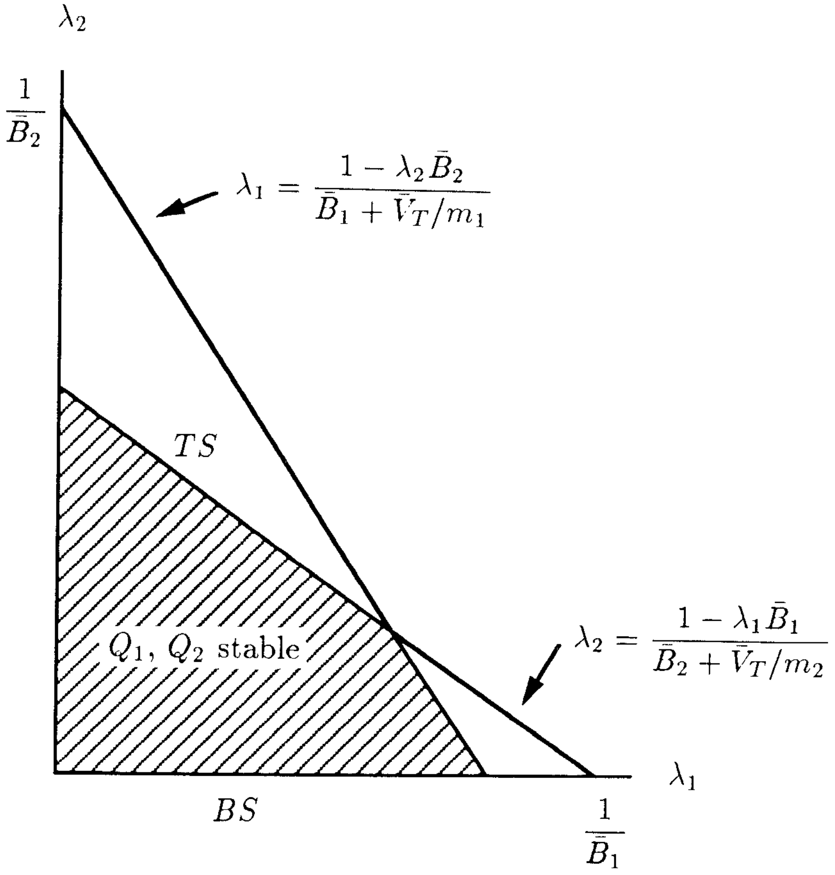

Figure 7.1: The Stability Region for $N = 2$

where $\rho_T = \rho_1 + \rho_2 = \lambda_1\bar{B}_1 + \lambda_2\bar{B}_2$. For both Q_1 and Q_2 to be stable, it is necessary that

$$\lambda_j\bar{C} = \frac{\lambda_j\bar{V}_T}{1-\rho_T} < m_j, \qquad j = 1, 2, \tag{7.3}$$

or equivalently

$$\lambda_j < \frac{1-\rho_T+\lambda_j\bar{B}_j}{\bar{B}_j+\bar{V}_T/m_j}, \qquad j = 1, 2. \tag{7.4}$$

These conditions were first derived by Kuehn [51]; Szpankowski and Rego [76] showed that these conditions were also sufficient. The stability region of a polling system under the limited policy with parameters (m_1, m_2) is shown in Figure 7.1. We see that the preceding derivation of stability conditions can be easily extended to the case N arbitrary.

7.3.2 Observations

We seek to approximate the mean waiting time $\bar{W}_1(\lambda_1, \lambda_2)$ at Q_1, where λ_1 and λ_2 are the rate of the Poisson arrival to Q_1 and Q_2, respectively. (The approximation of $\bar{W}_2$ can be carried out in exactly the same manner and so we only discuss $\bar{W}_1$ here.) First, we make the following observations:

1. If Q_2 is unstable, Q_1 can still be stable if and only if
$$\lambda_1 < \frac{m_1}{m_1\bar{B}_1 + m_2\bar{B}_2 + \bar{V}_T}. \tag{7.5}$$
Indeed, denoting the mean cycle time in this case by $\bar{C}_1$, we have
$$\bar{C}_1 = \lambda_1\bar{C}_1\bar{B}_1 + m_2\bar{B}_2 + \bar{V}_T, \tag{7.6}$$
and so
$$\bar{C}_1 = \frac{m_2\bar{B}_2 + \bar{V}_T}{1 - \lambda_1\bar{B}_1}. \tag{7.7}$$
The necessary and sufficient condition for Q_1 to be stable is $\lambda_1\bar{C}_1 < m_1$ which is equivalent to (7.5). As far as Q_1 is concerned, the system can be considered as a vacation model under the policy $L(m_1)$ with Q_1 as the queue and vacation periods being distributed as $B_2^1 + \cdots + B_2^{m_2} + V_T$ (where the additions are independent). The stability condition (7.5) can be seen to be consistent with (3.54).

2. If $\lambda_2 = 0$, then Q_1 can again be considered as a vacation model. In this case, the vacation periods are distributed as V_T, and Q_1 is stable if and only if
$$\lambda_1 < \frac{1}{\bar{B}_1 + \bar{V}_T/m_1}. \tag{7.8}$$

So, for the cases Q_2 unstable and $\lambda_2 = 0$, $\bar{W}_1$ can be approximated very accurately using the interpolation approximation for vacation models developed in Chapter 3. These cases are depicted in Figure 7.1 as the top side TS and bottom side BS of the quadrilateral representing the stability region. The approximation of $\bar{W}_1$ along the lines TS and BS is very accurate, and we would like to exploit this information to approximate $\bar{W}_1(\lambda_1^*, \lambda_2^*)$ for $(\lambda_1^*, \lambda_2^*)$ in the interior of the stability region. We consider below one interpolation method to do exactly this.

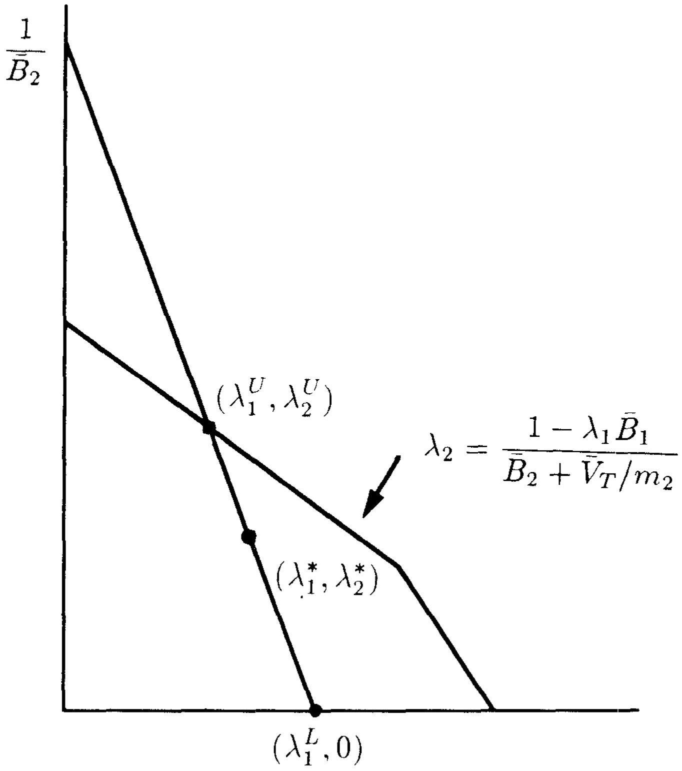
$\frac{1}{\bar{B}_2}$
$(\lambda_1^U, \lambda_2^U)$
$\lambda_2 = \frac{1 - \lambda_1 \bar{B}_1}{\bar{B}_2 + \bar{V}_T/m_2}$
$(\lambda_1^*, \lambda_2^*)$
$(\lambda_1^L, 0)$

Figure 7.2: An interpolation

7.3.3 An Interpolation

Given $(\lambda_1^*, \lambda_2^*)$ in the interior of the stability region, we use the information along TS and BS by interpolating along a line connecting a point $(\lambda_1^U, \lambda_2^U)$ on TS and a point $(\lambda_1^L, 0)$ on BS which crosses $(\lambda_1^*, \lambda_2^*)$. Obviously, there are infinitely many choices of $(\lambda_1^U, \lambda_2^U)$ and $(\lambda_1^L, 0)$ for any given $(\lambda_1^*, \lambda_2^*)$; we shall consider one choice which is obtained as follows. From the point $(0, \frac{1}{\bar{B}_2})$, draw a straight line crossing $(\lambda_1^*, \lambda_2^*)$ and take the point where it crosses TS as $(\lambda_1^U, \lambda_2^U)$ and where it crosses BS as $(\lambda_1^L, 0)$ (see Figure 7.2). Simple algebra will show that this line indeed crosses TS and BS and that furthermore the obtained λ_1^U and λ_1^L satisfy (7.5) and (7.8), respectively.

To be more specific, the approximation of $\bar{W}_1(\lambda_1^*, \lambda_2^*)$ is carried out as follows:

1. Given $(\lambda_1^*, \lambda_2^*)$, set

$$(\lambda_1^U, \lambda_2^U) = \Big(\frac{\lambda_1^* \bar{V}_T}{m_2\bar{B}_2(1-\rho_T) + \bar{V}_T(1-\lambda_2^*\bar{B}_2)}, \frac{m_2(1-\rho_T)}{m_2\bar{B}_2(1-\rho_T) + \bar{V}_T(1-\lambda_2^*\bar{B}_2)}\Big) \tag{7.9}$$

   and

$$(\lambda_1^L, 0) = \Big(\frac{\lambda_1^*}{1-\lambda_2^*\bar{B}_2}, 0\Big). \tag{7.10}$$

2. If $m_1 > 1$, approximate $\bar{W}_1(\lambda_1^U, \lambda_2^U)$ using (3.61) with

$$\bar{V} = m_2\bar{B}_2 + \bar{V}_T \qquad \text{and} \qquad \sigma_V^2 = m_2\sigma_{B_2}^2 + \sigma_{V_T}^2. \tag{7.11}$$

   Otherwise if $m = 1$, $\bar{W}_1(\lambda_1^U, \lambda_2^U)$ can be computed exactly using (3.60) with $p = 0$ and with vacation mean and variance given above. Denote the computed $\bar{W}_1(\lambda_1^U, \lambda_2^U)$ by $\bar{w}_1^U(\lambda_1^U)$.

3. Approximate or compute $\bar{W}_1(\lambda_1^L, 0)$ similarly, with

$$\bar{V} = \bar{V}_T \qquad \text{and} \qquad \sigma_V^2 = \sigma_{V_T}^2. \tag{7.12}$$

   Again, denote the computed $\bar{W}_1(\lambda_1^L, 0)$ by $\bar{w}_1^U(\lambda_1^L)$.

4. Approximate $\bar{W}_1(\lambda_1^*, \lambda_2^*)$ by $\bar{w}_1(\lambda_1^*, \lambda_2^*)$ given by

$$\bar{w}_1(\lambda_1^*, \lambda_2^*) = \bar{w}_1^U(\lambda_1^U) + \frac{\lambda_1 - \lambda_1^U}{\lambda_1^L - \lambda_1^U}\big[\bar{w}_1^U(\lambda_1^L) - \bar{w}_1^U(\lambda_1^U)\big]. \tag{7.13}$$

7.4 An Algorithm for N Arbitrary

In this section, we extend the interpolation approximation for the case $N = 2$ to the case N arbitrary. First, as mentioned earlier, the stability conditions for this case can be derived in the same manner as for the case $N = 2$. Thus, the system is stable if and only if

$$\lambda_j < \frac{1 - \rho_T + \lambda_j\bar{B}_j}{\bar{B}_j + \bar{V}_T/m_j}, \qquad j = 1, \ldots, N, \tag{7.14}$$

where $\rho_T = \sum_{i=1}^N \lambda_i\bar{B}_i$ and $\bar{V}_T = \sum_{i=1}^N \bar{V}_i$.

7.4.1 Observations

In the case $N = 2$ considered in the last section, $\bar{W}_1(\lambda_1^*, \lambda_2^*)$ is approximated by interpolating between two points: $(\lambda_1^U, \lambda_2^U)$ and $(\lambda_1^L, 0)$. At these two points, as far as $\bar{W}_1$ is concerned, the system can be considered as a vacation model. This idea can be naturally extended to the case N arbitrary as follows. Given $(\lambda_1^*, \ldots, \lambda_N^*)$ in the stability region, we find two points $(\lambda_1^U, \ldots, \lambda_N^U)$ and $(\lambda_1^L, \ldots, \lambda_{N-1}^L, 0)$. Assuming that $\bar{W}_1(\lambda_1^U, \ldots, \lambda_N^U)$ and $\bar{W}_1(\lambda_1^L, \ldots, \lambda_{N-1}^L, 0)$ (or their approximations) are available, we then approximate $\bar{W}_1(\lambda_1^*, \ldots, \lambda_N^*)$ by interpolating between these two points. The point $(\lambda_1^U, \ldots, \lambda_N^U)$ is chosen such that at this point Q_N is unstable. Obviously, as far as $\bar{W}_1$ is concerned, the system is equivalent to a polling system with $N-1$ queues (after modifying the total switchover times) at the two points mentioned above. As a result, the computation (approximation) of $\bar{W}_1(\lambda_1^U, \ldots, \lambda_N^U)$ is reduced to the computation of $\bar{W}_1$ of an equivalent polling system with $N-1$ queues and with arrival rates $(\lambda_1^U, \ldots, \lambda_{N-1}^U)$. Similarly, $\bar{W}_1(\lambda_1^L, \ldots, \lambda_{N-1}^L, 0)$ is exactly $\bar{W}_1$ of an equivalent polling system with $N-1$ queues and with arrival rates $(\lambda_1^L, \ldots, \lambda_{N-1}^L)$. For the point $(\lambda_1^U, \ldots, \lambda_N^U)$, the total switchover times of the equivalent system are distributed as $B_N^1 + \cdots + B_N^{m_N} + V_T$ (where the additions are independent); for the point $(\lambda_1^L, \ldots, \lambda_{N-1}^L, 0)$, the total switchover times are distributed as the ones in the original system.

Given $(\lambda_1^*, \ldots, \lambda_N^*)$ in the stability region, we obtain $(\lambda_1^U, \ldots, \lambda_N^U)$ and $(\lambda_1^L, \ldots, \lambda_{N-1}^L, 0)$ by drawing a straight line from $(0, \ldots, 0, \frac{1}{\bar{B}_N})$ to $(\lambda_1^*, \ldots, \lambda_N^*)$. Again, simple algebra will show that indeed this line and its extension crosses the stability "plane" of Q_N,

$$\lambda_N = \frac{1 - \sum_{i=1}^{N-1} \lambda_i \bar{B}_i}{\bar{B}_N + \bar{V}_T / m_N}, \tag{7.15}$$

and the plane $\lambda_N = 0$, and that the resulting $(\lambda_1^U, \ldots, \lambda_{N-1}^U)$ and $(\lambda_1^L, \ldots, \lambda_{N-1}^L)$ satisfy the stability conditions for their respective equivalent polling system.

To approximate $\bar{W}_1(\lambda_1^U, \ldots, \lambda_N^U)$ and $\bar{W}_1(\lambda_1^L, \ldots, \lambda_{N-1}^L, 0)$, the same approach used to approximate $\bar{W}_1(\lambda_1^*, \ldots, \lambda_N^*)$ can be applied. For example, the approximation of $\bar{W}_1(\lambda_1^L, \ldots, \lambda_{N-1}^L, 0)$ will involve the two points $(\lambda_1^{LU}, \ldots, \lambda_{N-1}^{LU}, 0)$ and

$(\lambda_1^{LL},\dots,\lambda_{N-2}^{LL},0,0)$. This line of argument is repeated $N-1$ times until we end up with only vacation models, at which point we can use the interpolation approximation for vacation models developed in Chapter 3.

7.4.2 The Algorithm

The algorithm consists of two phases: the *reduction phase* and the *interpolation phase*. In the reduction phase, we reduce the dimensionality of the system in a manner described above; in the interpolation phase, we perform interpolation between pairs of points. The reduction phase is subsequently divided into N *stages*: $\{R_j,\ j=0,\dots,N\}$. At stage R_j, we have polling systems with $N-j$ queues, i.e., we have "removed" $Q_N,\dots,Q_{N-j+1}$.

Let $\{S^{(j,k)},\ k=0,\dots,2^j-1\}$, be the polling systems (with $N-j$ queues) at stage R_j, $j=0,\dots,N-1$. The numbering of the systems within one stage is done as follows. Let the j-bit binary expansion of k be given by $d_0,\dots,d_{j-1}$. Then, $S^{(j,k)}$ is obtained by removing $Q_N,\dots,Q_{N-j+1}$ where the removal of Q_i, $i=N-j+1,\dots,N$, corresponds to either letting $\lambda_i=0$ (if $d_i=0$) or saturating Q_i (if $d_i=1$).

Let $x^{(j,k)}=(x_1^{(j,k)},\dots,x_{N-j}^{(j,k)})$, $j=0,\dots,N-1$; $k=0,\dots,2^j-1$, be the arrival rate vectors corresponding to $S^{(j,k)}$. For example, $x^{(0,0)}$ would be equal to $(\lambda_1^*,\dots,\lambda_N^*)$, $x^{(1,1)}=(\lambda_1^U,\dots,\lambda_{N-1}^U)$, $x^{(2,1)}=(\lambda_1^{LU},\dots,\lambda_{N-2}^{LU})$ and $x^{(2,2)}=(\lambda_1^{UU},\dots,\lambda_{N-2}^{UU})$. Also denote the (generic) total switchover time for $S^{(j,k)}$ by $V_T^{(j,k)}$ and the mean waiting time at Q_1 and its approximation by $\bar{W}_1^{(j,k)}$ and $\bar{w}_1^{(j,k)}$, respectively.

Given $x^{(j,k)}$, we generate $x^{(j+1,2k)}$ and $x^{(j+1,2k+1)}$ as follows. The parametric representation of the line connecting $(0,\dots,0,\frac{1}{\bar{B}_{N-j}})$ and $x^{(j,k)}$ is given by

$$\left(tx_1^{(j,k)},\dots,tx_{N-j-1}^{(j,k)},\frac{1}{\bar{B}_{N-j}}+\left(x_{N-j}^{(j,k)}-\frac{1}{\bar{B}_{N-j}}\right)t\right),\qquad t\geq 0. \tag{7.16}$$

Let $t_1^{(j,k)}$ be the value of t when this line crosses the stability plane of Q_{N-j} (as part of the system $S^{(j,k)}$), namely

$$x_{N-j}=\frac{1-\sum_{j=1}^{N-j-1}x_j\bar{B}_j}{\bar{B}_{N-j}+\bar{V}_T^{(j,k)}/m_{N-j}}, \tag{7.17}$$

and let $t_0^{(j,k)}$ be the value of t when the line crosses the plane $x_{N-j} = 0$. Thus, $x^{(j+1,2k)}$ and $x^{(j+1,2k+1)}$ are given by

$$x^{(j+1,2k)} = \left(t_0^{(j,k)} x_1^{(j,k)}, \ldots, t_0^{(j,k)} x_{N-j-1}^{(j,k)}\right) \tag{7.18}$$

and

$$x^{(j+1,2k+1)} = \left(t_1^{(j,k)} x_1^{(j,k)}, \ldots, t_1^{(j,k)} x_{N-j-1}^{(j,k)}\right), \tag{7.19}$$

respectively. It can be shown using simple algebra that $t_1^{(j,k)}$ and $t_0^{(j,k)}$ are given by

$$t_1^{(j,k)} = \frac{\bar{V}_T^{(j,k)}}{m_{N-j}\bar{B}_{N-j}(1-\rho_T^{(j,k)}) + \bar{V}_T^{(j,k)}(1-x_{N-j}^{(j,k)}\bar{B}_{N-j})} \tag{7.20}$$

and

$$t_0^{(j,k)} = \frac{1}{1-x_{N-j}^{(j,k)}\bar{B}_{N-j}}, \tag{7.21}$$

respectively, where $\rho_T^{(j,k)} = \sum_{i=1}^{N-j} x_i^{(j,k)}\bar{B}_i$.

The Algorithm:

```
** *Initialization* **
$x^{(0,0)} \leftarrow (\lambda_1^*, \ldots, \lambda_N^*)$
$\bar{V}_T^{(0,0)} \leftarrow \bar{V}_T$   and   $\sigma^2_{V_T^{(0,0)}} \leftarrow \sigma^2_{V_T}$
** *Reduction phase* **
do $j = 0,\ N-2$
    do $k = 0,\ 2^j - 1$
        compute $x^{(j+1,2k)}$ using (7.18) and (7.21)
        $\bar{V}_T^{(j+1,2k)} \leftarrow \bar{V}_T^{(j,k)}$
        $\sigma^2_{V_T^{(j+1,2k)}} \leftarrow \sigma^2_{V_T^{(j,k)}}$
        compute $x^{(j+1,2k+1)}$ using (7.19) and (7.20)
        $\bar{V}_T^{(j+1,2k+1)} \leftarrow \bar{V}_T^{(j,k)} + m_{N-j}\bar{B}_{N-j}$
        $\sigma^2_{V_T^{(j+1,2k+1)}} \leftarrow \sigma^2_{V_T^{(j,k)}} + m_{N-j}\sigma^2_{B_{N-j}}$
```

```
** Interpolation phase: **
do $k = 0,\ 2^{N-1} - 1$
    compute $\bar{w}_1^{(N-1,k)}$ using (3.61) or (3.60)
do $j = N-2,\ 0,\ -1$
    do $k = 0,\ 2^j - 1$
        $\bar{w}_1^{(j,k)} \leftarrow \bar{w}_1^{(j+1,2k)} + \left(\bar{w}_1^{(j+1,2k+1)} - \bar{w}_1^{(j+1,2k)}\right) x_{N-j}^{(j,k)} / x_{N-j}^{(j+1,2k+1)}$
```

7.5 Numerical Results

In this section, we study the performance of the algorithm proposed above. We compare approximations computed using the algorithm with simulation results. We also compare our algorithm to Fuhrmann and Wang's approximation [35]. This approximation uses the idea discussed in Section 6.7 with the pseudo–conservation law for the limited policy being approximated by that for the Bernoulli policy. This approximation is reasonable in view of the comparison result mentioned in Remark 6.1.

In the following tables, the simulated results are listed under the heading "$\bar{W}_1$"; our approximations under "$\bar{w}_1$"; and Fuhrmann & Wang's approximation under "$\bar{w}_1^{FW}$". As in Chapter 3, we use 90% confidence level for computing the confidence intervals of the simulated points.

System 1. Small N ($N = 3$); low load ($\rho_T = 0.3$); homogeneous service times, switchover times and arrival rates ($\bar{B}_i = 1.0$ exponential, $V_i = 0.05$ deterministic, $\lambda_i^* = 0.1$, $i = 1,2,3$).

System 2. Small N ($N = 3$); medium load ($\rho_T = 0.5$); homogeneous service times and switchover times ($\bar{B}_i = 1.0$ exponential, $V_i = 0.25$ deterministic, $i = 1,2,3$); nonhomogeneous arrival rates ($\lambda_1^* = 0.1$, $\lambda_2^* = \lambda_3^* = 0.2$).

System 3. Small N ($N = 3$); medium load ($\rho_T = 0.6$); homogeneous switchover times ($V_i = 0.25$, deterministic, $i = 1,2,3$); nonhomogeneous service times and

(m_1, m_2, m_3)	$\bar{W}_1$	$\bar{w}_1$	%Error	$\bar{w}_1^{FW}$
$(3,1,1)$	0.453 ± 0.008	0.412	−9.1	0.491
$(6,2,1)$	0.467 ± 0.008	0.460	−1.5	0.490
$(6,1,3)$	0.472 ± 0.008	0.510	+8.1	0.495
$(1,3,3)$	0.637 ± 0.015	0.644	+1.1	0.606
$(1,2,5)$	0.625 ± 0.012	0.695	+11.7	0.605

Table 7.1: System 1

(m_1, m_2, m_3)	$\bar{W}_1$	$\bar{w}_1$	%Error	$\bar{w}_1^{FW}$
$(3,1,1)$	1.265 ± 0.024	1.161	−8.2	1.487
$(6,2,1)$	1.294 ± 0.025	1.250	−3.4	1.474
$(6,1,3)$	1.377 ± 0.026	1.349	−2.0	1.508
$(1,3,3)$	2.550 ± 0.112	1.990	−22.0	2.342
$(1,2,5)$	2.432 ± 0.096	2.105	−13.4	2.329

Table 7.2: System 2

(m_1, m_2, m_3)	$\bar{W}_1$	$\bar{w}_1$	%Error	$\bar{w}_1^{FW}$
$(3,1,1)$	1.783 ± 0.052	1.694	-5.0	2.494
$(6,2,1)$	1.704 ± 0.037	1.704	0.0	2.355
$(6,1,3)$	2.044 ± 0.045	1.967	-6.0	2.417
$(2,3,3)$	2.827 ± 0.131	2.169	-23.3	3.427
$(2,2,5)$	3.308 ± 0.105	2.424	-26.7	3.407

Table 7.3: System 3

(m_1, m_2, m_3)	$\bar{W}_1$	$\bar{w}_1$	%Error	$\bar{w}_1^{FW}$
$(3,1,1)$	10.46 ± 0.28	9.11	-12.9	18.59
$(6,2,1)$	8.78 ± 0.30	9.07	$+3.3$	15.18
$(6,1,3)$	10.09 ± 0.36	9.15	-9.3	15.53
$(3,1,5)$	13.34 ± 0.50	10.45	-21.7	20.21
$(2,1,3)$	15.21 ± 0.54	10.83	-28.8	23.89

Table 7.4: System 4

arrival rates ($\bar{B}_1 = 2.0$, $\bar{B}_2 = \bar{B}_3 = 1.0$ all exponential, $\lambda_1^* = 0.1$, $\lambda_2^* = \lambda_3^* = 0.2$).

System 4. Small N ($N = 3$); heavy load ($\rho_T = 0.8$); homogeneous switchover times ($V_i = 0.25$, deterministic, $i = 1, 2, 3$); nonhomogeneous service times and arrival rates ($\bar{B}_1 = 5.0$, $\bar{B}_2 = 1.0$, $\bar{B}_3 = 3.0$ all exponential, $\lambda_1^* = 0.1$, $\lambda_2^* = 0.15$, $\lambda_3^* = 0.05$).

System 5. Large N ($N = 10$); low load ($\rho_T = 0.4$); homogeneous service times, switchover times and arrival rates ($\bar{B}_i = 1.0$ exponential, $V_i = 0.25$ deterministic, $\lambda_i^* = 0.04$, $i = 1, \ldots, 10$).

System 6. Large N ($N = 10$); medium load ($\rho_T = 0.594$); homogeneous switchover times ($V_i = 0.25$ deterministic, $i = 1, \ldots, 10$); nonhomogeneous service times and

$(m_1,\ldots,m_{10})$	$\bar{W}_1$	$\bar{w}_1$	%Error	$\bar{w}_1^{FW}$
$(3,1,1,1,1,1,1,1,1,1)$	2.529 ± 0.025	2.391	−5.5	2.807
$(6,2,2,2,2,1,1,1,1,1)$	2.534 ± 0.038	2.462	−2.8	2.704
$(5,1,1,1,1,3,3,3,3,3)$	2.585 ± 0.030	2.583	−0.1	2.744

Table 7.5: System 5

$(m_1,\ldots,m_{10})$	$\bar{W}_1$	$\bar{w}_1$	%Error	$\bar{w}_1^{FW}$
$(3,1,1,1,1,1,1,1,1,1)$	3.598 ± 0.067	3.367	−6.4	4.949
$(6,2,2,2,2,1,1,1,1,1)$	3.559 ± 0.056	3.354	−5.8	4.299
$(5,1,1,1,1,3,3,3,3,3)$	3.514 ± 0.046	3.374	−4.0	4.850

Table 7.6: System 6

arrival rates ($\bar{B}_i = 1.0$, $i = 1,2,3$, $\bar{B}_i = 0.6$, $i = 4,\ldots,10$, all exponential; $\lambda_i^* = 0.1$, $i = 1,2,3$, $\lambda_i^* = 0.07$, $i = 4,\ldots,10$).

System 7. Large N ($N = 10$); heavy load ($\rho_T = 0.718$); homogeneous switchover times ($V_i = 0.25$ deterministic, $i = 1,\ldots,10$); nonhomogeneous service times, switchover times and arrival rates ($\bar{B}_i = 0.6$, $i = 1,2,3,8,9,10$, $\bar{B}_i = 1.0$, $i = 4,\ldots,7$, all exponential; $\lambda_i^* = 0.1$, $i = 1,2,3$, $\lambda_i^* = 0.08$, $i = 4,5$, $\lambda_i^* = 0.09$, $i = 6,7$, $\lambda_i^* = 0.11$, $i = 8,9,10$).

In the examples above, we see that the algorithm indeed performs quite well.

$(m_1,\ldots,m_{10})$	$\bar{W}_1$	$\bar{w}_1$	%Error	$\bar{w}_1^{FW}$
$(3,1,1,1,1,1,1,1,1,1)$	5.031 ± 0.060	4.874	−3.1	7.915
$(6,2,2,2,2,1,1,1,1,1)$	4.731 ± 0.060	4.712	−0.4	6.482
$(5,1,1,1,1,3,3,3,3,3)$	5.122 ± 0.082	4.927	−3.8	8.822

Table 7.7: System 7

$(\lambda_1^*, \lambda_2^*)$	$\bar{W}_1$	$\bar{w}_1$	%Error	$\bar{w}_1^{FW}$
$(0.10, 0.20)$	0.424 ± 0.009	0.467	+10.1	0.455
$(0.10, 1.00)$	1.180 ± 0.019	1.092	−7.5	1.497
$(0.10, 1.40)$	1.561 ± 0.027	1.469	−5.9	2.948
$(0.50, 0.20)$	2.469 ± 0.119	2.083	−15.6	2.617
$(0.70, 0.20)$	16.826 ± 1.358	12.839	−23.7	17.008
$(0.30, 0.70)$	2.469 ± 0.132	1.516	−38.6	2.579

Table 7.8: System 8

We notice from Systems 1–4 that the accuracy of the algorithm generally tends to decrease as we increase the system load. However, this degradation is not as severe as in Fuhrmann & Wang's approximation. One surprising observation is that the algorithm performs well even for large N (Systems 5–7); we would expect the error to accumulate as we increase N. A possible explanation for this is illustrated by the following three examples (Systems 8–10). In these three examples, we consider a polling system with $N = 2$. We increase the mean total switchover times (System 8 has the smallest mean total switchover times, System 10 has the highest) while keeping other system parameters constant. For each system, we consider a set of six points whose relative locations in their respective stability region are kept roughly the same for all three systems. We observe that the accuracy of the approximations increases as we increase the mean total switchover times. This observation may provide an explanation to the fact that the algorithm performs well even for large N because as we increase N, the mean total switchover times increases.

System 8. $N = 2$; $m_1 = 2$, $m_2 = 4$; $\bar{B}_1 = 1.0$, $\bar{B}_2 = 0.5$ exponential; $V_1 = V_2 = 0.2$ deterministic.

System 9. $N = 2$; $m_1 = 2$, $m_2 = 4$; $\bar{B}_1 = 1.0$, $\bar{B}_2 = 0.5$ exponential; $V_1 = V_2 = 0.5$ deterministic.

$(\lambda_1^*, \lambda_2^*)$	$\bar{W}_1$	$\bar{w}_1$	%Error	$\bar{w}_1^{FW}$
$(0.10, 0.20)$	0.789 ± 0.014	0.815	+3.3	0.874
$(0.10, 0.50)$	1.096 ± 0.021	1.103	+0.6	1.236
$(0.10, 1.00)$	1.734 ± 0.029	1.632	−5.9	2.385
$(0.30, 0.20)$	1.659 ± 0.062	1.505	−9.3	1.934
$(0.50, 0.20)$	6.934 ± 0.222	5.475	−21.0	7.316
$(0.30, 0.50)$	2.879 ± 0.054	2.132	−25.9	3.281

Table 7.9: System 9

$(\lambda_1^*, \lambda_2^*)$	$\bar{W}_1$	$\bar{w}_1$	%Error	$\bar{w}_1^{FW}$
$(0.10, 0.20)$	1.439 ± 0.022	1.416	−1.6	1.653
$(0.10, 0.40)$	1.724 ± 0.029	1.679	−2.6	3.021
$(0.10, 0.70)$	2.225 ± 0.039	2.111	−5.1	2.837
$(0.25, 0.20)$	2.572 ± 0.092	2.369	−7.9	3.306
$(0.40, 0.20)$	12.048 ± 0.191	10.635	−11.7	14.961
$(0.25, 0.40)$	3.799 ± 0.096	3.062	−21.9	4.655

Table 7.10: System 10

System 10. $N = 2$; $m_1 = 2, m_2 = 4$; $\bar{B}_1 = 1.0$, $\bar{B}_2 = 0.5$ exponential; $V_1 = V_2 = 1.0$ deterministic.

CHAPTER 8

CONCLUSIONS

In this thesis, we showed that nonexact methods based on heavy and light traffic analysis and stochastic comparison techniques could indeed provide useful information about the performance of some nonexhaustive service policies in polling systems and vacation models. Most nonexhaustive policies, in particular the limited policy, have been deemed mathematically untractable; "traditional" exact approaches which have been very successfully used to study the exhaustive and gated policies have been shown to be unsuitable for these policies [82,83]. Numerous approximations have also been studied in the literature; more recent approximations are based on the pseudo–conservation laws. Unfortunately, pseudo–conservation laws for the limited policy and some other nonexhaustive policies have not been established either. Some approximation methods based on approximate pseudo–conservation laws have been studied [34,35]. In the first part of this thesis, we established heavy and light traffic results for vacation models. These results were then used as bases for interpolation approximations for vacation models. In the second part of the thesis, the interpolation approximations were extended to polling systems with an arbitrary number of queues. Because they are based on exact results for heavy and light traffic conditions, these approximations tend to be very accurate in the extreme traffic regions. Numerical examples also suggested that the approximations perform well on polling systems in which the queues are highly nonhomogeneous—systems which are not handled well by approximations based on (approximate) pseudo–conservation laws. One other advantage of interpolation approximations is that their accuracy can always be improved by additional information about the system which might be obtained in the future.

For service policies which cannot be analyzed exactly, it is useful to know if their performance is bounded by the performance of some other policies which are analyzable. Motivated by this idea, we developed a framework in which the performance of service policies in the context of vacation models can be stochastically compared. The comparison between the Bernoulli and limited policies is particularly important since the Bernoulli policy is more tractable than the limited policy. We showed that under some conditions, the Bernoulli policy upper–bounds the limited policy in terms of some performance measures of interest. Analysis of the Bernoulli policy yielded a similar but weaker comparison in the context of polling systems.

The approaches to performance evaluation of polling systems studied in this thesis have hardly been used by researchers in the field. A work on stochastic comparison of policies in the context of polling systems has just been started by Levy et al. [57]. Kella and Whitt [46] established heavy traffic results for vacation models where the vacation times and/or the time between vacations increase as the traffic intensity increases. They showed that the steady–state distributions of the heavy traffic limit processes exhibit decomposition properties. We believe that more results along these lines should be established to better understand the behavior of polling systems and vacation models under nonexhaustive policies. In the following, we propose some possible extensions to the results established in this thesis.

For the heavy and light traffic analysis and the resulting interpolation approximations, we considered mostly the limited and Bernoulli policies. We expect that a similar approach can be used for other nontractable service policies. In particular, we should be able to obtain heavy and light traffic results for the gated variants of the limited and Bernoulli policies using the methods used here with minor modifications.

In this thesis, we stochastically compared various service policies in vacation models in terms of waiting time processes, among others. We expect that these results, or at least the underlying techniques, can be extended to polling systems. In [57], Levy et al. compared service policies in polling systems in terms of the total amount of work in the system; it would be of interest to combine their approach with ours to obtain comparison results for waiting time processes in polling systems.

APPENDIX A

WEAK CONVERGENCE THEOREMS

In this Appendix, we collect theorems in the theory of weak convergence that are used in the thesis. Our primary source for terminology and basic theorems is Billingsley [6]. However, functional limit theorems that deal with double sequences are taken from Kyprianou [52]. Special cases of these results for single sequences can be found in Billingsley. We use $\mathcal{W}$ to denote a Wiener process on $[0,1]$.

Theorem A.1 (Continuous mapping) *Let $(S, \mathcal{S})$ and $(S', \mathcal{S}')$ be two probability spaces where both S and S' are metric spaces. Let $\{P_n,\ n = 0,1,\ldots\}$ and P be probability measures on $(S, \mathcal{S})$ and $h : S \to S'$ be a measurable map. Suppose $P(D_h) = 0$ where D_h is the set of discontinuities of h. If $P_n \xrightarrow{\mathcal{D}} P$ in $\mathcal{S}$, then $P_n h^{-1} \xrightarrow{\mathcal{D}} Ph^{-1}$ in $\mathcal{S}'$.*

Theorem A.2 (Converging together) *Let $\{X_n,\ n = 0,1,\ldots\}$ and $\{Y_n,\ n = 0,1,\ldots\}$ be two sequences of random variables with range $(S, \mathcal{S})$; S a separable metric space with metric ρ. Suppose that X_n and Y_n have a common domain so that $\rho(X_n, Y_n)$ is a real valued random variable. If $X_n \xrightarrow{\mathcal{D}} X$ and $\rho(X_n, Y_n) \xrightarrow{P} 0$, then $Y_n \xrightarrow{\mathcal{D}} X$.*

Theorem A.3 (Random Time Change) *Let $\{X_n,\ n = 0,1,\ldots\}$ be a sequence of $D[0,1]$-valued random functions and $\{\Phi_n,\ n = 0,1,\ldots\}$ a sequence of $D_0[0,1]$-valued random functions, where $D_0[0,1]$ consists of nondecreasing functions ϕ in $D[0,1]$ that satisfy $0 \le \phi(t) \le 1$. Assume that for each n, X_n and Φ_n have the same domain (which may vary with n). If $X_n \xrightarrow{\mathcal{D}} X$ with $P[X \in C[0,1]] = 1$ and $m(\Phi_n, \phi) \xrightarrow{\mathcal{D}} 0$ for some continuous function ϕ in D_0, then*

$$X_n(\Phi_n) \xrightarrow{\mathcal{D}} X(\phi). \tag{A.1}$$

Theorem A.4 (Prohorov) *If the double sequence $\{Z_n^r,\ r,n = 1,2,\ldots\}$ satisfies Condition A in Definition 2.1 with $Z_0^r = 0$ and $E[Z_n^r] = 0$ for all $r = 1,2,\ldots,$ then ξ^r defined by*

$$\xi^r(t) = \frac{1}{\sigma_Z\sqrt{r}}\sum_{i=0}^{\lfloor rt\rfloor} Z_i^r, \qquad 0 \le t \le 1 \tag{A.2}$$

converges weakly to $\mathcal{W}$ as $r \to \infty$.

Theorem A.5 (Random sums) *Let $\{Y_n^r,\ r,n = 1,2,\ldots\}$ be a sequence of zero–mean random variables and $\{X_n^r,\ r,n = 1,2,\ldots\}$ a sequence of nonnegative random variables, both satisfying Condition A. For $r = 1,2,\ldots,$ set $X_0^r = Y_0^r = 0$ and define*

$$N^r(t) = \max\{n \ge 0 : X_0^r + \cdots + X_n^r \le t\}, \qquad t \ge 0. \tag{A.3}$$

Then, ζ^r defined by

$$\zeta^r(t) = \frac{\sqrt{\bar{X}}}{\sigma_Y\sqrt{r}}\sum_{i=0}^{N^r(rt)} Y_i^r, \qquad 0 \le t \le 1 \tag{A.4}$$

converges weakly to $\mathcal{W}$ as $r \to \infty$.

Theorem A.6 (Renewal process) *Let $\{X_n^r,\ r,n = 1,2,\ldots\}$ be a sequence of nonnegative random variables satisfying Conditon A and define N^r as in Theorem A.5. Set*

$$\nu^r(t) = \frac{\bar{X}^{3/2}}{\sigma_X\sqrt{r}}\left[N^r(rt) - rt/\bar{X}^r\right], \qquad 0 \le t \le 1. \tag{A.5}$$

Then, ν^r converges weakly to $\mathcal{W}$ as $r \to \infty$.

REFERENCES

[1] Y. Aminetzah (1975), *An Exact Approach to the Polling System*, Ph.D. Thesis, Department of Electrical Engineering, McGill University, Montreal, Quebec.

[2] J.M. Appleton and M.M. Peterson (1986), "Traffic analysis of a token ring PBX," *IEEE Transactions on Communications* **COM–34** (5), pp. 417–422.

[3] B. Avi–Itzhak, W.L. Maxwell and L.W. Miller (1965), "Queueing with alternating priorities," *Operations Research* **13** (2), pp. 306–318.

[4] F. Baccelli and P. Brémaud (1987), *Palm Probabilities and Stationary Queues,* Springer–Verlag, Berlin.

[5] V. Beneš (1965), *Mathematical Theory of Connecting Networks and Telephone Traffic,* Academic Press, New York.

[6] P. Billingsley (1968), *Convergence of Probability Measures,* John Wiley & Sons, New York.

[7] O.J. Boxma (1986), "Models of two queues: A few new issues," in *Teletraffic Analysis and Computer Performance Evaluation,* O.J. Boxma, J.W. Cohen and H.C. Tijms, Editors, North–Holland, pp. 75–98.

[8] O.J. Boxma (1989), "Workloads and waiting times in single–server systems with multiple customer classes," *Queueing Systems—Theory & Applications* **5**, pp. 185–214.

[9] O.J. Boxma and W.P. Groenendijk (1987), "Pseudo–conservation laws in cyclic–service systems," *Journal of Applied Probability* **24** (4), pp. 949–964.

[10] O.J. Boxma and W.P. Groenendijk (1988), "Waiting times in discrete–time cyclic–service systems," *IEEE Transactions on Communications* **COM–36** (2), pp. 164–170.

[11] O.J. Boxma and B.W. Meister (1987), "Waiting time approximations for cyclic service systems with switchover times," *Performance Evaluation* **7**, pp. 299–308.

[12] D.Y. Burman and D.R. Smith (1983), "A light–traffic theorem for multi–server queues," *Mathematics of Operations Research* **8**, pp. 15–25.

[13] D.Y. Burman and D.R. Smith (1983), "Asymptotic analysis of a queueing model with bursty traffic," *Bell System Technical Journal* **62**, pp. 1433–1453.

[14] W. Bux (1981), "Local–area subnetworks: A performance comparison," *IEEE Transactions on Communications* **COM–29** (10), pp. 1465–1473.

[15] W. Bux (1984), "Performance issues in local–area networks," *IBM Systems Journal* **23** (4), pp. 351–374.

[16] W. Bux (1989), "Token–ring local–area networks and their performance," *Proceedings of the IEEE* **77** (2), pp. 238–256.

[17] E.G. Coffman, Jr. and M.I. Reiman (1984), "Diffusion approximation for computer/communication systems," in *Mathematical Computer Performance and Reliability,* G. Iazeolla, P.–J. Courtois and A. Hordijk, Editors, North–Holland, pp. 419–422.

[18] J.W. Cohen and O.J. Boxma (1981), "The $M/G/1$ queue with alternating service formulated as a Riemann–Hilbert problem," in *Performance '81*, F.J. Kylstra, Editor, North–Holland, pp. 181–199.

[19] R.B. Cooper (1970), "Queues served in cyclic order: Waiting times," *Bell System Technical Journal* **49** (3), pp. 399–413.

[20] R.B. Cooper and G. Murray (1969), "Queues served in cyclic order," *Bell System Technical Journal* **48** (3), pp. 675–689.

[21] B.T. Doshi (1985), "A note on stochastic decomposition in $GI/GI/1$ queue with vacations or set–up times," *Journal of Applied Probability* **22**, pp. 419–422.

[22] B.T. Doshi (1986), "Queueing system with vacations—A survey," *Queueing Systems—Theory & Applications* **1** (1), pp. 22–66.

[23] B.T. Doshi (1990) "Single server queues with vacations," in *Stochastic Analysis of Computer and Communication Systems,* H. Takagi, Editor, North–Holland, Amsterdam, pp. 217–265.

[24] M. Eisenberg (1971), "Two queues with changeover times," *Operations Research* **19** (2), pp. 386–401.

[25] M. Eisenberg (1972), "Queues with periodic service and changeover times," *Operations Research* **20** (2), pp. 440–451.

[26] D. Everitt (1986), "Simple approximations for token ring," *IEEE Transactions on Communications* **COM–34** (7), pp. 719–721.

[27] D. Everitt (1986), "A conservation–type law for the token ring with limited service," *British Telecom Technology Journal* **4** (2), pp. 51–61.

[28] D. Everitt (1988), "Approximations for asymmetric token rings with a limited service discipline," *British Telecom Technology Journal* **6** (3), pp. 5–10.

[29] D. Everitt, "A note on the pseudo–conservation laws for cyclic service systems with limited service disciplines," *IEEE Transactions on Communications* **COM–37** (7), pp. 781–783.

[30] F.J. Ferguson and Y.J. Aminetzah (1985), "Exact results for non–symmetric token–ring systems," *IEEE Transactions on Communications* **COM–33** (3), pp. 223–231.

[31] S.W. Fuhrmann (1985), "Symmetric queues served in cyclic order," *Operations Research Letters* **4** (3), pp. 139–144.

[32] S.W. Fuhrmann (1987), "Inequalities for cyclic service systems with limited service disciplines," in *Proceedings of IEEE/IEICE Global Telecommunications Conference 1987*, Tokyo, November 1987, pp. 182–186.

[33] S.W. Fuhrmann and R.B. Cooper (1985), "Stochastic decompositions in the $M/G/1$ queue with generalized vacations," *Operations Research* **33** (5), pp. 1117–1129.

[34] S.W. Fuhrmann and Y.T. Wang (1987), "Mean waiting time approximations of cyclic service systems with limited service," in *Performance '87*, P.-J. Courtois and G. Latouche, Editors, North-Holland, Amsterdam, pp. 253–265.

[35] S.W. Fuhrmann and Y.T. Wang (1988), "Analysis of cyclic service systems with limited service: Bounds and approximations," *Performance Evaluation* **9**, pp. 35–54.

[36] E. Gelenbe and R. Iasnogorodski (1980), "A queue with server of walking type," *Annales de l'Institut Henri Poincaré* **16**, pp. 63–73.

[37] J.M. Harrison (1973), "The heavy traffic approximation for single server queues in series," *Journal of Applied Probability* **10**, pp. 613–629.

[38] J.M. Harrison (1978), "Diffusion approximation for tandem queues in heavy traffic," *Advances in Applied Probability* **10**, pp. 886–905.

[39] J.M. Harrison (1985), *Brownian Motion and Stochastic Flow Systems,* John Wiley & Sons, New York.

[40] O. Hashida (1972), "Analysis of multiqueue," *Review of the Electrical Communication Laboratories* **20** (3–4), pp. 189–199.

[41] D.L. Iglehart and W. Whitt (1970), "Multiple channel queues in heavy traffic I," *Advances in Applied Probability* **2**, pp. 150–177.

[42] D.L. Iglehart and W. Whitt (1970), "Multiple channel queues in heavy traffic II: Sequences, networks, and batches," *AAP* **2**, pp. 355–369.

[43] T. Kamae, U. Krengel, and G.L. O'Brien (1977), "Stochastic inequalities on partially ordered spaces," *Annals of Probability* **5**, pp. 899–912.

[44] S. Karlin and H.M. Taylor (1975), *A First Course in Stochastic Processes*, 2nd ed. Academic Press, New York.

[45] J. Keilson and L.D. Servi (1986), "Oscillating random walk models for $GI/G/1$ vacation systems with Bernoulli schedules," *Journal of Applied Probability* **23**, pp. 790–802.

[46] O. Kella and W. Whitt (1989), "Heavy–traffic limits for queues with server vacations," submitted for publication.

[47] J.F.C. Kingman (1961), "The single server queue in heavy traffic," *Proceedings of the Cambridge Philosophical Society* **57**, pp. 902–904.

[48] J.F.C. Kingman (1962), "On queues in heavy traffic," *Journal of the Royal Statistical Society, Series B* **24**, pp. 383–392.

[49] L. Kleinrock (1976), *Queueing Systems Volume II: Computer Applications*, Wiley, New York.

[50] A.G. Konheim and B. Meister (1974), "Waiting lines and times in a system with polling," *Journal of the Association for Computing Machinery* **21** (3), pp. 470–490.

[51] P.J. Kuehn (1979), "Multiqueue systems with nonexhaustive cyclic service," *Bell System Technical Journal* **58** (3), pp. 671–698.

[52] E. Kyprianou (1971), "The virtual waiting time of the $GI/G/1$ queue in heavy traffic," *Advances in Applied Probability* **3**, pp. 249–268.

[53] A. M. Law and W. D. Kelton (1982), *Simulation Modeling and Analysis*, McGraw–Hill, New York.

[54] A.J. Lemoine (1978), "Networks of queues—A survey of weak convergence results," *Management Science* **24**, pp. 1175–1193.

[55] H. Levy (1988), "Analysis of cyclic–polling systems with binomial–gated service," in *Performance of Distributed and Parallel Systems*, T. Hasegawa, H. Takagi, Y. Takahashi, Editors, North–Holland, Amsterdam, pp. 127–139.

[56] H. Levy and L. Kleinrock (1986), "A queue with starter and a queue with vacations: delay analysis by decomposition," *Operations Research* **34** (3), pp. 426–436.

[57] H. Levy, M. Sidi and O.J. Boxma (1990), "Dominance relations in polling systems," *Queueing Systems—Theory & Applications* **6**, pp. 155–171.

[58] Y. Levy and U. Yechiali (1975), "Utilization of idle time in an $M/G/1$ queueing system," *Management Science* **22**, pp. 202–211.

[59] D. Lindley (1952), "Theory of queues with a single server," *Proceedings of the Cambridge Philosophical Society* **48**, pp. 277–289.

[60] R.M. Loynes (1962), "The stability of a queue with non–independent interarrival and service times," *Proceedings of the Cambridge Philosophical Society* **58**, pp. 497–520.

[61] C. Mack (1957), "The efficiency of N machines uni–directionally patrolled by one operative when walking time is constant and repair times are variable," *Journal of the Royal Statistical Society, Series B* **19** (1), pp. 173–178.

[62] C. Mack, T. Murphy and N.L. Webb (1957), "The efficiency of N machines uni–directionally patrolled by one operative when walking time is constant and repair times are constants," *Journal of the Royal Statistical Society, Series B* **19** (1), pp. 166–172.

[63] G.L. O'Brien (1975), "The comparison method for stochastic processes," *Annals of Probability* **3**, pp. 80–88.

[64] Yu. V. Prohorov (1956), "Convergence of random processes and limit theorems in probability theory," *Theory of Probability and Its Applications* **1**, pp. 157–214.

[65] Yu. V. Prohorov (1963), "Transient phenomena in processes of mass service," *Litovskii Matematicheskii Sbornik* **3**, pp. 199–205.

[66] R. Ramaswamy and L.D. Servi (1988), "The busy period of the M/G/1 vacation model with a Bernoulli schedule," *Stochastic Models* **4**, pp. 507–521.

[67] M.I. Reiman and B. Simon (1989), "An interpolation approximation for queueing systems with Poisson input," *Operations Research* **36** (3), pp. 454–469.

[68] M.I. Reiman and B. Simon (1989), "Open queueing system in light traffic," *Mathematics of Operations Research* **14** (1), pp. 26–59.

[69] S.M. Ross (1983), *Stochastic Processes,* John Wiley and Sons, New York.

[70] I. Rubin and L.F.M. de Moraes (1983), "Message delay analysis for polling and token multiple–access schemes for local communication networks," *IEEE Journal on Selected Areas in Communications* **SAC-1** (5), pp. 935–947.

[71] M. Scholl and L. Kleinrock (1983), "On the $M/G/1$ queue with rest period and certain service–independent queueing discipline," *Operations Research* **31**, pp. 705–719.

[72] L.D. Servi (1986), "Average delay approximation of *M/G/1* cyclic service queues with Bernoulli schedules," *IEEE Journal on Selected Areas in Communications* **SAC-4**, pp. 813–822.

[73] L.D. Servi and D.D. Yao (1989), "Stochastic bounds for queueing systems with limited service schedules," *Performance Evaluation* **9**, pp. 247–261.

[74] K.C. Sevcik and M.J. Johnson (1987), "Cycle time properties of the FDDI token ring protocol," *IEEE Transactions on Software Engineering* **SE–13** (3), pp. 376–385.

[75] W. Szpankowski (1990), "Towards computable stability criteria for some multidimensional stochastic process," in *Stochastic Analysis of Computer and Communication Systems,* H. Takagi, Editor, North–Holland, Amsterdam, pp. 131–172.

[76] W. Szpankowski and V. Rego (1987), "Ultimate stability conditions for some multidimensional distributed systems," Technical Report, Department of Computer Science, Purdue University, West Lafayette, Indiana.

[77] D. Stoyan (1983), *Comparison methods for queues and other stochastic models,* English Translation (D. J. Daley, Editor), John Wiley and Sons, New York.

[78] G.B. Swartz (1980), "Polling in a loop system," *Journal of the Association for Computing Machinery* **27** (1), pp. 42–59.

[79] J.S. Sykes (1970), "Simplified analysis of an alternating–priority queueing model with setup times," *Operations Research* **18** (6), pp. 1182–1192.

[80] H. Takagi (1985), "Mean message waiting times in symmetric multi–queue systems with cyclic service," *Performance Evaluation* **5** (4), pp. 271—277.

[81] H. Takagi (1986), *Analysis of Polling Systems*, MIT Press, Cambridge, MA.

[82] H. Takagi (1988), "Queueing analysis of polling models," *ACM Computing Surveys* **20** (1), pp. 5–28.

[83] H. Takagi (1990), "Queueing analysis of polling models: An update," in *Stochastic Analysis of Computer and Communication Systems,* H. Takagi, Editor, North–Holland, Amsterdam, pp. 267–318.

[84] Tedijanto (1988), "Stochastic comparison in vacation models," International Workshop on the Analysis of Polling Models, December 1988, Kyoto, Japan. Also to appear in *Stochastic Models* .

[85] Tedijanto (1990), "Exact results for the cyclic–service queue with a Bernoulli schedule," to appear in *Performance Evaluation* .

[86] Tedijanto (1990), "A note on the comparison between Bernoulli and limited policies in vacation models," submitted to *Performance Evaluation* .

[87] O. Viskov (1964), "Two asymptotic formulae in the theory of mass service," *Theory of Probability and Its Applications* **9**, pp. 177–178.

[88] K.S. Watson (1984), "Performance evaluation of cyclic service strategies—A survey," in *Performance '84*, E. Gelenbe, Editor, North–Holland, Amsterdam, pp. 521–533.

[89] W. Whitt (1968), *Weak Convergence Theorems for Queues in Heavy Traffic*, Ph.D Thesis, Cornell University.

[90] W. Whitt (1970), "Weak Convergence of Probability Measures on the Function Space $D[0,\infty)$," Technical Report, Department of Administrative Sciences, Yale University.

[91] W. Whitt (1974), "Heavy traffic limit theorems for queues: A survey," in *Mathematical Methods in Queueing Theory*, A. B. Clarke, Editor, Springer–Verlag, Berlin, pp. 307–350.

[92] W. Whitt (1980), "Some useful functions for functional limit theorems," *Mathematics of Operations Research* **5** (1), pp. 67–85.