VACATION QUEUEING MODELS

Theory and Applications

Naishuo Tian
Zhe George Zhang

Springer's INTERNATIONAL SERIES

Vacation Queueing Models
Theory and Applications

Recent titles in the **INTERNATIONAL SERIES IN OPERATIONS RESEARCH & MANAGEMENT SCIENCE**

Frederick S. Hillier, Series Editor, *Stanford University*

Talluri & van Ryzin/ *THE THEORY AND PRACTICE OF REVENUE MANAGEMENT*

Kavadias & Loch/*PROJECT SELECTION UNDER UNCERTAINTY: Dynamically Allocating Resources to Maximize Value*

Brandeau, Sainfort & Pierskalla/ *OPERATIONS RESEARCH AND HEALTH CARE: A Handbook of Methods and Applications*

Cooper, Seiford & Zhu/ *HANDBOOK OF DATA ENVELOPMENT ANALYSIS: Models and Methods*

Luenberger/ *LINEAR AND NONLINEAR PROGRAMMING, 2nd Ed.*

Sherbrooke/ *OPTIMAL INVENTORY MODELING OF SYSTEMS: Multi-Echelon Techniques, Second Edition*

Chu, Leung, Hui & Cheung/ *4th PARTY CYBER LOGISTICS FOR AIR CARGO*

Simchi-Levi, Wu & Shen/ *HANDBOOK OF QUANTITATIVE SUPPLY CHAIN ANALYSIS: Modeling in the E-Business Era*

Gass & Assad/ *AN ANNOTATED TIMELINE OF OPERATIONS RESEARCH: An Informal History*

Greenberg/ *TUTORIALS ON EMERGING METHODOLOGIES AND APPLICATIONS IN OPERATIONS RESEARCH*

Weber/ *UNCERTAINTY IN THE ELECTRIC POWER INDUSTRY: Methods and Models for Decision Support*

Figueira, Greco & Ehrgott/ *MULTIPLE CRITERIA DECISION ANALYSIS: State of the Art Surveys*

Reveliotis/ *REAL-TIME MANAGEMENT OF RESOURCE ALLOCATIONS SYSTEMS: A Discrete Event Systems Approach*

Kall & Mayer/ *STOCHASTIC LINEAR PROGRAMMING: Models, Theory, and Computation*

Sethi, Yan & Zhang/ *INVENTORY AND SUPPLY CHAIN MANAGEMENT WITH FORECAST UPDATES*

Cox/ *QUANTITATIVE HEALTH RISK ANALYSIS METHODS: Modeling the Human Health Impacts of Antibiotics Used in Food Animals*

Ching & Ng/ *MARKOV CHAINS: Models, Algorithms and Applications*

Li & Sun/ *NONLINEAR INTEGER PROGRAMMING*

Kaliszewski/ *SOFT COMPUTING FOR COMPLEX MULTIPLE CRITERIA DECISION MAKING*

Bouyssou et al/ *EVALUATION AND DECISION MODELS WITH MULTIPLE CRITERIA: Stepping stones for the analyst*

Blecker & Friedrich/ *MASS CUSTOMIZATION: Challenges and Solutions*

Appa, Pitsoulis & Williams/ *HANDBOOK ON MODELLING FOR DISCRETE OPTIMIZATION*

Herrmann/ *HANDBOOK OF PRODUCTION SCHEDULING*

Axsäter/ *INVENTORY CONTROL, 2nd Ed.*

Hall/ *PATIENT FLOW: Reducing Delay in Healthcare Delivery*

Józefowska & Węglarz/ *PERSPECTIVES IN MODERN PROJECT SCHEDULING*

**** A list of the early publications in the series is at the end of the book ****

Vacation Queueing Models
Theory and Applications

Naishuo Tian Zhe George Zhang

Springer

Naishuo Tian
Yanshan University
Qinhuangdao, China

Zhe George Zhang
Western Washington University
Bellingham, WA, USA

Library of Congress Control Number: 2006924559

ISBN-10: 0-387-33721-0 (HB) ISBN-10: 0-387-33723-7 (e-book)
ISBN-13: 978-0387-33721-0 (HB) ISBN-13: 978-0387-33723-4 (e-book)

Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1

springer.com

Contents

Preface

In the early twentieth century, A. K. Erlang's works on probability problems in telephone systems laid the groundwork for the development of queueing theory. During the past 100 years, queueing theory has always been one of the most important and active research areas in operations research and applied probability. Classical queueing theory has been well developed and applied as a fundamental performance evaluation tool in many fields such as computer and telecommunication, manufacturing and service, and transportation systems.

Since the mid-20th century, due to the rapid advance of computer technology, flexible manufacturing systems, telecommunication networks, and supply chain systems have been becoming more and more popular in many organizations. To evaluate and eventually improve the performance and efficiency, queueing models were developed to analyze the operations of these hi-tech systems. However, due to the increasing complexity of these stochastic systems, classical queueing theory, which was once quite successful in modeling telephone systems, became inadequate. Vacation queueing theory was developed in the 1970's as an extension of the classical queueing theory. In a queueing system with vacations, other than serving randomly arriving customers, the server is allowed to take vacations. The vacations may represent server's working on some supplementary jobs, performing server maintenance inspection and repairs, or server's failures that interrupt the customer service. Furthermore, allowing servers to take vacations makes queueing models more flexible in finding optimal service policies. Therefore, queues with vacations or simply called *vacation models* attracted great attentions of queueing researchers and became an active research area. Many studies on vacation models were published from the 1970's to the mid 1980's, and were summarized in two survey papers by Doshi and Teghem, respectively, in 1986. Stochastic decomposition theorems were established as the core of vacation queueing theory. In the early 1990's, Takagi published a set of three volume books entitled Queueing Analysis. One of Takagi's books was devoted to vacation models of both continuous and discrete time types and focus mainly on M/G/1 type and Geo/G/1 type queues with vacations. Takagi's book certainly advanced further research and wide applications of vacation models. In another book entitled Frontiers in Queueing edited by Dshalalow in 1997, various M/G/1 type vacation models were discussed as a category of queueing systems with state-dependent parameters.

The aim of this book is to provide an updated and comprehensive treatment of various vacation queueing systems including not only single-server vacation models of both M/G/1 and GI/M/1 types but also a variety of multiserver vacation models. There are several features of this book. Firstly, unconditional and conditional stochastic decomposition properties of stationary performance measures for all types of vacation models are established as the core of vacation queueing theory. Secondly, both performance evaluation and optimal control issues are addressed. In particular, the static and dynamic optimizations in vacation models are discussed. Finally, several practical systems are presented as a sample of wide applications of vacation models. The authors hope that

this book will facilitate further research and applications of vacation queueing models.

The book consists of eight chapters. Chapter 1 gives an introduction to vacation queueing models. The major components of a vacation model, the vacation policies, and the stochastic decomposition structures are described in this chapter. In Chapter 2, M/G/1 type vacation systems with exhaustive service are treated. This type of vacation model has been studied by many researchers using different methods. The system with multiple adaptive vacations is presented in details as a general model of this category. Some well-studied vacation models such as multiple vacation, single vacation, and setup time models are special cases of this general model. Batch arrival and batch service vacation models are discussed in this chapter. Other vacation models with finite buffer, threshold policy, and Markov arrival process (MAP) are also considered. Chapter 3 focuses on M/G/1 type vacation systems with non-exhaustive service including gated service, limited service, decremental service, and Bernoulli service. This chapter is mainly based on the materials from Takagi's book Queueing Analysis, Volume 1. Chapter 4 is devoted to GI/M/1 type vacation models. Compared to M/G/1 type vacation models which are analyzed by mainly using embedded Markov chain and supplementary variable methods, GI/M/1 type vacation models are treated by using the matrix analytical method developed by Neuts (see Neuts 1981). Some recent results about finite buffer or batch service GI/M/1 type vacation systems are also reported. In Chapter 5, Markovian multiserver vacation models are discussed. Multiserver vacation systems with various service policies are modelled as quasi-birth-and-death (QBD) processes and analyzed by using the matrix geometric solution method. Similar to unconditional stochastic decomposition properties in single-server vacation models, conditional stochastic decomposition properties when all servers are busy are established for multiserver models. Chapter 6 studies multiserver vacation models with general arrival process or of GI/M/c type. The stationary performance measures and the conditional stochastic decomposition properties are presented. In Chapter 7, the optimal control issue in vacation systems is addressed. For single-server vacation systems, both static optimization and dynamic control models under certain cost and revenue structures are developed. Searching method and proof of convexity for average cost function are presented in this chapter. Markov decision process is used to solve the dynamic control problems in single-server systems. For multiserver vacation systems with given cost and revenue structures, the optimal threshold policies are obtained by using the searching method. Finally, Chapter 8 provides a few examples that illustrate the applications of vacation models. A bibliographic notes is given at the end of each chapter.

Although the book contains a variety of vacation models that have been studied over the past thirty years, there are still some excellent past works, many successful applications, and open problems that are not included in this book. The topics that need further research include (but are not limited to) the diffusion approximation models, the queueing networks with vacations, the simulation-based models, and the multiserver vacation models with Markov arrival process.

Acknowledgements

We would like to thank our friends and colleagues at Western Washington University and Simon Fraser University for their support over the years and wish particularly to mention Drs. Floyd Lewis and Peter Haug at WWU and Drs. Ernie Love, Art Warburton, Eng Choo, and William Wedley at SFU. We are grateful to Gary Folven, our editor at Springer Science and Dr. Frederick S. Hillier, the series editor, for their support and encouragement; the editor assistant Carolyn Ford for her gracious and careful attention to the book's production. Our thanks also go to Deborah Doherty for her technical assistance of typesetting the manuscript and Gerry Geer for copyediting the manuscript.

The support from the research grants of BFR and CBE at Western Washington University, the NSERC research grant of Canada, and research grants of National Natural Science Foundation of China (No.19471012, 19871072 and 10271102) are gratefully acknowledged.

Finally, I wish to express my deepest appreciation to my wife Siping Sue, my daughters Nancy and Lucy for their love, understanding, and belief in me. I would also like to thank my parents Yuwen and Maoxi for their support and love. Nanshuo wishes to thank his wife Guihua Yang for her constant support and love.

Zhe George Zhang

At WWU/SFU

Chapter 1

INTRODUCTION

1.1 Queueing Systems with Server Vacations

In a classical queueing model, servers are always available. However, in many practical queueing systems, servers may become unavailable for a period of time due to a variety of reasons. This period of server absence may represent the server's working on some supplementary jobs, being checked for maintenance, or simply taking a break. To analyze these systems, we introduce the *server vacation* in queueing models to represent the period of temporary server absence. Allowing servers to take vacations makes queueing models more realistic and flexible in the study of real-world waiting-line systems. Below are some practical systems that can be modeled as queues with vacations.

Example 1.1 (call centers with multitask employees). The customer service hotline of a long distance calling card company may not be very busy all the time. The customer service representative's (CSR) main task is to answer customer calls for assistance. During the idle time, the CSR can make phone calls to potential customers to promote the company's service and products. In this situation, the inbound calls are queueing customers and the outbound calls are supplementary jobs that can be modeled as server vacations. A call center with multitask CSRs can be represented by a multiserver vacation model with "inbound calls" as customers and "outbound calls" as vacations.

Example 1.2 (Border-crossing stations). In a U.S.and Canada border-crossing station, the number of open lanes is determined by the level of congestion or the length of the waiting line of cars. When the queue length becomes zero, some of the open lanes are closed and the inspectors leave for other jobs. When the waiting line builds up to a certain

limit, these closed lanes are reopened to reduce the congestion level. In this situation, time spent on working on other jobs is considered to be a server vacation.

Example 1.3 (mixture of make-to-order and make-to-stock operations). A flexible manufacturing facility is mainly used for producing customer-specified products. When there are no customer backorders, the facility switches over to produce a variety of items in stock. Due to the considerable switchover cost between "make-to-order" and "make-to-stock" the facility is not switched back to process customer orders until the number of orders is more than a critical level. Once the facility switches back to serving customer orders, the service is exhaustive. In this system, the "make-to-order" operation is a queue service process and the "make-to-stock" operation can be modeled as a server vacation.

Example 1.4 (data transfer in computer/telecommunication networks). In an SVC (switched virtual connection)-based IP-over-ATM (asynchronous transfer mode) network, the SVC manager or IP controller can be considered to be as a server of a queueing model. The setup time corresponds to the time period needed to set up a new SVC by means of signaling protocols, and the shutdown time corresponds to an inactive time period during which the SVC resources (e.g., routing information and bandwidth) are reserved in anticipation of more customers (packets) from the same IP flow. The vacation time may be considered to be the time period required to release the SVC or the time during which the server sets up other SVCs.

Example 1.5 (maintenance activities as server vacations). Another example is the "repairman" problem in which the repairman's main duty is to repair broken machines. When no broken or malfunctioning machine exists, the repairman can do some maintenance or inspection jobs. In this situation, the broken machines are the customers forming a queue and the maintenance and inspection jobs are considered to be server vacations.

Many real-world systems can be modeled as queues with different vacation policies. Since the mid-twentieth century, due to the fast development of computer and communication networks and flexible manufacturing systems, the issue of performance evaluation and optimal control for these systems has become more and more important to users. Queueing models with vacations have been developed as useful performance analysis tools for these high-tech systems. Classical queueing models without vacations are not adequate for systems where servers may not be always available. Although, in the classical literature, queueing researchers have addressed some complex systems with polling service and priority service, most vacation queueing models have been studied and

reported only since the 1970s. Incorporating server vacations into queueing models reflects the fact that server(s) may become unavailable while working on secondary jobs in many practical queueing systems.

1.2 Vacation Policies

A classical queueing model consists of three parts: the arrival process, the service process, and queue discipline (see Gross and Harris (1985)). A vacation queueing model has an additional part: a vacation process governed by a vacation policy. A vacation policy can be characterized by three aspects:

(1) Vacation startup rule. This rule determines when the server starts a vacation. There are two major types, namely, exhaustive and nonexhaustive services. With an exhaustive service, the server cannot take a vacation until the system becomes empty. On the other hand, the server in a nonexhaustive service system can take a vacation even when the system is not empty. In a multiserver system, a semiexhaustive service rule may be used if some of the servers take vacations. Another vacation start-up rule is the service interruption during the progress of customer service. The service interruption may represent a machine failure during the operation.

(2) Vacation termination rule. This rule determines when the server resumes serving the queue. Two popular rules are the multiple vacation policy and the single vacation policy. A multiple vacation policy requires the server to keep taking vacations until it finds at least one customer waiting in the system at a vacation completion instant. In contrast, under a single vacation policy, the server takes only one vacation at the end of each busy period. After this single vacation, the server either serves the waiting customers, if any, or stays idle. More general rules, such as the threshold policy (also called N-policy) and the adaptive multiple vacation policy, will also be discussed in this book. In nonexhaustive service systems, more vacation termination rules are possible.

In multiserver systems, in addition to start-up and termination rules, there are other characteristics of a vacation policy. For example, all servers may take vacations together (synchronous vacations), or servers may take vacations individually and independently (asynchronous vacations). Another possible feature of a vacation policy is to allow some (but not all) servers to take vacations to ensure that at least a minimum number of servers are always available.

(3) Vacation duration distribution. Server vacations are often assumed to be independent and identically distributed (i.i.d.) random variables with a general distribution function, denoted by $V(x)$. How-

ever, some vacation models require different types of vacations and follow different distributions.

The many variations on the vacation policy will be discussed in this book.

1.3 Stochastic Decomposition in Vacation Models

The fundamental result of vacation models is the stochastic decomposition theorem. In most queueing systems with vacations, the stationary queue length or the stationary waiting time can be decomposed into the sum of two independent random variables. One of these is the queue length or waiting time of the corresponding classical queueing system without vacations, and the other is the additional queue length or delay due to vacations. These variables show clearly the effects of vacations on system performance. For a classical single-server queueing system that has reached the steady state, denote the number of customers in the system, the queue length, and the waiting time by L, Q, W, respectively, and denote the same performance measures by L_v, Q_v, W_v, respectively, for the corresponding steady-state vacation system. Let $X(z)$ and $X^*(s)$ be the z-transform, or probability generating function (p.g.f.), and the Laplace-Stieltjes transform (LST), respectively, of the stationary random variable X. With these notations, the stochastic decomposition properties can be written as

$$\begin{aligned} L_v &= L + L_d, \qquad & L_v(z) &= L(z)L_d(z), \\ Q_v &= Q + Q_d, \qquad & Q_v(z) &= Q(z)Q_d(z), \\ W_v &= W + W_d, \qquad & W_v^*(s) &= W^*(s)W_d^*(s), \end{aligned}$$

where L_d, Q_d, and W_d are the additional number of customers in the system, the additional queue length, and the additional delay, respectively, due to vacations. For M/G/1 type vacation systems, the stochastic decomposition properties have been proved by many researchers using different methods. Doshi (1985) presented the stochastic decomposition theorem for $GI/G/1$ type queues with vacations. Two excellent survey papers by Doshi (1986) and Teghem (1986) primarily focused on the stochastic decomposition properties in single server vacation models. Tian et al. (1989, 1990, 1993) studied GI/M/1 type queues with vacations and established the stochastic decomposition theorems. These stochastic decomposition theorems laid the foundation of analyzing single server vacation systems.

To expand the applications of vacation models, multiserver queues with vacations were also studied after numerous achievements in single server vacation models. However, it seems extremely difficult to estab-

lish the unconditional stochastic decomposition properties in multiserver models. When all servers in a multiserver system are busy, the conditional stochastic decomposition properties can be obtained. Consider a classical multiserver queue with c servers, and let J be the number of busy servers in a steady state. Define

$$Q_v^{(c)} = \{L_v - c | J = c\},\ W_v^{(c)} = \{W_v | L_v \geq c,\ J = c\}.$$

$Q_v^{(c)}$ is the number of customers waiting in line given that all servers are busy, and $Q^{(c)}$ is the same random variable for the corresponding queueing system without vacations. $W_v^{(c)}$ is the customer waiting time, given that all server are busy, and $W^{(c)}$ is the same random variable for the corresponding queueing system without vacations. The conditional stochastic decomposition properties are as follows:

$$\begin{aligned} Q_v^{(c)} &= Q^{(c)} + Q_d, \qquad & Q_v^{(c)}(z) &= Q^{(c)}(z)Q_d(z), \\ W_v^{(c)} &= W^{(c)} + W_d, \qquad & W_v^{(c)*}(s) &= W^{(c)*}(s)W_d^*(s), \end{aligned}$$

where Q_d and W_d are the additional queue length and additional delay due to server vacations, respectively.

These stochastic decomposition properties indicate the effects of vacations on system performance and play an important role in vacation model theory. In this book, we discuss various stochastic decomposition theorems as the fundamental theory of vacation models.

1.4 Bibliographic Notes

Since the early work by Erlang (1918) on modeling telephone traffic systems, queueing theory has been developed over almost 100 years. Due to its wide practical applications in many areas, queueing theory has been one of the most active research topics in operations research and management science over the past several decades. Some excellent books on classical queueing theory have been published, including these by Takacs (1962), Kleinrock (1975), Cooper (1981), Cohen (1982), Gross and Harris (1985), Saaty (1983), Wolff (1989), Prabhu (1997), and others. Some of the early work on queueing systems is relevant to queues with vacations. White and Christie (1958) studied queueing system with priority services and server breakdowns. Welch (1964) examined the system with exceptional service to the first customer starting a busy period. Jaiswal (1968) and Avi-Itzhak and Naor (1963) considered queues with server interruptions and different service-resumption priority rules. Cooper (1970) presented a study on queues served in a cyclic order, in which the time period of serving other queues can be considered a service interruption of the queue under consideration. However, significant

research results on vacation systems were published in the late twentieth century. Levy and Yechiali (1975) studied the issue of efficiently utilizing server idle time and introduced the concept of a server's taking vacations that represent the durations of the server's work on some supplementary project. Stochastic decomposition properties were discovered by Levy and Yechiali (1975). Afterwards, many research results on vacation models were published, including these by Courtois (1980), Fuhrmann (1984), Fuhrmann and Cooper (1985), Doshi (1985), Levy and Kleinrock (1986), Teghem (1985), Doshi (1990), Dshalalow (1997), etc. In these works, detailed analysis and stochastic decomposition theorems for M/G/1 type systems have been presented. Two excellent survey papers (Taghem (1986) and Doshi (1986)) summarized the major developments in this area. There are also a few books that contain chapters or sections on vacation models. Medhi (1991) discussed the M/G/1 queue with vacations. Takagi (1991,1993) published a set of books that provide a complete analysis of M/G/1 type and Geo/G/1 type vacation systems.

Stochastic decomposition properties were first observed in some early queueing studies, such as those by Gaver (1962), Miller (1964), Cooper (1970), and Levy and Yechiali (1975). After Levy and Yechiali's work, the stochastic decomposition theorems became the focus of most research papers including those of Shanthikumar (1980), Scholl and Kleinrock (1983), Ali and Neuts (1984), Neuts and Ramalhoto (1984), and Federgrune and Green (1986). Doshi (1985) extended the stochastic decomposition property for stationary waiting time into a GI/G/1 queue with vacations. Shanthikumar (1988, 1989) provided a proof for the stochastic decomposition theorem in an M/G/1 queue with a class of more general vacation policies. Takine and Hasegawa (1992) presented a stochastic decomposition property for the joint distribution of number of customers and elapsed service time. Rosberg and Gail (1991) studied the relationship between stochastic decomposition properties and PASTA. Keilson and Servi (1990) discussed the relationship between Little's law and stochastic decomposition in vacation models. Miyazawa (1994) used the work-conservation law to provide a unified treatment of various M/G/1 vacation models and established the stochastic decomposition theorems.

Tian et al. (1989) studied the GI/M/1 queue with exponentially distributed vacations and established the stochastic decomposition properties for stationary queue length and waiting time. Recently, Tian and Zhang (2003b) extended these properties to a GI/M/1 queue with PH-type setup times or vacations.

For multiserver vacation models, it has been proved by Tian et al. (1999), Zhang and Tian (2003a), and Tian and Zhang (2003a, 2003b) that there exists a set of conditional stochastic decomposition properties

for stationary queue length and waiting time, given that all servers are busy in a variety of M/M/c and GI/M/c type systems with different vacation policies.

Chapter 2

M/G/1 TYPE VACATION MODELS: EXHAUSTIVE SERVICE

This chapter focuses on single server vacation systems where the server follows an exhaustive-service policy: in other words, the server does not take any vacations until the system becomes empty. The systems considered are the M/G/1 type, where interarrival times are exponentially distributed i.i.d. random variables and service times are generally distributed i.i.d.random variables. The rules for resuming queue service at a vacation completion instant are numerous. However, they can be generally classified into two categories. The rules in the first category are mainly based on the number of vacations taken before the first customer arrives at the empty system. These rules usually require the server to serve the queue at a vacation completion instant if waiting customers exist. The rules in the second category are based on the number of waiting customers at a vacation completion instant. If the server returns to serve the queue only when the number of waiting customers reaches a critical value, the rule is called a *threshold policy*. In section 2.1, we consider the multiple adaptive vacation (MAV) policy, a general rule of the first category. In section 2.2, we demonstrate that several common vacation models are special cases of the MAV policy model. The threshold policy models are presented in section 2.3. Other variations of the M/G/1 type exhaustive-service models are also discussed in this chapter. Specifically, the discrete-time vacation models are presented in section 2.4. Vacation models with Markov arrival process (MAP) are considered in section 2.5. Vacation models with batch arrivals or batch services are discussed in section 2.6. Finally, the finite-buffer vacation models are given in section 2.7.

2.1 M/G/1 Queue with Multiple Adaptive Vacations

2.1.1 Classical M/G/1 Queue

We first present briefly some well-known results for a classical M/G/1 queue without vacations. The details of developing these results can be found in any queueing theory books (for example, see Gross and Harris (1985)). In such a system, customers arrive according to a Poisson process with rate λ and service times are i.i.d random variables with a general distribution function, denoted by $B(t)$. Let

$$\frac{1}{\mu} = \int_0^\infty t\mathrm{d}B(t),\ b^{(2)} = \int_0^\infty t^2\mathrm{d}B(t),\ B^*(s) = \int_0^\infty e^{-st}\mathrm{d}B(t).$$

Assume that the service order is first-come-first-served (FCFS) and that interarrival times and service times are independent.

Denote by L_n the number of customers in the system at the nth customer departure instant, $\{L_n,\, n \geq 1\}$ is an embedded Markov chain of the queueing process, satisfying

$$L_{n+1} = \begin{cases} L_n - 1 + A_{n+1}, & L_n \geq 1, \\ A_{n+1}, & L_n = 0, \end{cases}$$

where A_{n+1} is the number of arrivals during the $(n+1)$ service time. Obviously these numbers are i.i.d. random variables and can be denoted by A, with respective probability distribution and mean

$$a_j = P(A = j) = \int_0^\infty \frac{(\lambda t)^j}{j!} e^{-\lambda t}\mathrm{d}B(t),\ j \geq 0, \qquad E(A) = \frac{\lambda}{\mu} = \rho.$$

ρ is called the *traffic intensity* of the system and is the ratio of arrival rate to service rate. The probability generating function (p.g.f.) of A is $A(z) = B^*(\lambda(1-z))$, and the transition probability matrix of the embedded Markov chain is

$$\mathbf{P} = \begin{bmatrix} a_0 & a_1 & a_2 & a_3 & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots \\ & a_0 & a_1 & a_2 & \cdots \\ & & a_0 & a_1 & \cdots \\ & & & \vdots & \vdots \end{bmatrix}. \tag{2.1.1}$$

It can be proved that $\{L_n,\, n \geq 1\}$ is positive recurrent and the system reaches the steady state if and only if $\rho < 1$. Therefore, when $\rho < 1$, the p.g.f.s of the stationary number of customers in the system, L, and the

stationary number of customers waiting in line, Q, and the LST of the stationary waiting time, W, are as follows:

$$\begin{aligned} L(z) &= \frac{(1-\rho)(1-z)B^*(\lambda(1-z))}{B^*(\lambda(1-z))-z}, \\ Q(z) &= \frac{(1-\rho)(1-z)}{B^*(\lambda(1-z))-z}, \\ W^*(s) &= \frac{(1-\rho)s}{s-\lambda(1-B^*(s))}. \end{aligned} \tag{2.1.2}$$

The means of these stationary random variables are, respectively,

$$\begin{aligned} E(L) &= \rho + \frac{\lambda^2 b^{(2)}}{2(1-\rho)}, \\ E(Q) &= \frac{\lambda^2 b^{(2)}}{2(1-\rho)}, \\ E(W) &= \frac{\lambda b^{(2)}}{2(1-\rho)} = \frac{1}{\lambda}E(Q). \end{aligned} \tag{2.1.3}$$

These formulas are called *Pollaczek-Khinthin formulas*. Note that (2.1.2) gives the p.g.f. of the queue length distribution at a customer departure instant, called the *departure distribution*. It can be shown that the departure distribution is the same as the distribution seen by an arriving customer, called the *arrival distribution*. Furthermore, due to the well-known Poisson Arrivals See Time Averages (PASTA) property (see Wolff (1982)), the arrival distribution is the same as the distribution of the queue length at any time t. Therefore, the departure distribution obtained in (2.1.2) is the same as the distribution at any time. This important property holds in all M/G/1 vacation models discussed in this chapter.

A busy period, denoted by D, is defined as the period from the arrival instant of the first customer at an empty system to the departure instant of a customer that leaves an empty system. It is well known that the LST of D satisfies the functional relation

$$D^*(s) = B^*(s + \lambda(1 - D^*(s))).$$

Based on this relation, the mean of the busy period is obtained as

$$E(D) = \frac{1}{\mu(1-\rho)} = \frac{1}{\lambda-\mu}. \tag{2.1.4}$$

2.1.2 Multiple Adaptive Vacation Model

In an M/G/1 queue, the server follows the following vacation policy. When the server finishes serving all customers in the system, it starts to take a vacation. The server will take vacations consecutively until either a customer has arrived at a vacation completion instant or a maximum number, denoted by H, of vacations have been taken. In the case of arrivals occurred during a vacation, the server resumes serving the queue immediately at that vacation completion instant. In the case of no arrivals occuring after the server has completed H vacations, the server stays idle and waits to serve the next arrival. H, called the *stages of vacations*, is assumed to be a discrete random variable, with respective distribution and p.g.f.

$$P\{H = j\} = h_j,\ j \geq 1; \quad H(z) = \sum_{j=1}^{\infty} h_j z^j.$$

The consecutive vacations, denoted by V_k, $k = 1, 2, ..., H$, are i.i.d. random variables with the distribution function of $V(x)$, the LST of $v^*(s)$, and the finite first and second moments. The queueing system of this policy is called a *vacation model with exhaustive service*, *multiple adaptive vacations* (*MAV*), or simply an *E-MAV model*, denoted by M/G/1 (E, MAV). The E-MAV policy reflects the flexibility of allowing the server to work on both the primary randoml-arrival jobs (the queue) and a random number of secondary jobs (the vacations) during the idle time. Assume that the interarrival times, the service times, the vacation times, and the stages of vacations are mutually independent and the service order is FCFS.

Define two events

$$A_I = \{\text{a busy period starts with the ending of an idle period}\},$$
$$A_v = \{\text{a busy period starts with the ending of a vacation}\},$$

we have

$$\begin{aligned} P\{A_I\} &= \sum_{j=1}^{\infty} P\{H = j\} P\{T > V_1 + \cdots + V_j\} \\ &= \sum_{j=1}^{\infty} h_j \int_0^{\infty} e^{-\lambda t} \mathrm{d}V^{(j)}(t) \\ &= \sum_{j=1}^{\infty} h_j [v^*(\lambda)]^j = H[v^*(\lambda)], \end{aligned}$$

where $V^{(j)}(t)$ is the jth convolution of $V(t)$. Obviously,

$$P\{A_v\} = 1 - H[v^*(\lambda)].$$

Letting L_n be the number of customers left behind by the nth customer, we have

$$L_{n+1} = \begin{cases} L_n - 1 + A, & \text{for } L_n \geq 1, \\ Q_b - 1 + A, & \text{for } L_n = 0, \end{cases}$$

where Q_b is the number of customers in the system when a busy period starts. Note that the case of $Q_b = 1$ is for M/G/1 queue without vacations.

Lemma 2.1.1. The p.g.f. and the mean of Q_b are, respectively,

$$Q_b(z) = H[v^*(\lambda)]z + \frac{1 - H[v^*(\lambda)]}{1 - v^*(\lambda)}\{v^*(\lambda(1-z)) - v^*(\lambda)\},$$
$$E(Q_b) = H[v^*(\lambda)] + \frac{1 - H[v^*(\lambda)]}{1 - v^*(\lambda)}\lambda E(V). \tag{2.1.5}$$

Proof: The event $\{Q_b = 1\}$ occurs if either of two mutually exclusive cases happens: (1) the busy period starts with a customer arriving at an idle server; or (2) the busy period starts with the ending of a vacation during which only one customer arrives. Hence, we have

$$P\{Q_b = 1\} = H[v^*(\lambda)] + \frac{1 - H[v^*(\lambda)]}{1 - v^*(\lambda)} v_1,$$

where $v_j = \int_0^\infty \frac{(\lambda t)^j}{j!} e^{-\lambda t} \mathrm{d}V(t)$ is the probability that j customers arrive during a vacation time. For $j \geq 2$, $\{Q_b = j\}$ represents the case in which the busy period starts with the ending of a vacation during which j customers have arrived. Thus,

$$P\{Q_b = j\} = \frac{1 - H[v^*(\lambda)]}{1 - v^*(\lambda)} v_j, \qquad j \geq 2.$$

Taking the p.g.f. of the distribution of Q_b yields $Q_b(z)$ and computing $Q_b'(1)$ gives $E(Q_b)$. □

Under the E-MAV policy, the transition probability matrix of the embedded chain of $\{L_n, n \geq 1\}$ becomes

$$\mathbf{P} = \begin{bmatrix} b_0 & b_1 & b_2 & b_3 & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots \\ & a_0 & a_1 & a_2 & \cdots \\ & & a_0 & a_1 & \cdots \\ & & & \vdots & \vdots \end{bmatrix}, \tag{2.1.6}$$

where

$$b_j = P\{Q_b - 1 + A = j\}$$
$$= H[v^*(\lambda)]a_j + \frac{1 - H[v^*(\lambda)]}{1 - v^*(\lambda)} \sum_{i=1}^{j+1} v_i a_{j+1-i}, \quad j \geq 0. \tag{2.1.7}$$

Similar to the classical M/G/1 queue, from (2.1.6) it can be proved that the embedded chain $\{L_n, n \geq 1\}$ is positive recurrent if and only if $\rho = \lambda\mu^{-1} < 1$. When $\rho < 1$, let L_v be the limiting (or stationary) random variable of L_n as $n \to \infty$, with the stationary distribution

$$\Pi = (\pi_0, \pi_1, \cdots, \pi_n, \cdots),$$

where $\pi_j = P\{L_v = j\} = \lim_{n\to\infty} P\{L_n = j\}$, for $j \geq 0$. We now give the stochastic decomposition property for the stationary queue length.

Theorem 2.1.1. For $\rho < 1$, L_v can be decomposed into the sum of two independent random variables,

$$L_v = L + L_d,$$

where L is the queue length of a classical M/G/1 queue without vacations with its p.g.f. given in (2.1.2). L_d is the additional queue length due to the vacation effect, with the p.g.f.

$$L_d(z) = \frac{1 - Q_b(z)}{E(Q_b)(1 - z)}, \tag{2.1.8}$$

where $Q_b(z)$ is given in Lemma 2.1.1.

Proof: Based on the equilibrium equation of $\Pi\mathbf{P} = \Pi$ and (2.1.6), we have

$$\pi_k = \pi_0 b_k + \sum_{j=1}^{k+1} \pi_j a_{k+1-j}, \qquad k \geq 0. \tag{2.1.9}$$

From (2.1.7), we obtain the p.g.f of $\{b_k, k \geq 0\}$:

$$\sum_{k=0}^{\infty} z^k b_k = \frac{1}{z} B^*(\lambda(1 - z)) Q_b(z).$$

Multiplying both sides of (2.1.9) by z^k and summing over k gives

$$\begin{aligned} L_v(z) &= \sum_{k=0}^{\infty} z^k \pi_k \\ &= \pi_0 \frac{1}{z} B^*(\lambda(1-z)) Q_b(z) + \sum_{k=0}^{\infty} z^k \sum_{j=1}^{k+1} \pi_j a_{k+1-j} \\ &= \pi_0 \frac{1}{z} B^*(\lambda(1-z)) Q_b(z) + \frac{1}{z} B^*(\lambda(1-z))[L_v(z) - \pi_0]. \end{aligned}$$

Solving the equation above for $L_v(z)$, we get

$$L_v(z) = \frac{\pi_0 B^*(\lambda(1-z))[1 - Q_b(z)]}{B^*(\lambda(1-z)) - z}. \tag{2.1.10}$$

Using the normalization condition and the L'Hopital rule, we have

$$\pi_0 = \frac{1-\rho}{E(Q_b)},$$

and substituting it into (2.1.10) gives

$$\begin{aligned} L_v(z) &= \frac{(1-\rho)(1-z)B^*(\lambda(1-z))}{B^*(\lambda(1-z)) - z} \frac{1 - Q_b(z)}{E(Q_b)(1-z)} \\ &= L(z) L_d(z). \end{aligned}$$

This completes the proof. □

Note that $L_d(z)$ in (2.1.8) is a p.g.f of a probability distribution. Define a distribution as

$$q_j = \frac{1}{E(Q_b)} \sum_{n=j+1}^{\infty} P\{Q_b = n\}, \qquad j = 0, 1, ...$$

Then the p.g.f. of $\{q_j,\, j \geq 0\}$ is

$$\begin{aligned} \overline{Q}_b(z) &= \sum_{j=0}^{\infty} q_j z^j \\ &= \frac{1}{E(Q_b)} \sum_{j=0}^{\infty} z^j \sum_{n=j+1}^{\infty} P\{Q_b = n\} \\ &= \frac{1}{E(Q_b)(1-z)} \sum_{n=1}^{\infty} P\{Q_b = n\}(1 - z^n) \\ &= \frac{1 - Q_b(z)}{E(Q_b)(1-z)}. \end{aligned}$$

Based on Theorem 2.1.1, the following expected value formulas are obtained:

$$E(L_d) = \frac{E(Q_b^2)}{2E(Q_b)},$$
$$E(L_v) = \rho + \frac{\lambda^2 b^{(2)}}{2(1-\rho)} + \frac{E(Q_b^2)}{2E(Q_b)}. \tag{2.1.11}$$

Using $Q_b(z)$ in (2.1.5), we have

$$E(Q_b^2) = \frac{1 - H[v^*(\lambda)]}{1 - v^*(\lambda)} \lambda^2 E(V^2).$$

For the stationary waiting time, there exists a similar stochastic decomposition property.

Theorem 2.1.2. For $\rho < 1$, the stationary waiting time, denoted by W_v, can be decomposed into the sum of the two independent random variables,

$$W_v = W + W_d,$$

where W is the waiting time of a classical M/G/1 queue without vacations, with its LST given in (2.1.2). W_d is the additional delay due to the vacation effect, with the LST

$$W_d^*(s) = \frac{H[v^*(\lambda)]}{E(Q_b)} + \frac{\lambda E(V)}{E(Q_b)} \frac{1 - H[v^*(\lambda)]}{1 - v^*(\lambda)} \frac{1 - v^*(s)}{E(V)s}, \tag{2.1.12}$$

where $E(Q_b)$ is given in Lemma 2.1.1.

Proof: Based on the independent increment property of Poisson arrivals and the fact that the number of customers left behind by a departing customer is the same as the number of arrivals during this customer's time (waiting and service) in the system, we have

$$\begin{aligned} L_v(z) &= \sum_{k=0}^{\infty} z^k \int_0^{\infty} \int_0^{\infty} \frac{[\lambda(x+y)]^k}{k!} e^{-\lambda(x+y)} \mathrm{d}W_v(x) \mathrm{d}B(y) \\ &= \int_0^{\infty} \int_0^{\infty} e^{-\lambda(x+y)(1-z)} \mathrm{d}W_v(x) \mathrm{d}B(y) \\ &= W_v^*(\lambda(1-z)) B^*(\lambda(1-z)). \end{aligned}$$

Substituting $L_v(z)$ into the formula above gives

$$W_v^*(\lambda(1-z)) = \frac{(1-\rho)(1-z)}{B^*(\lambda(1-z)) - z} \frac{1 - Q_b(z)}{E(Q_b)(1-z)}. \tag{2.1.13}$$

Letting $\lambda(1-z)=s$, we have

$$\begin{aligned} W_v^*(s) &= \frac{(1-\rho)s}{s-\lambda(1-B^*(s))}\frac{\lambda[1-Q_b(1-\frac{s}{\lambda})]}{E(Q_b)s} \\ &= W^*(s)W_d^*(s). \end{aligned}$$

Using (2.1.2), we find that the additional delay W_d has an LST of

$$W_d^*(s) = \frac{\lambda[1-Q_b(1-\frac{s}{\lambda})]}{E(Q_b)s}. \tag{2.1.14}$$

Substituting $Q_b(z)$ from (2.1.5) into (2.1.14) and simplifying yields (2.1.12). □

Formula (2.1.12) indicates that the additional delay W_d is zero with probability of $p=H[v^*(\lambda)][E(Q_b)]^{-1}$ and is equal to the residual vacation time with probability of $1-p$. It is easy to verify that the number of arrivals during W_d is the additional queue length due to the vacation effect, L_d. The means of the additional delay and the waiting time can be obtained as

$$\begin{aligned} E(W_d) &= \frac{\{1-H[v^*(\lambda)]\}\lambda E(V^2)}{2(1-v^*(\lambda))E(Q_b)}, \\ E(W_v) &= \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{\{1-H[v^*(\lambda)]\}\lambda E(V^2)}{2(1-v^*(\lambda))E(Q_b)}. \end{aligned} \tag{2.1.15}$$

Let us now provide the busy-period analysis of the M/G/1 (E,MAV) model. Denote by D_v the busy period of the vacation system and by D the busy period of the classical M/G/1 system. Note that the only difference between D_v and D is the number of customers present in the system when the busy period starts. Due to the memoryless property of the exponential interarrival times, the busy period starting with k customers in the system is equal to the sum of k independent M/G/1 queue busy periods D. It follows immediately that

$$D_v^*(s) = Q_b[D^*(s)],$$

where $D^*(s)$ is the LST of D. Thus

$$E(D_v) = \frac{1}{\mu(1-\rho)}E(Q_b).$$

Let J be the number of consecutive vacations taken by the server. Based on the MAV policy, we have

$$J = \min\{H, k : V^{(k-1)} < T < V^{(k)}\}.$$

It is easy to verify that

$$P\{J \geq 1\} = 1,$$
$$P\{J \geq j\} = P\{H \geq j\}P\{V^{(j-1)} \geq T\} = [v^*(\lambda)]^{j-1}\sum_{k=j}^{\infty} h_k, \quad j \geq 2.$$

Therefore, we have

$$\sum_{j=1}^{\infty} P\{J \geq j\}z^j = \frac{z(1-J(z))}{1-z}$$
$$= \sum_{j=1}^{\infty} z^j[v^*(\lambda)]^{j-1}\sum_{k=j}^{\infty} h_k = z\frac{1-H[v^*(\lambda)z]}{1-v^*(\lambda)z}.$$

From this relation, we obtain

$$J(z) = 1 - \frac{1-z}{1-v^*(\lambda)z}\{1-H[v^*(\lambda)z]\},$$
$$E(J) = \frac{1-H[v^*(\lambda)]}{1-v^*(\lambda)}.$$

Denote the total length of J consecutive vacations by V_G. Then

$$E(V_G) = E(J)E(V) = \frac{1-H[v^*(\lambda)]}{1-v^*(\lambda)}E(V). \tag{2.1.16}$$

The idle period, denoted by I_v, occurs only when event A_I happens. Hence,

$$E(I_v) = H[v^*(\lambda)]\frac{1}{\lambda}. \tag{2.1.17}$$

Define the busy cycle B_c as the time period between two consecutive busy-period ending instants. Then we have

$$E(B_c) = E(D_v) + E(V_G) + E(I_v)$$
$$= \frac{1}{\mu(1-\rho)}E(Q_b) + \frac{1-H[v^*(\lambda)]}{1-v^*(\lambda)}E(V) + H[v^*(\lambda)]\frac{1}{\lambda}$$
$$= \frac{1}{\lambda(1-\rho)}E(Q_b). \tag{2.1.18}$$

Let p_b, p_v, and p_i be the probabilities of the server's being busy, on vacation, and idle, respectively. We then have

$$\begin{aligned} p_b &= \frac{E(D_v)}{E(B_c)} = \rho, \\ p_v &= \frac{E(V_G)}{E(B_c)} = \frac{1 - H[v^*(\lambda)]}{(1 - v^*(\lambda))E(Q_b)}\lambda(1-\rho)E(V), \\ p_i &= \frac{E(I_v)}{E(B_c)} = \frac{1}{E(Q_b)}(1-\rho)H[v^*(\lambda)]. \end{aligned} \tag{2.1.19}$$

2.2 Some Classical M/G/1 Vacation Models

In this section, we show that several classical vacation models are the special cases of the E-MAV model presented in the previous section.

2.2.1 Multiple Vacation Model

Consider an M/G/1 queue where the server follows an exhaustive-service and multiple vacation (E, MV) policy. This policy requires the server to keep serving customers until the system is empty and then to take vacations for as long as the system is empty. The server returns to serve the queue when there are some customers waiting in the system at a vacation completion instant. This type of system, denoted by M/G/1 (E, MV), has been extensively studied. The multiple vacation policy allows the server to maximize the use of idle time for supplementary work. However, the server does not have any idle time in such a system (where idle time means either serving the queue or being on vacation), if taking a vacation represents doing productive work. Obviously, this situation is the $H = \infty$ case for the E-MAV model.

If $H = \infty$, $H(z) = 0$. From (2.1.5), the busy period starts with Q_b customers in the system. The p.g.f. and the mean of Q_b are, respectively,

$$\begin{aligned} Q_b(z) &= \frac{v^*(\lambda(1-z)) - v^*(\lambda)}{1 - v^*(\lambda)}, \\ E(Q_b) &= \frac{\lambda E(V)}{1 - v^*(\lambda)}. \end{aligned} \tag{2.2.1}$$

As a special case, it follows directly from Theorem 2.1.1 that the stochastic decomposition properties exist in the M/G/1 (E,MV).

Theorem 2.2.1. For $\rho < 1$, in an M/G/1 (E, MV) system, the queue length L_v can be decomposed into the sum of two independent random variables,

$$L_v = L + L_d,$$