Wuyi Yue
Yutaka Takahashi
Hideaki Takagi
Editors

Advances in Queueing Theory and Network Applications

Springer

Advances in Queueing Theory and Network Applications

Wuyi Yue • Yutaka Takahashi • Hideaki Takagi
Editors

Advances in Queueing Theory and Network Applications

Springer

Editors

Wuyi Yue
Department of Intelligence and Informatics
Konan University
Kobe
658-8501 Japan
yue@konan-u.ac.jp

Hideaki Takagi
Graduate School of Systems
and Information Engineering
University of Tsukuba
Ibaraki
305-8573 Japan
takagi@sk.tsukuba.ac.jp

Yutaka Takahashi
Graduate School of Informatics
Kyoto University
Kyoto
606-8501 Japan
takahashi@i.kyoto-u.ac.jp

ISBN: 978-0-387-09702-2 e-ISBN: 978-0-387-09703-9
DOI: 10.1007/978-0-387-09703-9

Library of Congress Control Number: 2008939851

AMS Codes (2000): 60K25, 60K30, 60K15, 68M10, 68M12, 68M20, 90B15, 90B18, 90B20, 90B22

Printed on acid-free paper

springer.com

The book will prove useful to academics and industrial scientists as well as engineers engaged in research. It will also benefit those involved in post-graduate courses in communications, computer science or engineering, who have interests in the development and application of queueing theory.

This book may be used as pre-requisite reading for other more advanced courses like network design and management which are based on performance modeling with example applications for modern communication and computer networks. It could also be used as a course book on stochastic models in mathematics and operations research departments.

Contents

Preface

A considerable amount of efforts needs to be devoted to performance modeling and analysis of emerging technologies and their applications in order to develop proper design and operation management of future multi-service networks where application-dependent Quality of Service (QoS) is ensured. To this end, extensive interdisciplinary research in performance analysis and system management of communication networks is essential.

Included in this book are 16 chapters of high quality. All the manuscripts were selected, after peer-review, from among those submitted by prominent researchers working on queueing theory and network applications. The reviewers' reports not only helped the editors qualify the articles for inclusion in the book, but also improved the quality of the chapters.

The chapters cover a wide variety of recent topics on queueing theory and network applications. These include single-server queues, finite-buffer queues, retrial/balking queues, multiple queues as well as optimization in queues. They further present new theoretical results on timely topics related to protocols, application services and routing algorithms in the Internet and wireless-related issues.

We believe that all of these chapters not only provide novel ideas, new analytical models, simulation and experimental results in this field but also enhance the future research activities in the area of Queueing Theory and Network Applications.

This book consists of five parts, which cover topics in Queueing Processes, Single-Server Queues, Multiple Queues, Finite-Buffer Queues, and Network Applications. A brief summary of each chapter is listed as follows [Chapters 1–16, this book].

Part I: Queueing Processes

Part I discusses *Queueing Processes* in two chapters, Chapters 1 and 2.

Chapter 1 considers a two-sided doubly quasi-birth-and-death process. Under a discrete time setting, this is a two dimensional skip free random walk on the half space whose second component is non-negative integer valued and whose first component may take positive or negative integers. The major interest of this chapter is

in the tail decay rate of stationary distribution as one of the components goes either to infinity or to minus infinity, provided the stationary distribution exists. Two kinds of decay rates, called 'weak' and 'exact' for the doubly QBD (or DQBD) and characterized in the previous publication by the author are extended to the two sided DQBD, and are applied to the generalized shortest queue. This chapter shows that a weak decay rate, that is, the decay rate in the logarithmic sense, is completely specified in terms of the primitive data for the generalized shortest queue.

Chapter 2 proposes an analytical model based on renewal reward theory to investigate the dynamics of on-demand streaming service, deriving the average download rate. This chapter uses a simple method, combining multicast method and unicast method that can reduce the download rate from the streaming server effectively. By modeling requests as Poisson arrivals, the dynamics of this streaming service are studied and the optimal sharing of unicast and multicast methods is derived. This chapter also shows how to estimate the fluctuation of download rates of a streaming service.

Part II: Single-Server Queues

Part II on *Single-Server Queues* includes four chapters, Chapters 3–6.

In Chapter 3, a Geom/G/1 queue with a pure decrementing service policy and multiple adaptive vacations is analyzed. The Probability Generating Function (P.G.F.) of the queue length is obtained by using an embedded Markov chain method. The P.G.F. of the waiting time is then derived and the probabilities for the system being in various states such as a busy state, an idle state or a vacation state, are also derived. Finally, some special cases for the queueing model are given to demonstrate the general properties of the queue models.

Chapter 4 investigates an M/M/1 working vacation queue with setup times, using a quasi birth and death process and a matrix-geometric solution method to derive the distributions for the stationary queue length and the waiting time of a customer in the system. Also presented in this chapter are stochastic decomposition structures of stationary Indices.

In Chapter 5, a single-machine production system with early setup and extra job operations is considered. It is controlled by two thresholds. The first is used to control the setup starting time and the second is used to control the production starting time. The system is modeled by the BMAP/G/1 queue and the manufacturing lead time is analyzed. The factorization principle is used to derive its distribution and mean value.

Chapter 6 presents an analysis of a state-dependent M/E_k/1 queue with balking and single vacations. Customers are served at two different rates depending on the number of customers in the system. If customers on arrival find any other customers in the system, they decide to either enter the queue or balk with a constant probability. The server takes a single vacation when the system becomes empty. First, a quasi-birth and death process is formulated. Then, the equilibrium condition of the system is obtained. By using the matrix geometric solution method, the steady-state probability vectors are obtained. The computation of the steady-state probability

vectors is also discussed. Then, some performance measures are derived explicitly. Based on these performance analyses, a mathematical model is developed to optimize the cost of the system. Finally, some discussion on the sensitivity of the model is given through numerical experiments.

Part III: Multiple Queues

Part III discusses *Multiple Queues* in two chapters, Chapters 7 and 8.

Chapter 7 considers Markovian polling systems in which a single server serves J stations with Poisson arrivals and general service times. After a service period at station i, the server selects station j with probability p_{ij} and visits the station after spending a switchover time. This chapter uses the functional method that has been proposed in a previous research on multiclass M/G/1 type systems. The advantages of the functional method are (1) its wide applicability to the analysis of M/G/1 type multi-class queues, and (2) its rather small computational complexity compared with the buffer occupancy method.

Chapter 8 considers a two-station hybrid system which handles make-to-order (MTO) and make-to-stock products (MTS). The first station represents an MTS system producing standard products for ordinary demands, which also can be semi-finished products for specific demands processed in the second station. The second station performs some additional works on the standard products for the specific demands. In the system considered in this chapter, the MTS system is controlled under the base-stock policy. To evaluate the performance of the system, the fill rate of the ordinary demands and the response time of the specific demands are considered. The objective is to study the relation between base-stock level and the fill rate of the ordinary demands and the response time of the specific demands. The system is analyzed by modeling it as an inventory-queue model. Based on these analyses, one can determine the optimal base-stock level numerically under the constraints on the fill rate of the ordinary demands and the response time of the specific demands.

Part IV: Finite-Buffer Queues

Part IV on *Finite-Buffer Queues* includes four chapters, Chapters 9-12.

Chapter 9 presents an analysis of an M/M/c/N queueing system with balking, reneging and synchronous vacations of servers. By using the blocked matrix method, the steady-state probability vector is obtained in terms of the inverse of two matrices, whose computation is discussed. Then, the steady-state probabilities are calculated by using the elements of the inverse of the two matrices. The conditional stationary distribution of the queue length and waiting time is also derived.

An M/M/m queue with mixed loss and delay calls was analyzed by J.W. Cohen half a century ago (1956), where the two types of calls had identical constant arrival and service rates. It is straightforward to extend his analysis to an M/M/m/K queue. In Chapter 10, the model is further generalized such that the call arrival rates depend on the number of calls present in the system upon arrival. This model includes the balking and the finite population size models as special cases. A method is presented

to calculate the blocking probability for lost calls as well as the distribution of the waiting time for accepted delayed calls.

Chapter 11 considers a feedback finite fluid queue (FFFQ, for short) with downward jumps, where the fluid flow rate and the jump size are controlled by a background Markov chain with a finite state space. The feedback means that the background process may change according to the level of the buffer, which is used for modeling TCP/IP flow. The matrix analytic technique has been successfully applied to an FFFQ without jumps. This chapter incorporates downward jumps into this FFFQ, and shows that its loss probability decays exponentially as the buffer size gets large under a negative drift condition.

In Chapter 12, a closed-form explicit expression is derived for the probability density function of the length of a busy period starting with i customers in an M/M/1/K queue, where K is the capacity of the system. The density function is given as a weighted sum of K exponential distributions with coefficients calculated from K distinct zeros of a polynomial that involves Chebyshev polynomials of the second kind. The mean and second moment of the busy period are also shown explicitly.

Part V: Network Applications

Part V includes 4 chapters, Chapters 13–16 on network applications.

Chapter 13 presents a method to analyze the performance of Automatic Repeat reQuest (ARQ) schemes in self-similar traffic. A batch arrival queueing model is built by taking into account the self-similar nature of a massive-scale wireless multimedia service and by supposing that the batch size is a random variable following a Pareto distribution. A setup strategy in the model is built by considering the delay in the setting up procedure of a data link. Thus a batch arrival Geom^X/G/1 queueing system with setup is built in this chapter. A discrete-time imbedded Markov chain is used to analyze the stationary distribution of the queueing system and derive the PGFs of the queue length and the waiting time of the system. Performance measures are given in terms of the response time of data frames, setup ratios and offered loads for different ARQ schemes. Numerical results are given to evaluate the performance of the system and to show the influence of the self-similar degree and the delay of the setup procedure on these performance measures

Chapter 14 analyzes a peer-to-peer (P2P) file sharing system by means of a so-called level-dependent Quasi-Birth-and-Death (QBD) process. The dissemination of a single file consisting of different segments is considered, and a model is proposed for the upload queue management mechanism with peers competing for bandwidth. By applying an efficient matrix-analytic algorithm, the performance of P2P file diffusion can be evaluated in terms of the corresponding extinction probability, i.e., the probability that the sharing process ends.

Chapter 15 considers the performance of a decentralized content delivery system where video data is simultaneously delivered without duplication by multiple streaming video servers, resulting in a low sending rate per video server. By focusing on a multiple-server video streaming service reinforced by forward error correction

(FEC), the system is modeled as a set of independent GI+M/M/1/K queues, and the block-level loss probability is derived. Numerical results show that the decentralized content delivery system with FEC recovery is significantly effective to guarantee video quality even when the background traffic intensity is high.

Chapter 16 studies a mathematical model for calculating blocking probabilities with optimal bandwidth allocation and QoS routing on multi-class communication networks. This scheme consists of two procedures. The first step determines optimal paths under network constraints. The second step computes the blocking probability with predetermined optimal solutions. The blocking is due to the failure of meeting the demand of end-to-end paths for each class.

All the above chapters highlight the scientific and technical challenges inspired by current and future networks, and enrich novel modeling and performance evaluation techniques.

We are deeply indebted to many people for their great help during the multiple phases of publishing this book. We first would like to express our sincere gratitude to all reviewers for their valuable comments concerning all the chapters. The reviewers' reports not only helped us qualify the articles in the book, but also improved their quality in presentation. Then we are heartily grateful to all the authors for their contribution to the book. Their tremendous efforts in providing excellent chapters made the book very attractive and informative. We would like to express our appreciation to the staff members Loew, Elizabeth, Kostant, AnnBelanger, Jessica and others at Springer Publishers, Inc. for their excellent support to complete our editorial works. Last but not the least, we thank Dr. Mark S. K. Lau for assisting us in a part of editorial work of this book.

Japan,
October 2008

Wuyi Yue
Yutaka Takahashi
Hideaki Takagi

Chapter 1
Two Sided DQBD Process and Solutions to the Tail Decay Rate Problem and Their Applications to the Generalized Join Shortest Queue

Masakiyo Miyazawa

Abstract We are concerned with a two sided doubly quasi-birth-and-death process. Under a discrete time setting, this is a two dimensional skip free random walk on the half space whose second component is a nonnegative integer valued while its first component may take positive or negative integers. Our major interest is in the tail decay rate of the stationary distribution of this two sided process as either one of the components goes either to infinity or to minus infinity, provided the stationary distribution exists. The author [1] recently obtained two kinds of decay rates, called weak and exact for the doubly QBD, DQBD for short, in terms of the transition kernel of the DQBD. We extends those results to the two sided DQBD, and apply to the generalized shortest queue. The tail decay rate problem for this queueing model has been only partially answered in the literature. We show that a weak decay rate, that is, the decay rate in the logarithmic sense, is completely specified in terms of the primitive data for the generalized shortest queue. This refines results in Miyazawa [2] and corrects some results in Li, Miyazawa and Zhao [3].

1.1 Introduction

A quasi birth-and-death process, QBD process for short, is a continuous time Markov chain which has a main state, called level, and a background state in such a way that the level is nonnegative integer valued, and its increments are ± 1 at most and controlled by the background state. This model has been well studied when the background state space is finite (see, e.g. [4], [5]).

We are concerned with the case that the background space is infinite. Li, Miyazawa and Zhao [3] recently proposed a double sided QBD process for the generalized join shortest queue with two waiting lines, by extending the level of

Masakiyo Miyazawa
Department of Information Sciences, Tokyo University of Sciences, Chiba 278-8510, Japan
e-mail: miyazawa@is.noda.tus.ac.jp

W. Yue et al. (eds.), *Advances in Queueing Theory and Network Applications,*

such a QBD process to be integer valued. This queue is a service system with two parallel queues that have three arrival streams, two of which are dedicated to each queue and the other of which chooses the shortest queue with tie breaking. Assume that those arrival streams are independent and subject to Poisson processes, and service times are independently, identically and exponentially distributed at each queue. Then, this queue can be formulated as the QBD process or the two sided QBD process. In particular, the latter model is required when we take the difference of the two queues as level.

It is notable that the transition structure may change in the double sided QBD when the level process goes through zero. This is crucial to formulate the generalized join shortest queue as the double sided QBD. In this chapter, we specialize this double sided QBD in such a way that its background process is birth-and-death. We refer to this process as a two sided doubly quasi birth-and-death process, a two sided DQBD for short. Since those QBD and DQBD can be formulated as discrete time Markov chains, we are only concerned with the discrete time processes throughout the chapter.

We are interested in the asymptotic behaviors of the stationary distributions of the level and background state as their values go to infinity, provided it exists. Due to the special structure of the two sided DQBD, the QBD structure is preserved when the level and background are exchanged. So, we mainly consider the asymptotics for the level. We are concerned with two types of the asymptotic decays of the stationary probabilities as the level goes to infinity.

One type is called a weak decay, which is meant that the logarithm of the stationary probability dived by the level n converges to a constant, say $-a$, as n goes to infinity. Then, e^{-a} is simply referred to as a weak decay rate. Another type is called an exactly geometric decay, which is meant that the stationary probability multiplied by a power constant to the level n, say α^n, converges to an another constant as n goes to infinity. Then, α^{-1} is referred to as an exactly geometric decay rate. In [1], more general types of exact decay rates are considered, but we are only concerned with these two types of decay rates in this chapter.

The purpose of this chapter is twofold. We first study the decay rate problem for the two sided DQBD process, by extending the approach for the DQBD process in [1]. We completely characterize the weak decay rates in terms of the transition probabilities (Theorems 1.3 and 1.4). For the exactly geometric decay, we find sufficient conditions, which are close to necessary conditions (Theorem 1.3). We secondly apply these results to find the decay rates of the stationary distributions of the minimum of the two queues and their difference in the generalized join shortest queue with two waiting lines.

The decay rates for this queue have been studied in [3] and [6], but they are obtained only for certain limited cases, e.g., under a so called strongly pooled condition. We completely answer to this problem for the weak decay rates, and give weaker sufficient conditions for the exactly geometric decay rates (Theorem 1.5 and Corollary 1.2). In particular, it turns out that the strongly pooled condition still plays an important role for finding the decay rate for the minimum of two queues, which may not be the square of the total traffic intensity in general.

The two sided DQBD is a special case of the double sided QBD introduced in [3] since the latter allows the background process to be a general Markov chain. The exactly geometric decays are studied in [3], but only sufficient conditions are obtained. Furthermore, those sufficient conditions require the stationary probabilities at the boundaries, i.e., at level 0, so they are not easy to verify. Not for the two sided DQBD but for the DQBD, Miyazawa [1] completely solves the decay rate problem recently, developing the ideas in [2].

We here extend this approach in [1]. Thus, many arguments are parallel to those in [1]. Namely, the approach heavily depends on the QBD structure and the Wiener Hopf factorization for the Markov additive process that generate the QBD process, and the key idea is to formulate the decay rate problem as a multidimensional optimization problem. However, the level and background states are not symmetric in the two sided DQBD while they are symmetric in the DQBD. So, we need some further effort to get the decay rates, which is a main contribution of this chapter for a general QBD model.

For the join shortest queue and its generalized versions, the decay rate problem has been widely studied in the literature. One possible approach is to use the large deviation principle. Puhalskii and Vladimirov [7] recently obtained the weak decay rates as the solutions of the variational problem for a much more general class of the generalized join shortest queue with an arbitrary number of parallel queues. However, this variational problem is very hard to not only analytically but also numerically solve even for the case of two queues.

Another approach is either to use the random walk structure or the QBD formulation. For example, Foley and McDonald [6] took the former formulation while Li, Miyazawa and Zhao [3] took the latter formulation. An interesting sufficient condition, i.e., so called strongly pooled condition, is found in [6]. However, those papers mainly consider the decay rate under this limited condition for the case of the two queues. So far, the decay rate problem has not been well answered for the generalized join shortest queue. In this chapter, we completely solve this problem for the case of the two queues (Theorem 1.5 and Corollary 1.2). For simpler arrival processes, there are many other studies on the join shortest queues and the decay rate problem has been relatively well answered (see references in [3], [6]).

This chapter is made up by seven sections. In Sect. 1.2, we introduce the two sided DQBD process formally, and consider its basic property, particularly on the rate matrices for representing the stationary distribution in a matrix geometric form. In Sect. 1.3, we characterize the set of positive eigenvectors of the rate matrices using the moment generating functions of the transition kernels insides and on the boundaries. The weak decay rates are completely answered in Sect. 1.4. We also give sufficient conditions for those decay rates to be exactly geometric. In Sect. 1.5, we consider the generalized join the shortest queue with two queues, and answer to the decay rate problems. We finally give some remarks on the existence results in Sect. 1.6. Conclusions are drawn in Sect. 1.7.

1.2 Two Sided DQBD Process

Let $\{(L_{1t}, L_{2t}); t = 0, 1, \ldots\}$ be a two dimensional Markov chain taking values in $S \equiv \mathbb{Z} \times \mathbb{Z}_+$, where $\mathbb{Z}$ is the set of all integers and $\mathbb{Z}_+ = \{\ell \in \mathbb{Z}; \ell \geq 0\}$, with the following transition probabilities (see Fig. 1.1).

$$P(L_{1(t+1)} = i', L_{2(t+1)} = j' | L_{1t} = i, L_{2t} = j)$$
$$= \begin{cases} p^{+}_{(i'-i)(j'-j)}, & i \geq 1, j \geq 1, i'-i, j'-j = 0, \pm 1 \\ p^{-}_{(i'-i)(j'-j)}, & i \leq -1, j \geq 1, i'-i, j'-j = 0, \pm 1 \\ p^{(1+)}_{(i'-i)j'}, & i \geq 1, j = 0, i'-i = 0, \pm 1, j' = 0, 1 \\ p^{(1-)}_{(i'-i)j'}, & i \leq -1, j = 0, i'-i = 0, \pm 1, j' = 0, 1 \\ p^{(2)}_{i'(j'-j)}, & i = 0, j \geq 1, i' = 0, 1, j'-j = 0, \pm 1 \\ p^{(0)}_{i'j'}, & i = j = 0, i' = 0, \pm 1, j' = 0, 1 \\ 0, & \text{otherwise}, \end{cases}$$

where $\sum_{i,j} p_{ij} = \sum_{i,j} p^{(k)}_{ij} = 1$ for $k = 0, \pm, 1\pm, 2$. Thus, $\{(L_{1t}, L_{2t})\}$ is a skip free random walk in all directions, and reflected at the boundary $\partial S_1 \equiv \{(i, j) \in S; j = 0\}$ and has discontinuous statistics at $\partial S_2 \equiv \{(i, j) \in S; i = 0\}$.

We first take L_{1t} as level, and L_{2t} as background state, and refer to this Markov chain as a discrete-time two sided DQBD (doubly quasi-birth-and-death) process. In the random walk terminology, this process is two dimensional reflected random walk on the half space $\{(m, n) \in \mathbb{Z}^2; n \geq 0\}$ with discontinuous statistics at the boundaries where either one of components vanishes. We also note that this model is a special case of the double sided QBD in [3] whose background process is not necessary to be birth-and-death.

To present the transition probability matrix of this Markov chain, we first introduce the following matrices. For $k = 0, \pm 1$ and $s = \pm$,

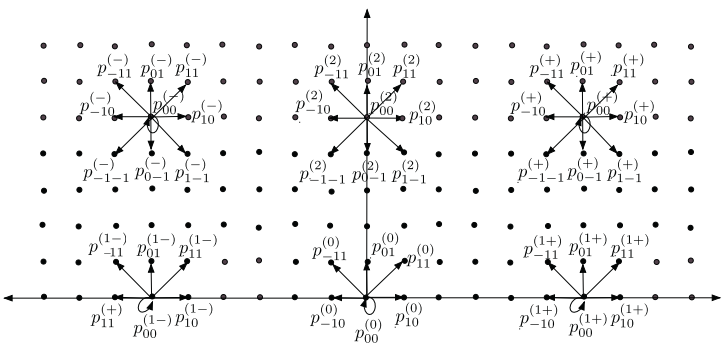


Fig. 1.1 State transitions for the two sided DQBD process.

$$A_k^{(s)} = \begin{pmatrix} p_{k0}^{(1s)} & p_{k1}^{(1s)} & 0 & \cdots & & & & \\ p_{k(-1)}^{(s)} & p_{k0}^{(s)} & p_{k1}^{(s)} & 0 & \cdots & & & \\ 0 & p_{k(-1)}^{(s)} & p_{k0}^{(s)} & p_{k1}^{(s)} & 0 & \cdots & & \\ 0 & 0 & p_{k(-1)}^{(s)} & p_{k0}^{(s)} & p_{k1}^{(s)} & 0 & \cdots & \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix},$$

and for $k = 0, \pm 1$,

$$B_k^{(1)} = \begin{pmatrix} p_{k0}^{(0)} & p_{k1}^{(0)} & 0 & \cdots & & & & \\ p_{k(-1)}^{(2)} & p_{k0}^{(2)} & p_{k1}^{(2)} & 0 & \cdots & & & \\ 0 & p_{k(-1)}^{(2)} & p_{k0}^{(2)} & p_{k1}^{(2)} & 0 & \cdots & & \\ 0 & 0 & p_{k(-1)}^{(2)} & p_{k0}^{(2)} & p_{k1}^{(2)} & 0 & \cdots & \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}.$$

Then, the two sided DQBD has the following tridiagonal transition matrix $P^{(1)}$.

$$P^{(1)} = \begin{pmatrix} \ddots & \ddots & \ddots & \ddots & \ddots & \cdots & & & & \\ \cdots & 0 & A_{-1}^{(-)} & A_0^{(-)} & A_1^{(-)} & 0 & \cdots & & & \\ & \cdots & 0 & A_{-1}^{(-)} & A_0^{(-)} & A_1^{(-)} & 0 & \cdots & & \\ & & \cdots & 0 & B_{-1}^{(1)} & B_0^{(1)} & B_1^{(1)} & 0 & \cdots & \\ & & & \cdots & 0 & A_{-1}^{(+)} & A_0^{(+)} & A_1^{(+)} & 0 & \cdots \\ & & & & \cdots & 0 & A_{-1}^{(+)} & A_0^{(+)} & A_1^{(+)} & 0 \\ & & & & & \cdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}.$$

Throughout this chapter, we assume that $P^{(1)}$ is irreducible and aperiodic, and positive recurrent. The unique stationary distribution of P is denoted by probability row vector:

$$\nu = (\ldots, \nu_{-1}, \nu_0, \nu_1, \ldots),$$

where $\nu_{\mathbf{n}}$ for $n \in \mathbb{Z}$ are row vectors for background states in level n. We also write ν as $\{\nu_{ij}; i \in \mathbb{Z}, j \in \mathbb{Z}_+\}$. We assume that

(i) For each $s = \pm, 2$, $A^{(s)} \equiv A_{-1}^{(s)} + A_0^{(s)} + A_1^{(s)}$ is irreducible and aperiodic;

(ii) For each $s = \pm, 2$, Markov additive process driven by kernel $\{A_n^{(s)}; n = 0, \pm 1\}$ is 1-arithmetic in the sense that for every pair $(i, j) \in S_1 \times S_1$, the greatest common divisor of $\{n \in \mathbb{Z}; A_n^{(s)}(i, j) > 0\}$ is one, where $\mathbb{Z}$ is the set of all integers (see, e.g., [8]).

Remark 1.1. The irreducibility of $A^{(s)}$ in (i) is satisfied by many applications, but it is stronger than the irreducibility of P. Our arguments in this chapter can be modified

so as to be valid without that irreducibility, and the same results are obtained. However, proofs becomes complicated just because we need to consider each case separately depending on the irreducibility or the non irreducibility. So, we here do not consider the non irreducible case, which will be detailed in a technical note.

It is well-known that, for each $s = \pm$, there exists a nonnegative matrix $R^{(s)}$ uniquely determined as a minimal nonnegative solution of the matrix equation:

$$R^{(-)} = A_{-1}^{(-)} + R^{(-)}A_0^{(-)} + \left(R^{(-)}\right)^2 A_1^{(-)}, \tag{1.1}$$

$$R^{(+)} = \left(R^{(+)}\right)^2 A_{-1}^{(+)} + R^{(+)}A_0^{(+)} + A_1^{(+)}, \tag{1.2}$$

and the stationary distribution has the following matrix geometric form.

$$\nu_n = \begin{cases} \nu_1 \left(R^{(+)}\right)^{n-1}, & n \geq 1 \\ \nu_{-1} \left(R^{(-)}\right)^{-n-1}, & n \leq -1. \end{cases} \tag{1.3}$$

Note that $R^{(s)}$ may not be irreducible, but has a single irreducible class due to (i) and (ii).

We also consider the case that L_2 is taken as level. In this case, the transition matrix is denoted by $P^{(2)}$, and given by

$$P^{(2)} = \begin{pmatrix} B_0^{(2)} & B_1^{(2)} & 0 & \cdots & & \\ A_{-1}^{(2)} & A_0^{(2)} & A_1^{(2)} & 0 & \cdots & \\ 0 & A_{-1}^{(2)} & A_0^{(2)} & A_1^{(2)} & 0 & \cdots \\ 0 & 0 & A_{-1}^{(2)} & A_0^{(2)} & A_1^{(2)} & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix},$$

where, for $k = 0, \pm 1$,

$$A_k^{(2)} = \begin{pmatrix} \ddots & \ddots & \ddots & \ddots & \ddots & \cdots & & & \\ \cdots & 0 & p_{(-1)k}^{(-)} & p_{0k}^{(-)} & p_{1k}^{(-)} & 0 & \cdots & & \\ & \cdots & 0 & p_{(-1)k}^{(-)} & p_{0k}^{(-)} & p_{1k}^{(-)} & 0 & \cdots & \\ & & \cdots & 0 & p_{-1k}^{(2)} & p_{0k}^{(2)} & p_{1k}^{(2)} & 0 & \cdots \\ & & & \cdots & 0 & p_{(-1)k}^{(+)} & p_{0k}^{(+)} & p_{1k}^{(+)} & 0 & \cdots \\ & & & & \cdots & 0 & p_{(-1)k}^{(+)} & p_{0k}^{(+)} & p_{1k}^{(+)} & 0 & \cdots \\ & & & & & \cdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix},$$

and for $k = 0, 1$,

$$B_k^{(2)} = \begin{pmatrix} \ddots & \ddots & \ddots & \ddots & \ddots & \cdots & & & & \\ \cdots & 0 & p_{(-1)k}^{(1-)} & p_{0k}^{(1-)} & p_{1k}^{(1-)} & 0 & \cdots & & & \\ & \cdots & 0 & p_{(-1)k}^{(1-)} & p_{0k}^{(1-)} & p_{1k}^{(1-)} & 0 & \cdots & & \\ & & \cdots & 0 & p_{-1k}^{(2)} & p_{0k}^{(2)} & p_{1k}^{(2)} & 0 & \cdots & \\ & & & \cdots & 0 & p_{(-1)k}^{(1+)} & p_{0k}^{(1+)} & p_{1k}^{(1+)} & 0 & \cdots \\ & & & & \cdots & 0 & p_{(-1)k}^{(1+)} & p_{0k}^{(1+)} & p_{1k}^{(1+)} & 0 & \cdots \\ & & & & & \cdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}.$$

In this case, the stationary distribution $\nu = \{\nu_{ij}\}$ is partitioned as

$$\nu = \left(\nu_0^{(2)}, \nu_1^{(2)}, \ldots\right),$$

where $\nu_n^{(2)} = \{\nu_{in}; i \in \mathbb{Z}\}$. Viewing L_{1t} as the background process, we have the standard process. Then, as is well known, there exists a minimal nonnegative solution $R^{(2)}$ of

$$R^{(2)} = \left(R^{(2)}\right)^2 A_{-1}^{(2)} + R^{(2)} A_0^{(2)} + A_1^{(2)}, \tag{1.4}$$

and the stationary distribution ν has the following form:

$$\nu_n^{(2)} = \nu_1^{(2)} \left(R^{(2)}\right)^{n-1}, \qquad n \geq 1. \tag{1.5}$$

We are interested in the geometric decay behaviors of the stationary vector ν_n as $n \to \pm\infty$ and $\nu_n^{(2)}$ as $n \to \infty$. We are interested in two different types of asymptotics. If there are constant $\alpha_+ > 1$ and constant positive vector $\mathbf{c}_+$ such that

$$\lim_{n\to\infty} \alpha_+^n \nu_n = \mathbf{c}_+,$$

then ν_n is said to asymptotically have exactly geometric decay rate α_+^{-1} as $n \to \infty$. Another decay rate is of logarithmic type, which is defined through

$$\log r_+(i) = \lim_{n\to\infty} \frac{1}{n} \log \nu_{ni}, \qquad i \in \mathbb{Z}_+, \tag{1.6}$$

where $r_+(i) \leq 1$. If $r_+(i)$ does not depend on i, we write it as r_+. In this case, ν_n is said to asymptotically have weak geometric decay rate r_+. Those decay rates are also defined for ν_n as $n \to -\infty$ and for $\nu_n^{(2)}$ as $n \to \infty$, which are denoted by r_- and r_2, respectively. Since those decay rates may not exist, we also use the following notation:

$$\log \underline{r}_+(i) = \liminf_{n\to\infty} \frac{1}{n} \log \nu_{ni}, \qquad \log \overline{r}_+(i) = \limsup_{n\to\infty} \frac{1}{n} \log \nu_{ni}, \qquad i \in \mathbb{Z}_+.$$

Similarly, $\underline{r}_s(i)$ and $\bar{r}_s(i)$ are defined for $s = -, 2$. These decay rates are referred to as the weak lower and weak upper decay rates, respectively.

It is noticed that $\bar{r}_+(i)$ in (1.6) is bounded as

$$\underline{r}_+(i)^{-1} \le \sup\left\{ z \ge 1; \sum_{n=0}^{\infty} z^n \nu_{ni} < \infty \right\}, \qquad i \in \mathbb{Z}_+.$$

Then, from (1.3) and (1.5), it might be expected that the weak decay rate r_+^{-1} is obtained as the reciprocal of the convergence parameter $c_p(R^{(+)})$ of $R^{(+)}$, which is defined as

$$c_p(R^{(+)}) = \sup\left\{ z \ge 0; \sum_{n=0}^{\infty} z^n \left(R^{(+)}\right)^n < \infty \right\}.$$

This is true under certain situations, but generally not true. In general, we only have the following lower bounds for the decay rates from this information.

Lemma 1.1. The decay rates are bounded below by the corresponding convergence parameters of the rate matrices. That is, we have

$$\underline{r}_s(i) \ge c_p(R^{(s)})^{-1}, \qquad s = \pm, 2, \qquad i \in \mathbb{Z}_s,$$

where $\mathbb{Z}_2 = \mathbb{Z}$.

The proof of this lemma is exactly the same as Lemma 2.1 of [1], so it is omitted. This lemma just gives the lower bounds, but it turns out that they are very useful to identify the decay rates as well as to prove their existence.

We next prepare some useful facts for the convergence parameters.

Proposition 1.1 (Theorem 6.3 of [9]). For a nonnegative square matrix T, let $\mathbf{X}$ be the set of all nonnegative and nonzero row vectors whose size is the same as that of T. Then we have

$$c_p(T) = \sup\{z \ge 0; z\mathbf{x}T \le \mathbf{x}, \mathbf{x} \in \mathbf{X}\}.$$

We will consider all eigenvectors of $R^{(s)}$, $s = \pm, 2$, to find the decay rate. For this, we use the Markov additive process generated by $\{A_k^{(s)}; k = 0, \pm 1\}$. Note that (1.1) and (1.2) and the corresponding equations of $R^{(-)}$ and $R^{(2)}$ are equivalent to

$$I - A_*^{(s)}(z^{u(s)}) = (I - zR^{(s)})(I - G_*^{(s)}(z)), \qquad z \ne 0, s = \pm, 2, \tag{1.7}$$

where $u(s) = -1$ for $s = -$, $u(s) = 1$ for $s = +, 2$, and $A_*(z)$ and $G_*^{(s)}(z)$ are defined as

$$A_*^{(s)}(z) = z^{-1}A_{-1}^{(s)} + A_0^{(s)} + zA_1^{(s)}, \qquad G_*^{(s)}(z) = A_0^{(s)} + R^{(s)}A_{-1}^{(s)} + z^{-1}A_{-1}^{(s)}.$$

The decomposition formula (1.7) is known as the Wiener Hopf factorization (see, e.g. [8]). Then Proposition 1.1 concludes

Lemma 1.2. $c_p(R^{(s)}) = \sup\{z \geq 1; \mathbf{x}A_*^{(s)}(z) \leq \mathbf{x}, \mathbf{x} \in \mathbf{X}^{(s)}\}$ for $s = \pm, 2$, where $\mathbf{X}^{(s)}$ is the set of all nonnegative and nonzero vectors in $R^{\mathbb{Z}_+}$ for $s = \pm$ and $R^{\mathbb{Z}}$ for $s = 2$.

Remark 1.2. The proof that $\mathbf{x}A_*^{(s)}(z) \leq \mathbf{x}$ implies $z \leq c_p(R^{(s)})$ is immediate from (1.7). However, it is not so obvious to find z such that $c_p(R^{(s)}) \leq z$ and $\mathbf{x}A_*^{(s)}(z) \leq \mathbf{x}$. The proof of this can be found in [1].

Similarly to the case of the doubly QBD process in [1], we compute each entry of $A_*^{(s)}(z)$ using the following notations. Let (X_1, X_2) be a random vector subject to distribution $\{p_{ij}\}$, and let $(X_1^{(s)}, X_2^{(s)})$ be those subject to distribution $\{p_{ij}^{(s)}\}$ for $s = 0, \pm, 1\pm, 2$. Define generating functions as

$$
\begin{aligned}
&p_{**}^{(s)}(u,v) = \mathrm{E}[u^{X_1^{(s)}} v^{X_2^{(s)}}], \quad p_{*j}^{(1s)}(u) = \mathrm{E}[u^{X_1^{(1s)}} 1(X_2^{(1s)} = j)], \quad s = \pm, j = 0, 1,\\
&p_{i*}^{(s)}(v) = \mathrm{E}[v^{X_2^{(s)}} 1(X_1^{(s)} = i)], \qquad s = \pm, i, j = 0, \pm 1,\\
&p_{*j}^{(s)}(u) = \mathrm{E}[u^{X_1^{(s)}} 1(X_2^{(s)} = j)], \qquad s = \pm, i, j = 0, \pm 1,\\
&p_{i*}^{(2)}(u) = \mathrm{E}[v^{X_2^{(2)}} 1(X_1^{(2)} = i)], \qquad i = 0, \pm 1.
\end{aligned}
$$

Then, $A_*^{(s)}(z)$ for $s = \pm$ has the following QBD structure.

$$
A_*^{(s)}(z) = \begin{pmatrix}
p_{*0}^{(1s)}(z) & p_{*1}^{(1s)}(z) & 0 & \cdots & & & \\
p_{*(-1)}^{(s)}(z) & p_{*0}^{(s)}(z) & p_{*1}^{(s)}(z) & 0 & \cdots & & \\
0 & p_{*(-1)}^{(s)}(z) & p_{*0}^{(s)}(z) & p_{*1}^{(s)}(z) & 0 & \cdots & \\
0 & 0 & p_{*(-1)}^{(s)}(z) & p_{*0}^{(s)}(z) & p_{*1}^{(s)}(z) & 0 & \cdots \\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots
\end{pmatrix}. \tag{1.8}
$$

Similarly, we have

$$
A_*^{(2)}(z) = \begin{pmatrix}
\ddots & \ddots & \ddots & \ddots & \ddots & \cdots & & & & \\
\cdots & 0 & p_{(-1)*}^{(-)}(z) & p_{0*}^{(-)}(z) & p_{1*}^{(-)}(z) & 0 & \cdots & & & \\
 & \cdots & 0 & p_{(-1)*}^{(-)}(z) & p_{0*}^{(-)}(z) & p_{1*}^{(-)}(z) & 0 & \cdots & & \\
 & & \cdots & 0 & p_{-1*}^{(2)}(z) & p_{0*}^{(2)}(z) & p_{1*}^{(2)}(z) & 0 & \cdots & \\
 & & & \cdots & 0 & p_{(-1)*}^{(+)}(z) & p_{0*}^{(+)}(z) & p_{1*}^{(+)}(z) & 0 & \cdots \\
 & & & & \cdots & 0 & p_{(-1)*}^{(+)}(z) & p_{0*}^{(+)}(z) & p_{1*}^{(+)}(z) & 0 \\
 & & & & & \cdots & \ddots & \ddots & \ddots & \ddots
\end{pmatrix}.
$$

1.3 Eigenvectors of Rate Matrices

If we take L_1 as level, then the background process $\{L_{2t}\}$ is the birth and death process in each half lie $[1,\infty)$ or $(-\infty,-1]$. Hence, this case is easier, so we consider $R^{(s)}$ with $s=\pm$ first. In what follows, we use the following notations for $s=\pm,2$.

$$\mathscr{V}_R^{(s)} = \left\{(z,\mathbf{x}); z\mathbf{x}R^{(s)} = \mathbf{x}, z \geq 1, \mathbf{x} \in \mathbf{X}^{(s)}\right\},$$
$$\mathscr{V}_A^{(s)} = \left\{(z,\mathbf{x}); \mathbf{x}A^{(s)}(z) = \mathbf{x}, z \geq 1, \mathbf{x} \in \mathbf{X}^{(s)}\right\}.$$

We first note the following facts, which are easily concluded by the Wiener Hopf factorization.

Lemma 1.3. Let s be either one of $-$, $+$ or 2. For $z>1$, $(z,\mathbf{x}) \in \mathscr{V}_R^{(s)}$ if and only if $(z,\mathbf{x}) \in \mathscr{V}_A^{(s)}$. If there is no $(z,\mathbf{x})$ in $\mathscr{V}_A^{(s)}$ with $z>1$, then $c_p(R^{(s)})=1$.

Then, the following result is immediate from Theorem 3.1 in [1].

Theorem 1.1. Let $\mathscr{D}_1^{(-)}$ denote the subset of all $(-\theta_1,\theta_2)$ in $\mathbb{R}^2$ such that

$$\mathrm{E}\left[e^{\theta_1 X_1^{(-)}+\theta_2 X_2^{(-)}}\right] = 1, \tag{1.9}$$
$$\varphi_0^{(1-)}(\theta_1) + \varphi_1^{(1-)}(\theta_1)e^{\theta_2} \leq 1, \tag{1.10}$$
$$\theta_1 \leq 0, \theta_2 \in \mathbb{R},$$

where $\varphi_i^{(1-)}(\theta_1) = \mathrm{E}\left[e^{\theta_1 X_1^{(1-)}}; X_2^{(1-)} = j\right]$ for $j=0,1$. Similarly, let $\mathscr{D}_1^{(+)}$ denote the subset of all (θ_1,θ_2) in $\mathbb{R}^2$ such that

$$\mathrm{E}\left[e^{\theta_1 X_1^{(+)}+\theta_2 X_2^{(+)}}\right] = 1, \tag{1.11}$$
$$\varphi_0^{(1+)}(\theta_1) + \varphi_1^{(1+)}(\theta_1)e^{\theta_2} \leq 1, \tag{1.12}$$
$$\theta_1 \geq 0, \theta_2 \in \mathbb{R},$$

where $\varphi_i^{(1+)}(\theta_1) = \mathrm{E}\left[e^{\theta_1 X_1^{(1+)}}; X_2^{(1+)} = j\right]$ for $j=0,1$. Then, for each $s=\pm$, there exists a $(z,\mathbf{x}) \in \mathscr{V}_A^{(s)}$ if and only if there exists a $(\theta_1,\theta_2) \in \mathscr{D}_1^{(s)}$. Furthermore, we have the following facts.

(1a) For this (θ_1,θ_2), $(z,\mathbf{x}) \in \mathscr{V}_A^{(-)}$ $\left(\text{res., } \mathscr{V}_A^{(+)}\right)$ is given by $z=e^{\theta_1}$ and $\mathbf{x}=\{x_n\}$:

$$x_n = \begin{cases} c_1 e^{-\underline{\theta}_2(n-1)} + c_2 e^{-\overline{\theta}_2(n-1)}, & \underline{\theta}_2 \neq \overline{\theta}_2, \\ (c_1' + c_2'(n-1))\, e^{-\underline{\theta}_2(n-1)}, & \underline{\theta}_2 = \overline{\theta}_2, \end{cases} \qquad n \geq 1, \tag{1.13}$$

where $\underline{\theta}_2, \overline{\theta}_2$ are the two solutions of (1.9) (res., (1.11)) for the given θ_1 such that $\underline{\theta}_2 \leq \overline{\theta}_2$, and c_i, c_i' are nonnegative constants satisfying $c_1+c_2 \neq 0$ and $c_1'+$

$c_2' \neq 0$. Furthermore, both of c_1 and c_2 are positive only if the strict inequality holds in (1.9) (res., (1.11)).

(1b) The convergence parameter $c_p(R^{(s)})$ is obtained as the supremum of e^{θ_1} over $\mathscr{D}_1^{(s)}$ for each $s = \pm$.

Similarly to Theorem 1.1, we can prove the following theorem for $R^{(2)}$. Since this result is the core of our arguments, we give its detailed proof in Appendix A.

Theorem 1.2. Let $\mathscr{D}_2$ denote the subset of all $(-\eta_1^{(-)}, \eta_1^{(+)}, \eta_2)$ in $\mathbb{R}^3$ such that

$$\mathrm{E}\left[e^{\eta_1^{(-)}X_1^{(-)}+\eta_2 X_2^{(-)}}\right] = 1, \tag{1.14}$$

$$\mathrm{E}\left[e^{\eta_1^{(+)}X_1^{(+)}+\eta_2 X_2^{(+)}}\right] = 1, \tag{1.15}$$

$$\varphi_{-1}^{(2)}(\eta_2)e^{-\eta_1^{(-)}} + \varphi_0^{(2)}(\eta_2) + \varphi_1^{(2)}(\eta_2)e^{\eta_1^{(+)}} \leq 1, \tag{1.16}$$

$$\eta_2 \geq 0, \eta_1^{(-)}, \eta_1^{(+)} \in \mathbb{R},$$

where $\varphi_i^{(2)}(\eta_2) = \mathrm{E}\left[e^{\eta_2 X_2^{(2)}}; X_1^{(2)} = i\right]$ for $i = 0, \pm 1$. Then, there exists a $(z, \mathbf{x}) \in \mathscr{V}_A^{(2)}$ if and only if there exists a $(-\eta_1^{(-)}, \eta_1^{(+)}, \eta_2) \in \mathscr{D}_2$. Furthermore, we have the following facts.

(2a) For this $(-\eta_1^{(-)}, \eta_1^{(+)}, \eta_2)$, $(z, \mathbf{x}) \in \mathscr{V}_A^{(2)}$ is given by $z = e^{\eta_2}$ and $\mathbf{x} = \{x_n\}$:

$$x_n^{(s)} = \begin{cases} c_1^{(s)} e^{-\underline{\eta}_1^{(s)}(n-1)} + c_2^{(s)} e^{-\overline{\eta}_1^{(s)}(n-1)}, & \underline{\eta}_1^{(s)} \neq \overline{\eta}_1^{(s)}, \\ \left(d_1^{(s)} + d_2^{(s)}|n-1|\right) e^{-\underline{\eta}_1^{(s)}(n-1)}, & \underline{\eta}_1^{(s)} = \overline{\eta}_1^{(s)} \end{cases} \quad n \geq 1, s = \pm, \tag{1.17}$$

where $\underline{\eta}_1^{(-)}, \overline{\eta}_1^{(-)}$ (res., $\underline{\eta}_1^{(+)}, \overline{\eta}_1^{(+)}$) are the two solutions of (1.14) (res., (1.15)) for the given η_2 such that $\underline{\eta}_1^{(-)} \leq \overline{\eta}_1^{(-)}$ (res., $\underline{\eta}_1^{(+)} \leq \overline{\eta}_1^{(+)}$), and for each $s = \pm$, $c_i^{(s)}, d_i^{(s)}$ are nonnegative constants satisfying $c_1^{(s)} + c_2^{(s)} \neq 0$ and $d_1^{(s)} + d_2^{(s)} \neq 0$. Furthermore, both of $c_1^{(s)}$ and $c_2^{(s)}$ are positive only if the strict inequality holds in (1.16).

(2b) The convergence parameter $c_p(R^{(2)})$ is obtained as the supremum of e^{η_2} over $\mathscr{D}_2$.

For convenience, we also introduce the following projections of $\mathscr{D}_2$, which will be used in Lemma 1.7.

$$\mathscr{D}_2^{(-)} = \{(\eta_1^{(-)}, \eta_2); (\eta_1^{(-)}, \eta_1^{(+)}, \eta_2) \in \mathscr{D}_2\},$$
$$\mathscr{D}_2^{(+)} = \{(\eta_1^{(+)}, \eta_2); (\eta_1^{(-)}, \eta_1^{(+)}, \eta_2) \in \mathscr{D}_2\}.$$

An important observation in these theorems is that z satisfying $(z, \mathbf{x}) \in \mathscr{V}(R^{(s)})$ can be found through θ_1 or η_2 in sets $\mathscr{D}_1^{(-)}$, $\mathscr{D}_1^{(+)}$ and $\mathscr{D}_2$, which are in the boundary

of convex sets. Furthermore, $\mathscr{D}_1^{(s)}$, $\mathscr{D}_2^{(2)}$ and $\mathscr{D}_2$ are compact and connected sets for $s=\pm$. This observation is expected to extend Corollary 3.1 of [1] for the two sided DQBD process. However, we have to check the two sided version of Proposition 3.1 of [1]. That is, we need the following lemmas. For convenience, we denote the set of non-positive integers by $\mathbb{Z}_-$.

Lemma 1.4. For $s=\pm$, if there exist a positive vector $\mathbf{x}^{(s)}=\{x_n^{(s)};n\in\mathbb{Z}_s\}$ such that $(\alpha_s,\mathbf{x}^{(s)})\in\mathscr{V}_A^{(s)}$ and some finite $\underline{d}_s(\mathbf{x}),\overline{d}_s(\mathbf{x})\ge 0$ such that

$$\liminf_{n\to\infty}\frac{\nu_{sn}}{x_n^{(s)}}=\underline{d}_s(\mathbf{x}^{(s)}),\qquad \limsup_{n\to\infty}\frac{\nu_{sn}}{x_n^{(s)}}=\overline{d}_s(\mathbf{x}^{(s)}),$$

then, for any nonnegative column vector $\mathbf{u}^{(s)}$ satisfying $\mathbf{x}^{(s)}\mathbf{u}^{(s)}<\infty$, there are nonegative and finite $\underline{d}_s^\dagger(\mathbf{x}^{(s)})$ and $\overline{d}_s^\dagger(\mathbf{x}^{(s)})$ such that

$$\begin{aligned}\alpha_s\underline{d}_s^\dagger(\mathbf{x}^{(s)})\mathbf{x}^{(s)}\mathbf{u}^{(s)}&\le\liminf_{n\to s\infty}\alpha_s^{|n|}\nu_n\mathbf{u}^{(s)}\\&\le\limsup_{n\to s\infty}\alpha_s^{|n|}\nu_n\mathbf{u}^{(s)}\le\alpha_s\overline{d}_s^\dagger(\mathbf{x}^{(s)})\mathbf{x}^{(s)}\mathbf{u}^{(s)}.\end{aligned}\tag{1.18}$$

In particular, if $\underline{d}_s^\dagger(\mathbf{x}^{(s)})=\overline{d}_s^\dagger(\mathbf{x}^{(s)})$ and $0\le d_s^\dagger\equiv\underline{d}_s^\dagger(\mathbf{x}^{(s)})<\infty$, then

$$\lim_{n\to s\infty}\alpha_s^n\nu_n\mathbf{u}^{(s)}=\alpha_s d_s^\dagger\mathbf{x}^{(s)}\mathbf{u}^{(s)}.\tag{1.19}$$

That is, $\nu_n\mathbf{u}^{(s)}$ decays geometrically with rate α_s^{-1} as $n\to s\infty$.

Lemma 1.5. If there exist a positive vector $\mathbf{x}=\{x_n;n\in\mathbb{Z}\}$ such that $(\alpha,\mathbf{x})\in\mathscr{V}_A^{(2)}$ and some finite $\underline{d}^-(\mathbf{x}),\overline{d}^-(\mathbf{x}),\underline{d}^+(\mathbf{x}),\overline{d}^+(\mathbf{x})\ge 0$ such that

$$\begin{aligned}\liminf_{n\to-\infty}\frac{\nu_{n1}}{x_n}&=\underline{d}^-(\mathbf{x}),&\limsup_{n\to-\infty}\frac{\nu_{n1}}{x_n}&=\overline{d}^-(\mathbf{x}),\\\liminf_{n\to+\infty}\frac{\nu_{n1}}{x_n}&=\underline{d}^+(\mathbf{x}),&\limsup_{n\to+\infty}\frac{\nu_{n1}}{x_n}&=\overline{d}^+(\mathbf{x}),\end{aligned}$$

then, for any nonnegative column vector $\mathbf{u}$ satisfying $\mathbf{x}\mathbf{u}<\infty$, there are nonegative and finite $\underline{d}^\dagger(\mathbf{x})$ and $\overline{d}^\dagger(\mathbf{x})$ such that

$$\alpha\underline{d}^\dagger(\mathbf{x})\mathbf{x}\mathbf{u}\le\liminf_{n\to\infty}\alpha^n\nu_n^{(2)}\mathbf{u}\le\limsup_{n\to\infty}\alpha^n\nu_n^{(2)}\mathbf{u}\le\alpha\overline{d}^\dagger(\mathbf{x})\mathbf{x}\mathbf{u}.\tag{1.20}$$

In particular, if $\underline{d}^\dagger(\mathbf{x})=\overline{d}^\dagger(\mathbf{x})$ and $0\le d^\dagger\equiv\underline{d}^\dagger(\mathbf{x})<\infty$, then

$$\lim_{n\to\infty}\alpha^n\nu_n^{(2)}\mathbf{u}=\alpha d^\dagger\mathbf{x}\mathbf{u}.\tag{1.21}$$

That is, $\nu_n^{(2)}\mathbf{u}$ decays geometrically with rate α^{-1} as n goes to infinity.

Since this lemma can be proved in a similar way to Proposition 3.1 of [1], we omit its proof. For each $n \geq 0$, let

$$\nu_{-n}^{(-)} = \{\nu_{(-n)k}; k \geq 0\}, \qquad \nu_n^{(+)} = \{\nu_{nk}; k \geq 0\}, \qquad \nu_n^{(2)} = \{\nu_{kn}; k \in \mathbb{Z}\}.$$

Then, the next corollary follows from Theorem 1.1, Theorem 1.2 and Lemmas 1.4 and 1.5 similarly to Corollary 3.1 of [1].

Corollary 1.1. Define β_s for $s\pm, 2$ as

$$\beta_- = \sup\left\{\theta_1; \limsup_{n\to\infty} \nu_{(-1)n} e^{\underline{\theta_2} n} < \infty, (\theta_1, \theta_2) \in \mathscr{D}_1^{(-)}\right\},$$
$$\beta_+ = \sup\left\{\theta_1; \limsup_{n\to\infty} \nu_{1n} e^{\underline{\theta_2} n} < \infty, (\theta_1, \theta_2) \in \mathscr{D}_1^{(+)}\right\},$$
$$\beta_2 = \sup\left\{\eta_2; \limsup_{n\to\infty} \nu_{(-n)1} e^{\underline{\eta_1^{(-)}} n} < \infty, \right.$$
$$\left. \limsup_{n\to\infty} \nu_{n1} e^{\underline{\eta_1^{(+)}} n} < \infty, (\eta_1^{(-)}, \eta_1^{(+)}, \eta_2) \in \mathscr{D}_2\right\}.$$

Then, the weak upper decay rates $\overline{r}_-(i)$, $\overline{r}_+(i)$ and $\overline{r}_2(j)$ of ν_{-ni}, ν_{ni} and ν_{jn}, respectively, as $n \to \infty$ are uniformly bounded by $e^{-\beta_-}$, $e^{-\beta_+}$ and $e^{-\beta_2}$. In particular, for each $s = \pm, 2$, if $\beta_s = \log c_p(R^{(s)})$, then the weak decay rate r_s exists and $r_s = e^{-\beta_s}$. Furthermore, if the asymptotic decay of ν_{1n}, $\nu_{(-1)n}$ or ν_{n1} and $\nu_{-n(-1)}$ is exactly geometric as $n \to \infty$, then the corresponding stationary level distribution asymptotically decays in the exactly geometric form.

1.4 Answers to Decay Rate Problem

We are now in a position to answer to the decay rate problem. Since $\mathscr{D}_1^{(s)}$ for $s = \pm$ and $\mathscr{D}_2$ are compact sets, we can define, for $s = \pm$,

$$\theta_1^{(sc)} = \max\{\theta_1; (\theta_1, \theta_2) \in \mathscr{D}_1^{(s)}\}, \qquad \theta_2^{(sc)} = \min\{\theta_2; (\theta_1^{(sc)}, \theta_2) \in \mathscr{D}_1^{(s)}\},$$
$$\eta_2^{(c)} = \max\{\eta_2; (\eta_1^{(-)}, \eta_1^{(+)}, \eta_2) \in \mathscr{D}_2\},$$
$$\eta_1^{(sc)} = \max\{\eta_1^{(s)}; (\eta_1^{(-)}, \eta_1^{(+)}, \eta_2^{(c)}) \in \mathscr{D}_2\}.$$

Note that $\theta_1^{(sc)} = \log c_p(R^{(s)})$ for $s = \pm$ and $\eta_2^{(c)} = \log c_p(R^{(2)})$. Furthermore, $(\eta_1^{(-c)}, \eta_1^{(+c)}, \eta_2^{(c)})$ and $(\theta_1^{(sc)}, \theta_2^{(sc)})$ are in $\mathscr{D}_2$ and $\mathscr{D}_1^{(s)}$ for $s = \pm$, respectively.

Similarly to Theorem 4.1 of [1], we consider the following nonlinear optimization problems. Let, for $s = \pm$,

$$\alpha_s = \sup\{e^{\theta_1^{(s)}}; \theta_2^{(s)} \le \eta_2, \eta_1^{(-)} \le \theta_1^{(-)}, \eta_1^{(+)} \le \theta_1^{(+)}, \\ (\theta_1^{(-)}, \theta_2^{(-)}) \in \mathscr{D}_1^{(-)}, (\theta_1^{(+)}, \theta_2^{(+)}) \in \mathscr{D}_1^{(+)}, (\eta_1^{(-)}, \eta_1^{(+)}, \eta_2) \in \mathscr{D}_2\}, \quad (1.22)$$

$$\alpha_2 = \sup\{e^{\eta_2}; \theta_2^{(-)} \le \eta_2, \theta_2^{(+)} \le \eta_2, \eta_1^{(-)} \le \theta_1^{(-)}, \eta_1^{(+)} \le \theta_1^{(+)}, \\ (\theta_1^{(-)}, \theta_2^{(-)}) \in \mathscr{D}_1^{(-)}, (\theta_1^{(+)}, \theta_2^{(+)}) \in \mathscr{D}_1^{(+)}, (\eta_1^{(-)}, \eta_1^{(+)}, \eta_2) \in \mathscr{D}_2\}. \quad (1.23)$$

We can find solutions α_s for $s = \pm, 2$ in the following way.

Lemma 1.6. For the two sided DQBD process satisfying the assumptions (i) and (ii), suppose that its stationary distribution exists, which denoted by $\nu = \{\nu_{ij}\}$. Then, we have

$$\bar{r}_s \equiv \sup_i \bar{r}_s(i) \le \alpha_s^{-1}, \qquad s = \pm, 2. \quad (1.24)$$

Proof. We define the following functions of $u, u_-, u_+ \ge 0$.

$$f_-(u) = \sup\left\{\theta_1; \theta_2 \le u, (\theta_1, \theta_2) \in \mathscr{D}_1^{(-)}\right\},$$
$$f_+(u) = \sup\left\{\theta_2; \theta_1 \le u, (\theta_1, \theta_2) \in \mathscr{D}_1^{(+)}\right\},$$
$$f_2(u_-, u_+) = \sup\left\{\eta_2; \eta_1^{(-)} \le u_-, \eta_1^{(-)} \le u_+, (\eta_1^{(-)}, \eta_2^{(+)}, \eta_2) \in \mathscr{D}_2\right\}.$$

For convenience, let $\sigma_s = -\log \bar{r}_s$ for $s = \pm, 2$. Suppose that $0 \le u_s \le \sigma_s$, which implies that $\bar{r}_s(1) \le e^{-u_s}$ and $\bar{r}_2(s) \le e^{-u_2}$ for $s = \pm$. Then, Corollary 1.1 leads that

$$f_-(u_2) \le \sigma_-, \qquad f_+(u_2) \le \sigma_+, \qquad f_2(u_-, u_+) \le \sigma_2. \quad (1.25)$$

We next inductively define $u_s^{(n)}$ for $n = 0, 1, \ldots$ and $s = \pm, 2$ in the following way. Let $u_s^{(0)} = 0$, and

$$u_-^{(n+1)} = f_-\left(u_2^{(n)}\right), \qquad u_+^{(n+1)} = f_+\left(u_2^{(n)}\right), \qquad u_2^{(n+1)} = f_2\left(u_-^{(n+1)}, u_+^{(n+1)}\right).$$

Then, it is easy to see that $u_s^{(n)}$ is non decreasing in n, and satisfies (1.25) for $u_s = u_s^{(n)}$ for $s = \pm, 2$. Hence, Corollary 1.1 concludes

$$u_s^{(n)} \le \sigma_s, \qquad n = 0, 1, \ldots, \qquad s = \pm, 2.$$

On the other hand, from the definitions of α_s, it is easy to prove by induction that

$$u_s^{(\infty)} \equiv \lim_{n\to\infty} u_s^{(n)} \le \log \alpha_s, \qquad s = \pm, 2.$$

Then, it can be shown that the limits $u_s^{(\infty)}$ are attained in finitely many steps. The detailed proof of this can be found in the proof of Theorem 4.1 of [1]. Hence, we have

$$\log \alpha_s = \lim_{n\to\infty} u_s^{(n)} \le \sigma_s, \qquad s = \pm, 2.$$

Thus, we get (1.24). □

For each $s = \pm$, define the following four sets of conditions.

(sC1) $\eta_1^{(sc)} < \theta_1^{(sc)}$ and $\theta_2^{(sc)} < \eta_2^{(c)}$, ($s$C2) $\eta_1^{(sc)} < \theta_1^{(sc)}$ and $\eta_2^{(c)} \le \theta_2^{(sc)}$,
(sC3) $\theta_1^{(sc)} \le \eta_1^{(sc)}$ and $\theta_2^{(sc)} < \eta_2^{(c)}$, ($s$C4) $\theta_1^{(sc)} \le \eta_1^{(sc)}$ and $\eta_2^{(c)} \le \theta_2^{(sc)}$.

These conditions are exclusive and cover all the cases for each $s = \pm$. Furthermore, (sC4) is impossible since $\theta_1^{(sc)} \le \eta_1^{(sc)}$ implies that $\eta_2^{(c)} > \theta_2^{(sc)}$ due to the convexity of the set with boundary (1.9) and (1.11). For the other three cases for each $s = \pm$, we have to consider their combinations, so nine cases in total. For convenience, we denote the condition that ($-$Ci) and ($+$Cj) hold by C(i, j) for $i, j = 1, 2, 3$.

The next lemma shows how we can compute α_s for $s = \pm, 2$.

Lemma 1.7. Under the assumptions of Lemma 1.6, the α_-, α_+ and α_2 of (1.22) and (1.23) are obtained in either one of the following nine ways.

(c1) If C(1,1) holds, then $\alpha_- = \exp(\theta_1^{(-c)})$, $\alpha_+ = \exp(\theta_1^{(+c)})$ and $\alpha_2 = \exp(\eta_2^{(c)})$.
(c2) If C(1,2) holds, then $\alpha_- = \exp(\theta_1^{(-c)})$, $\alpha_2 = \exp(\eta_2^{(c)})$ and α_+ is the maximum value satisfying $(\log \alpha_+, \eta_2^{(c)}) \in \mathscr{D}_1^{(+)}$.
(c3) If C(2,1) holds, then $\alpha_+ = \exp(\theta_1^{(+c)})$, $\alpha_2 = \exp(\eta_2^{(c)})$ and α_- is the maximum value satisfying $(\log \alpha_-, \eta_2^{(c)}) \in \mathscr{D}_1^{(-)}$.
(c4) If C(1,3) holds, then $\alpha_+ = \exp(\theta_1^{(+c)})$, α_2 is the maximum value satisfying $(\theta_1, \log \alpha_2) \in \mathscr{D}_2^{(+)}$ with $\theta_1 \le \theta_1^{(+c)}$, and α_- is the maximal value satisfying $(\log \alpha_-, \theta_2) \in \mathscr{D}_1^{(-)}$ with $\theta_2 \le \alpha_2$.
(c5) If C(3,1) holds, then $\alpha_- = \exp(\theta_1^{(-c)})$, α_2 is the maximum value satisfying $(\theta_1, \log \alpha_2) \in \mathscr{D}_2^{(-)}$ with $\theta_1 \le \theta_1^{(-c)}$, and α_+ is the maximal value satisfying $(\log \alpha_+, \theta_2) \in \mathscr{D}_1^{(+)}$ with $\theta_2 \le \alpha_2$.
(c6) If C(2,2) holds, then $\alpha_2 = \exp(\eta_2^{(c)})$ and α_s is the maximum value satisfying $(\log \alpha_s, \eta_2^{(c)}) \in \mathscr{D}_1^{(s)}$ for $s = \pm$.
(c7) If C(2,3) holds, then $\alpha_+ = \exp(\theta_1^{(+c)})$, α_2 is the maximum value satisfying $(\theta_1, \log \alpha_2) \in \mathscr{D}_2^{(+)}$ with $\theta_1 \le \theta_1^{(+c)}$, and α_- is the maximum value satisfying $(\log \alpha_-, \log \alpha_2) \in \mathscr{D}_1^{(-)}$.
(c8) If C(3,2) holds, then $\alpha_- = \exp(\theta_1^{(-c)})$, α_2 is the maximum value satisfying $(\theta_1, \log \alpha_2) \in \mathscr{D}_2^{(-)}$ with $\theta_1 \le \theta_1^{(-c)}$, and α_+ is the maximum value satisfying $(\log \alpha_+, \log \alpha_2) \in \mathscr{D}_1^{(+)}$.
(c9) If C(3,3) holds, then $\alpha_- = \exp(\theta_1^{(-c)})$, $\alpha_- = \exp(\theta_1^{(+c)})$ and α_2 is the maximum value satisfying $(\theta_1^{(-)}, \theta_1^{(+)}, \log \alpha_2) \in \mathscr{D}_2$ with $\theta_1^{(-)} \le \theta_1^{(-c)}$ and $\theta_1^{(+)} \le \theta_1^{(+c)}$.

This theorem can be proved in the same way as Lemma 4.2 of [1]. So, instead of proving it, we give figures to explain how those decay rates are obtained. They can be found in Figs. 1.2, 1.3 and 1.4. Since cases (c3), (c5) and (c7) are symmetric with (c2), (c4) and (c6), respectively, we omit their figures. We shall see more figures for specific examples in Sect. 1.5.

Theorem 1.3. Under the assumptions of Lemma 1.6, we have $r_s = \alpha_s^{-1}$ for $s = \pm, 2$. Namely, α_-^{-1}, α_+^{-1} and α_2^{-1} are the weak decay rates of $\nu_{-n}^{(-)}$, $\nu_n^{(+)}$ and $\nu_n^{(2)}$, respectively, as $n \to \infty$. Furthermore, the marginal probabilities, $\nu_{-n}^{(-)}\mathbf{1}$, $\nu_n^{(+)}\mathbf{1}$ and $\nu_n^{(2)}\mathbf{1}$, have the same decay rates α_-^{-1}, α_+^{-1} and α_2^{-1}, respectively, if they are less than 1, respectively.

Proof. We first consider $\bar{r}_s(1)$ for $s = \pm, 2$, which are the weak upper decay rates of ν_{-n1}, ν_{n1} and ν_{1n} as $n \to \infty$, respectively, are obtained by (1.24). Hence, Lemmas 1.1, 1.4 and 1.5 yield

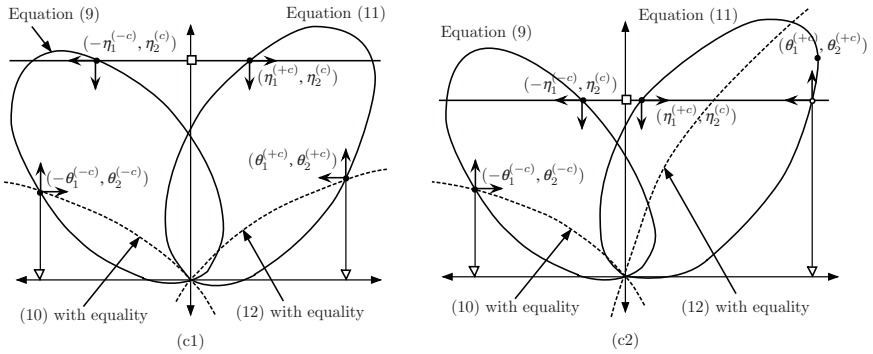


Fig. 1.2 Typical examples for (c1) and (c2).

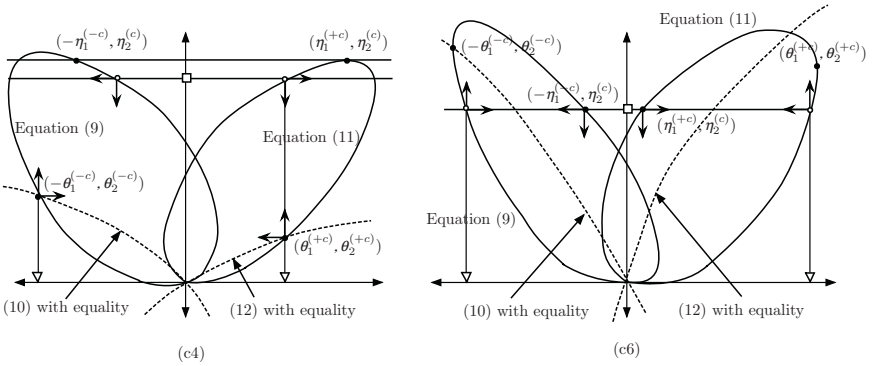


Fig. 1.3 Typical examples for (c4) and (c6).

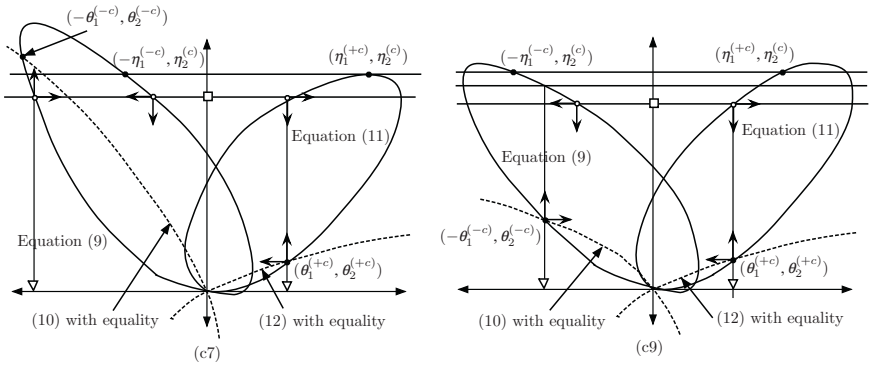


Fig. 1.4 Typical examples for (c7) and (c9).

$$c_p\left(R^{(s)}\right)^{-1} \le \underline{r}_s(i) \le \overline{r}_s(i) \le \alpha_s^{-1}, \quad i \in \mathbb{Z}_+ \text{ for } s = \pm \text{ and } i \in \mathbb{Z} \text{ for } s = 2.$$

From Lemma 1.7, at least one of α_-, α_+ and α_2 agree with the corresponding convergence parameter $c_p(R^{(s)})$. Hence, we have $r_s = \alpha_s^{-1}$ at least for one s. This together with Corollary 1.1 and Lemmas 1.4 and 1.5 conclude that the same equality must hold for the other s's. This completes the proof. □

We can refine the decay rates in this theorem from weak to exact ones in a similar way as Theorem 4.2 of [1] using Proposition 3.1 of [1] and Lemmas 1.4 and 1.5 for the case that the decay rates are exactly geometric. However, for the other cases, we can not directly use Theorem 5 of [10] which was used in [1] since the level or background state is two sided. Thus, we here only present the case that the exactly geometric decay occurs. We omit its proof since it is similar to Theorem 4.2 of [1].

Theorem 1.4. Under the assumptions of Theorem 1.3 with $\alpha_- > 1$, $\alpha_+ > 1$ or $\alpha_2 > 1$, let, for $s = \pm$,

$$\mathscr{D}_0^{(s)} = \left\{ (s\theta_1, \theta_2) \in \mathbb{R}^2; \mathrm{E}\left[e^{\theta_1 X_1^{(s)} + \theta_2 X_2^{(s)}} \right] = 1 \right\},$$
$$\theta_i^{s\max} = \arg\max_{(\theta_1,\theta_2) \in \mathscr{D}_0^{(s)}} \{\theta_i\}, \qquad i = 1, 2.$$

Then, we have the exactly geometric decay rates for the following cases.

(d1) If either (-C2) or (-C3) holds, then both asymptotic decays of $\{\nu_{(-n)k}\}$ and $\{\nu_{\ell n}\}$ as $n \to \infty$ are exactly geometric with the decay rates α_-^{-1} and α_2^{-1}, respectively.

(d2) If either (+C2)) or (+C3) holds, then both asymptotic decays of $\{\nu_{nk}\}$ and $\{\nu_{\ell n}\}$ as $n \to \infty$ are exactly geometric with the decay rates α_+^{-1} and α_2^{-1}, respectively.

(d3) If (C1) holds and if $\theta_1^{s\max} \notin \mathscr{D}_1^{(s)}$ and $\eta_1^{(c)} < \theta_1^{(sc)}$, then the asymptotic decay of $\{\nu_{(sn)k}\}$ ($\{\nu_{\ell n}\}$) as $n \to \infty$ is exactly geometric with the rate α_s^{-1} (α_2^{-1}) for $s = \pm$.

1.5 Generalized Join Shortest Queue

Let us apply Theorems 1.3 and 1.4 to the generalized join shortest queue which is studied in [3], [6] and explained in Sect. 1.1. We first introduce notations for this model. It has two parallel queues, numbered as queues 1 and 2. For each $i = 1, 2$, queue i serves customers in the First-Come First-Served manner with *i.i.d.* service times subject to the exponential distribution with rate μ_i. There are three exogenous Poisson arrival streams. The first and second streams go to queues 1 and 2 with the mean arrival rate λ_1 and λ_2, respectively, while arriving customers in the third stream with the mean rate δ choose the shorter queue with tie breaking. The decay rates does not depend on the probability that customer with tie breaking choose queue 1, so we simply assume it to be $1/2$.

We are interested to see how the stationary tail probabilities of the shorter queue lengths and the difference of the two queues decay. Due to the dedicated stream to each queue, this problem is much harder than the one for the standard joining the shortest queue. Since we only consider the stationary distribution, we can formulate this continuous time model as a discrete time Markov chain. For this, we assume without loss of generality that

$$\lambda_1 + \lambda_2 + \mu_1 + \mu_2 + \delta = 1.$$

Let Q_{1t} and Q_{2t} be the queue lengths including customers being served at time $t = 0, 1, \ldots$, and let $L_{1t} = Q_{2t} - Q_{1t}$ and $L_{2t} = \min(Q_{1t}, Q_{2t})$. It is not hard to see that (L_{1t}, L_{2t}) is a skip free random walk on each region $(\mathbb{Z}_+ \cup \setminus\{0\}) \times (\mathbb{Z}_+ \setminus \{0\})$ reflected at the boundary $\mathbb{Z} \times \{0\}$ and has different transitions at $\{0\} \times \mathbb{Z}_+$ (see Fig. 1.5).

Then, the transition probabilities are give by

$$p^{(-)}_{(-1)0} = \lambda_1, \quad p^{(-)}_{(-1)(-1)} = \mu_2, \quad p^{(-)}_{10} = \mu_1, \quad p^{(-)}_{11} = \lambda_2 + \delta,$$
$$p^{(+)}_{10} = \lambda_2, \quad p^{(+)}_{1(-1)} = \mu_1, \quad p^{(+)}_{(-1)0} = \mu_2, \quad p^{(+)}_{(-1)1} = \lambda_1 + \delta,$$
$$p^{(2)}_{10} = \lambda_2 + \frac{\delta}{2}, \quad p^{(2)}_{1(-1)} = \mu_1, \quad p^{(2)}_{(-1)(-1)} = \mu_2, \quad p^{(2)}_{(-1)0} = \lambda_1 + \frac{\delta}{2},$$

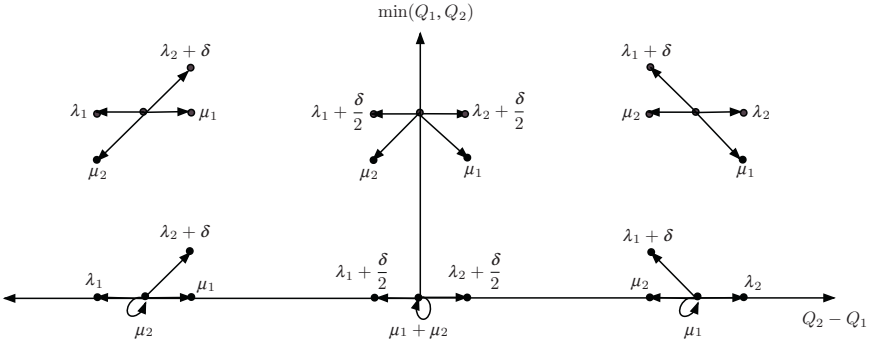


Fig. 1.5 State transitions for the generalized shortest queue.

$$p^{(1-)}_{(-1)0} = \lambda_1, \quad p^{(1-)}_{00} = \mu_2, \quad p^{(1+)}_{10} = \mu_1, \quad p^{(1-)}_{11} = \lambda_2 + \delta,$$
$$p^{(1+)}_{10} = \lambda_2, \quad p^{(1+)}_{00} = \mu_1, \quad p^{(1+)}_{(-1)0} = \mu_2, \quad p^{(1+)}_{(-1)1} = \lambda_1 + \delta,$$
$$p^{(0)}_{10} = \lambda_2 + \frac{\delta}{2}, \quad p^{(0)}_{00} = \mu_1 + \mu_2, \quad p^{(0)}_{(-1)0} = \lambda_1 + \frac{\delta}{2},$$

where all other transitions are null. To exclude obvious cases, we assume that δ, μ_1, μ_2 are all positive.

Denote traffic intensities by

$$\rho_1 = \frac{\lambda_1}{\mu_1}, \qquad \rho_2 = \frac{\lambda_2}{\mu_2}, \qquad \rho = \frac{\lambda_1 + \lambda_2 + \delta}{\mu_1 + \mu_2}.$$

Then, it is known that this generalized join shortest queue is stable if and only if $\rho_1 < 1, \rho_2 < 1$ and $\rho < 1$ (e.g., see [6]). This stability condition is assumed throughout this section. We will also use the following notation, which were introduced and shown to be very useful in computations in [3].

$$\gamma_1 = \mu_1 \rho^2 + \lambda_2, \qquad \gamma_2 = \mu_2 \rho^2 + \lambda_1.$$

We apply Theorem 1.3 to this model. For this, we need to compute $\theta^{(-c)}$, $\theta^{(+c)}$ and $\eta_2^{(c)}$. In the view of Theorems 1.1 and 1.2, they are obtained if we can solve the following three sets of equations.

$$\mathrm{E}\left[e^{\theta_1 X_1^{(-)} + \theta_2 X_2^{(-)}}\right] = 1, \quad \varphi_0^{(1-)}(\theta_1) + \varphi_1^{(1-)}(\theta_1) e^{\theta_2} = 1, \tag{1.26}$$

$$\mathrm{E}\left[e^{\theta_1 X_1^{(+)} + \theta_2 X_2^{(+)}}\right] = 1, \quad \varphi_0^{(1+)}(\theta_1) + \varphi_1^{(1+)}(\theta_1) e^{\theta_2} = 1, \tag{1.27}$$

$$\mathrm{E}\left[e^{\eta_1^{(-)} X_1^{(-)} + \eta_2 X_2^{(-)}}\right] = 1, \quad \mathrm{E}\left[e^{\eta_1^{(+)} X_1^{(+)} + \eta_2 X_2^{(+)}}\right] = 1,$$
$$\varphi_{-1}^{(2)}(\eta_2) e^{-\eta_1^{(-)}} + \varphi_0^{(2)}(\eta_2) + \varphi_1^{(2)}(\eta_2) e^{\eta_1^{(+)}} = 1. \tag{1.28}$$

For convenience, let $z = e^{-\theta_1}$ and $\xi = e^{\theta_2}$ in (1.26). Then, we have

$$\lambda_1 z + \mu_2 z \xi^{-1} + \mu_1 z^{-1} + (\lambda_2 + \delta) z^{-1} \xi = 1, \tag{1.29}$$
$$\lambda_1 z + \mu_2 + \mu_1 z^{-1} + (\lambda_2 + \delta) z^{-1} \xi = 1. \tag{1.30}$$

Solving these equations for $z \neq 1$, we have $z = \xi = \rho_1^{-1}$. For $z = \rho_1^{-1}$, (1.29) yields $\xi = \rho_1^{-1}, \frac{\mu_2}{\lambda_2 + \delta} \rho_1^{-1}$. Note that $\rho_1^{-1} < \frac{\mu_2}{\lambda_2 + \delta} \rho_1^{-1}$ if and only if $\mu_2 > \lambda_2 + \delta$. Hence, reminding the definitions of $\theta_i^{-\max}$:

$$\theta_1^{-\max} = \max\{\log z;\ (1.29) \text{ holds.}\}, \qquad \theta_2^{-\max} = \max\{\log \xi;\ (1.29) \text{ holds.}\},$$

we have

$$\left(\theta_1^{(-c)},\theta_2^{(-c)}\right)=\begin{cases}(\log\rho_1^{-1},\log\rho_1^{-1}), & \mu_2>\lambda_2+\delta\\ (\theta_1^{-\max},\theta_2^{-\max}), & \mu_2\le\lambda_2+\delta.\end{cases}\tag{1.31}$$

It is also notable that $\theta_1^{-\max}\ge\log\rho_1^{-1}$, so we always have that $\theta_1^{(-c)}\ge\log\rho_1^{-1}$.

Remark 1.3. The $\theta_i^{-\max}$ for $i=1,2$ are computed from their definitions as

$$\theta_1^{-\max}=\log\frac{1}{2\lambda_1}\left(1-2\sqrt{\mu_2(\lambda_2+\delta)}+\zeta_1^{(-)}\right),$$
$$\theta_2^{-\max}=\log\frac{1-4(\lambda_1\mu_1+(\lambda_2+\delta)\mu_2)+\zeta_2^{(-)}}{8\lambda_1(\lambda_2+\delta)},$$

where

$$\zeta_1^{(-)}=\sqrt{1+4(\mu_2(\lambda_2+\delta)-\sqrt{\mu_2(\lambda_2+\delta)}-\lambda_1\mu_1)}\,,$$
$$\zeta_2^{(-)}=\sqrt{(1-4(\lambda_1\mu_1+(\lambda_2+\delta)\mu_2))^2-64(\lambda_2+\delta)\lambda_1\mu_1\mu_2}\,.$$

Similarly, letting $z=e^{\theta_1}$ and $\xi=e^{\theta_2}$ in (1.27),

$$\lambda_2 z+\mu_1 z\xi^{-1}+\mu_2 z^{-1}+(\lambda_1+\delta)z^{-1}\xi=1,\tag{1.32}$$
$$\lambda_2 z+\mu_1+\mu_2 z^{-1}+(\lambda_1+\delta)z^{-1}\xi=1.\tag{1.33}$$

Solving these equations for $z\ne1$, we have $z=\xi=\rho_2^{-1}$. For $z=\rho_2^{-1}$, (1.32) yields $\xi=\rho_2^{-1},\frac{\mu_1}{\lambda_1+\delta}\rho_2^{-1}$. Reminding that

$$\theta_1^{+\max}=\max\{\log z;\ (1.32)\text{ holds.}\},\qquad \theta_2^{+\max}=\max\{\log\xi;\ (1.32)\text{ holds.}\},$$

we have that $\theta_1^{(+c)}\ge\log\rho_2^{-1}$ and

$$\left(\theta_1^{(+c)},\theta_2^{(+c)}\right)=\begin{cases}(\log\rho_2^{-1},\log\rho_2^{-1}), & \mu_1>\lambda_1+\delta\\ (\theta_1^{+\max},\theta_2^{+\max}), & \mu_1\le\lambda_1+\delta.\end{cases}\tag{1.34}$$

We also consider to solve (1.28). In this case, let $\xi=e^{\eta_2}$, $z_1=e^{-\eta_1^{(-)}}$ and $z_2=e^{\eta_1^{(+)}}$. Then, (1.28) becomes

$$\lambda_1 z_1+\mu_2 z_1\xi^{-1}+\mu_1 z_1^{-1}+(\lambda_2+\delta)z_1^{-1}\xi=1,\tag{1.35}$$
$$\lambda_2 z_2+\mu_1 z_2\xi^{-1}+\mu_2 z_2^{-1}+(\lambda_1+\delta)z_2^{-1}\xi=1,\tag{1.36}$$
$$\left(\lambda_1+\frac{\delta}{2}\right)z_1+\mu_2 z_1\xi^{-1}+\mu_1 z_2\xi^{-1}+\left(\lambda_2+\frac{\delta}{2}\right)z_2=1.\tag{1.37}$$

These equations have been solved in [3]. That is, if $z\ne1$, then $\xi=\rho^{-2}$ and $z_1=z_2=\rho^{-1}$. For $\xi=\rho^{-2}$, the first equation has solutions $z_1=\rho^{-1},\frac{\gamma_1+\delta}{\gamma_2}\rho^{-1}$, and

the second equation yields $z_2 = \rho^{-1}, \frac{\gamma_2+\delta}{\gamma_1}\rho^{-1}$. In this case, $\eta_2^{(c)}$ is obtained as the maximum ξ that satisfies (1.35), (1.36) and

$$\left(\lambda_1 + \frac{\delta}{2}\right) z_1 + \mu_2 z_1 \xi^{-1} + \mu_1 z_2 \xi^{-1} + \left(\lambda_2 + \frac{\delta}{2}\right) z_2 \leq 1. \tag{1.38}$$

Thus, we need to solve a convex optimization problem. We here already know that $(z_1, z_2, \xi) = (1,1,1), (\rho^{-1}, \rho^{-1}, \rho^{-2})$ are the extreme points of the constrains. To identify the latter point on the convex curves (1.35) and (1.36), it is convenient to introduce the following classifications:

$$\gamma_2 + \delta > \gamma_1, \qquad \gamma_1 + \delta > \gamma_2, \tag{1.39}$$
$$\gamma_2 + \delta \leq \gamma_1, \qquad \gamma_1 + \delta > \gamma_2, \tag{1.40}$$
$$\gamma_2 + \delta > \gamma_1, \qquad \gamma_1 + \delta \leq \gamma_2, \tag{1.41}$$

where we exclude the case that $\gamma_2 + \delta \leq \gamma_1$ and $\gamma_1 + \delta \leq \gamma_2$, which is impossible since $\delta > 0$. Note that (1.39) is equivalent to

$$|\gamma_1 - \gamma_2| < \delta,$$

which is introduced and called strongly pooled in [6].

We now find $\eta_2^{(c)}$ by solving the convex optimization problem.

Lemma 1.8. If the strongly pooled condition (1.39) holds, then

$$\eta_2^{(c)} = \log \rho^{-2}, \qquad \eta_1^{(-c)} = \eta_1^{(+c)} = \log \rho^{-1}.$$

Otherwise, if (1.40) holds, then

$$(\eta_2^{(c)}, \eta_1^{(-c)}, \eta_1^{(+c)}) = \left(\theta_2^{-\max}, \log \frac{e^{\eta_2^{(c)}}}{2(\lambda_1 e^{\eta_2^{(c)}} + \mu_2)}, \arg \max_{(\theta_1, \eta_2^{(c)}) \in \mathscr{D}_0^{(+)}} \theta_1\right),$$

and, if (1.41) holds, then

$$(\eta_2^{(c)}, \eta_1^{(-c)}, \eta_1^{(+c)}) = \left(\theta_2^{+\max}, \arg \max_{(\theta_1, \eta_2^{(c)}) \in \mathscr{D}_0^{(-)}} \theta_1, \log \frac{e^{\eta_2^{(c)}}}{2(\lambda_2 e^{\eta_2^{(c)}} + \mu_1)}\right).$$

We defer the proof of this lemma to Appendix B.

We next consider to apply Theorem 1.3 to the generalized join shortest queue. To this end, we introduce another classifications.

$$\rho_1 < \rho, \quad \rho_2 < \rho, \tag{1.42}$$
$$\rho_1 \geq \rho, \quad \rho_2 < \rho, \tag{1.43}$$
$$\rho_1 < \rho, \quad \rho_2 \geq \rho, \tag{1.44}$$

where we do not consider the case that $\rho_1 \geq \rho$ and $\rho_2 \geq \rho$, which is impossible since $\delta > 0$. The condition (1.42) is referred to as a weakly pooled condition in [6].

Under the conditions (1.39) and (1.42), the asymptotic decay of

$$P(\min(Q_1, Q_2) = n, Q_1 - Q_2 = \ell), \qquad n \to \infty$$

is shown to be exactly geometric with decay rate ρ^2 for each fixed ℓ in [3], [6]. This is the only known results for the decay rate for the minimum of the two queues. Using the two sets of the classifications, we can answer to the decay rate problem for all the cases but for the weak decay rates.

Theorem 1.5. For the generalized join shortest queue with two queues, suppose that the stability conditions $\rho < 1$, $\rho_1 < 1$ and $\rho_2 < 1$ are satisfied. Then, the weak decay rate r_2 exists for the minimum of the two queues in the sense of marginal distribution as well as jointly with each fixed difference of the two queues, and one of the following three cases occurs.

(g1) If (1.39) holds, then either one of the following cases happens.

(g1a) (1.42) implies $r_2 = \rho^2$.

(g1b) (1.43) implies $r_2 = \dfrac{\lambda_2 + \delta}{\mu_2}\rho_1$.

(g1c) (1.44) implies $r_2 = \dfrac{\lambda_1 + \delta}{\mu_1}\rho_2$.

(g2) If (1.40) holds, then either one of the following cases happens.

(g2a) (1.42) implies $r_2 = \begin{cases} e^{-\theta_2^{-}\max}, & \eta_1^{(+c)} \leq \theta_1^{(+c)} \\ \dfrac{\lambda_1 + \delta}{\mu_1}\rho_2, & \eta_1^{(+c)} > \theta_1^{(+c)}. \end{cases}$

(g2b) (1.43) implies

$$r_2 = \begin{cases} e^{-\theta_2^{-}\max}, & \eta_1^{(-c)} < \log \rho_1^{-1}, \eta_1^{(+c)} < \theta_1^{(+c)} \\ \dfrac{\lambda_2 + \delta}{\mu_2}\rho_1, & \eta_1^{(-c)} \geq \log \rho_1^{-1}, \eta_1^{(+c)} < \theta_1^{(+c)} \\ \dfrac{\lambda_1 + \delta}{\mu_1}\rho_2, & \eta_1^{(-c)} < \log \rho_1^{-1}, \eta_1^{(+c)} \geq \theta_1^{(+c)} \\ \min\left(\dfrac{\lambda_2 + \delta}{\mu_2}\rho_1, \dfrac{\lambda_1 + \delta}{\mu_1}\rho_2\right), & \eta_1^{(-c)} \geq \log \rho_1^{-1}, \eta_1^{(+c)} \geq \theta_1^{(+c)}. \end{cases}$$

(g2c) (1.44) implies $r_2 = \dfrac{\lambda_1 + \delta}{\mu_1}\rho_2$.

(g3) If (1.41) holds, then either one of the following cases happens.

(g3a) (1.42) implies $r_2 = \begin{cases} e^{-\theta_2^{+}\max}, & \eta_1^{(-c)} \leq \theta_1^{(-c)} \\ \dfrac{\lambda_2 + \delta}{\mu_2}\rho_1, & \eta_1^{(-c)} > \theta_1^{(-c)}. \end{cases}$

(g3b) (1.43) implies $r_2 = \dfrac{\lambda_2+\delta}{\mu_2}\rho_1$.

(g3c) (1.44) implies

$$r_2 = \begin{cases} e^{-\theta_2^{+\max}}, & \eta_1^{(-c)} < \theta_1^{(-c)}, \eta_1^{(+c)} < \log\rho_2^{-1} \\ \dfrac{\lambda_2+\delta}{\mu_2}\rho_1, & \eta_1^{(-c)} \geq \theta_1^{(-c)}, \eta_1^{(+c)} < \log\rho_2^{-1} \\ \dfrac{\lambda_1+\delta}{\mu_1}\rho_2, & \eta_1^{(-c)} < \theta_1^{(-c)}, \eta_1^{(+c)} \geq \log\rho_2^{-1} \\ \min\left(\dfrac{\lambda_2+\delta}{\mu_2}\rho_1, \dfrac{\lambda_1+\delta}{\mu_1}\rho_2\right), & \eta_1^{(-c)} \geq \theta_1^{(-c)}, \eta_1^{(+c)} \geq \log\rho_2^{-1}. \end{cases}$$

Furthermore, the decay rates are exactly geometric for the cases (g1), (g2) unless $\eta_1^{(-c)} = \theta_1^{-\max}$ and (g3) unless $\eta_1^{(+c)} = \theta_1^{+\max}$.

Proof. This theorem is concluded applying Theorem 1.3 together with Lemma 1.7 for $(\theta_1^{(sc)}, \theta_2^{(sc)})$ for $s = \pm$ and Lemma 1.8. We first consider case (g1a). In this case, we suppose that the strongly pooled condition (1.39) and the weakly pooled condition (1.42) hold, then $\theta_1^{(sc)} \geq \eta_1^{(sc)}$ for $s = \pm$ from (1.31), (1.34) and Lemma 1.6. Hence, either one of C(1,1). C(1,2) or C(2,1) occurs in Lemma 1.7, which implies that $r_2 = \alpha_2^{-1} = e^{-\eta_2^{(c)}} = \rho^2$.

We next consider (g1b). In this case, (1.39) and (1.43) are assumed. Note that $\rho_1 \geq \rho$ in (1.43) implies that

$$\mu_2 \geq \frac{\mu_1}{\lambda_1}(\lambda_2+\delta) > \lambda_2+\delta.$$

Hence, we always have $\theta_1^{(-c)} = \log\rho_1^{-1}$ from (1.31) in this case. Since $\log\rho_1^{-1} \leq \log\rho^{-1} = \eta_1^{(-c)}$ and $\eta_1^{+c} = \log\rho^{-1} < \log\rho_2^{-1} \leq \theta_1^{(+c)}$, we have (g1b) from (c5) or (c8) of Lemma 1.7.

The other cases are similarly proved. So, we omit their details. □

To visualize the results of Theorem 1.5, we draw equations (1.26) and (1.27) on the (θ_1, θ_2) plane simultaneously for some examples. We here consider the four cases (g1a), (g1b), (g2a) and (g2b).

These four cases are given in Figs. 1.6 and 1.7. In case (g1a) of Fig. 1.6,

$$\lambda_1 = \frac{1}{16}, \quad \lambda_2 = \frac{3}{16}, \quad \delta = \frac{1}{8}, \quad \mu_1 = \frac{1}{4}, \quad \mu_2 = \frac{3}{8},$$

which implies that $\rho_1 = \frac{1}{4}$, $\rho_2 = \frac{1}{2}$ and $\rho = \frac{3}{5}$. In case (g1b),

$$\lambda_1 = \frac{6}{29}, \quad \lambda_2 = \frac{4}{29}, \quad \delta = \frac{1}{29}, \quad \mu_1 = \frac{10}{29}, \quad \mu_2 = \frac{8}{29},$$

which implies that $\rho_1 = 0.6$, $\rho_2 = 0.5$ and $\rho = \frac{11}{18}$.

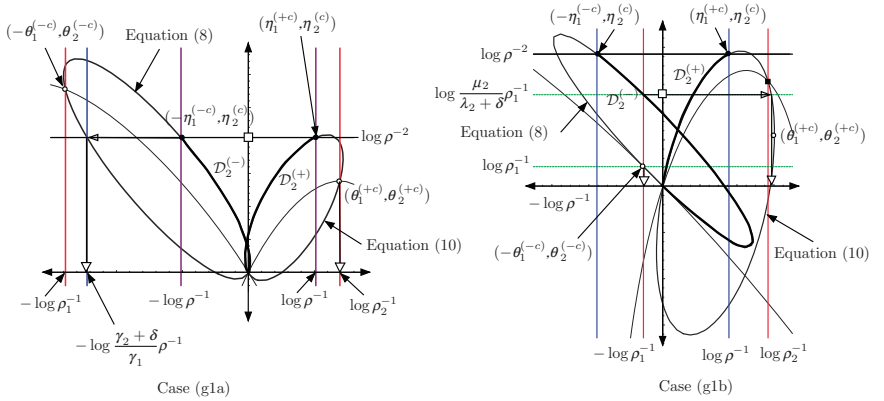


Fig. 1.6 The decay rates for strongly pooled (1.39): case (g1a) for (1.42) and case (g1b) for (1.43).

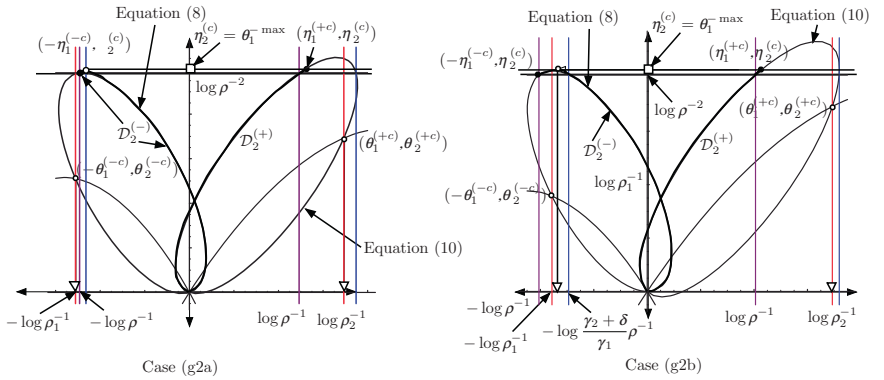


Fig. 1.7 The decay rates for not strongly pooled (1.40): case (g2a) for (1.42) and case (g2b) for (1.43).

Figure 1.7 shows the case where the weakly pooled condition (1.39) does not hold. In case (g2a), we set

$$\lambda_1 = \frac{9}{170}, \quad \lambda_2 = \frac{51}{170}, \quad \delta = \frac{1}{17}, \quad \mu_1 = \frac{1}{17}, \quad \mu_2 = \frac{9}{17},$$

which implies that $\rho_1 = 0.9$, $\rho_2 = \frac{17}{30}$ and $\rho = 0.7$. This example shows that the strongly pooled condition (1.39) does not imply the weakly pooled condition (1.42). In case (g2b),

$$\lambda_1 = \frac{7}{30}, \quad \lambda_2 = \frac{2}{15}, \quad \delta = \frac{1}{30}, \quad \mu_1 = \frac{10}{30}, \quad \mu_2 = \frac{8}{30},$$

which implies that $\rho_1 = 0.7$, $\rho_2 = 0.5$ and $\rho = \frac{2}{3}$.

Similarly to Theorem 1.5, we can get the following corollary for the decay rates for the difference of the two queues. We omit its proof since it is parallel to the arguments in Theorem 1.5.

Corollary 1.2. Under the assumptions of Theorem 1.5, the weak decay rates r_- and r_+ for the difference $Q_2 - Q_1$ in the negative and positive directions, respectively, exist in the sense of marginal distributions as well as jointly with each fixed minimum of the two queues, and we have the following cases, where $(\theta_1^{(-c)}, \theta_2^{(-c)})$ and $(\theta_1^{(+c)}, \theta_2^{(+c)})$ are given by (1.31) and (1.34), respectively, and

$$t_-(v) = \min\{z^{-1};\ (1.29) \text{ for } \xi = v^{-1}\}, \quad t_+(v) = \min\{z^{-1};\ (1.32) \text{ for } \xi = v^{-1}\}.$$

(h1) If (1.39) holds, then either one of the following cases happens.

(h1a) (1.42) implies

$$r_- = \begin{cases} e^{-\theta_1^{(-c)}}, & \theta_2^{(-c)} \le \log \rho^{-2} \\ \dfrac{\gamma_2}{\gamma_1 + \delta}\rho, & \theta_2^{(-c)} > \log \rho^{-2}, \end{cases} \tag{1.45}$$

$$r_+ = \begin{cases} e^{-\theta_1^{(+c)}}, & \theta_2^{(+c)} \le \log \rho^{-2} \\ \dfrac{\gamma_1}{\gamma_2 + \delta}\rho, & \theta_2^{(+c)} > \log \rho^{-2}. \end{cases} \tag{1.46}$$

(h1b) (1.43) implies with $r_2 = \frac{(\lambda_2+\delta)}{\mu_2}\rho_1$ that

$$r_- = \rho_1, \qquad r_+ = \begin{cases} e^{-\theta_1^{(+c)}}, & \theta_2^{(+c)} \le \log r_2^{-1} \\ t_+(r_2), & \theta_2^{(+c)} > \log r_2^{-1}. \end{cases} \tag{1.47}$$

(h1c) If (1.44) implies with $r_2 = \frac{(\lambda_1+\delta)}{\mu_1}\rho_2$ that

$$r_- = \begin{cases} e^{-\theta_1^{(-c)}}, & \theta_2^{(-c)} \le \log r_2^{-1}, \\ t_-(r_2), & \theta_2^{(-c)} > \log r_2^{-1}, \end{cases} \qquad r_+ = \rho_2. \tag{1.48}$$

(h2) If (1.40) holds, then either one of the following cases happens.

(h2a) (1.42) implies

$$(r_-, r_+) = \begin{cases} \left(e^{-\theta_1^{(-c)}}, \min(e^{-\theta_1^{(+c)}}, t_+(e^{-\eta_2^{(c)}}))\right), & \theta_1^{(+c)} \ge \eta_1^{(+c)} \\ \left(e^{-\theta_1^{(-c)}}, \rho_2\right), & \theta_1^{(+c)} < \eta_1^{(+c)}. \end{cases} \tag{1.49}$$

(h2b) (1.43) implies

$$(r_-, r_+) = \begin{cases} (\rho_1, e^{-\theta_1^{(+c)}}), & \eta_1^{(-c)} < \log \rho_1^{-1}, \eta_1^{(+c)} < \theta_1^{(+c)} \\ (\rho_1, \min(e^{-\theta_1^{(+c)}}, t_+(r_2)), & \eta_1^{(-c)} \geq \log \rho_1^{-1}, \eta_1^{(+c)} < \theta_1^{(+c)} \\ (\min(e^{-\theta_1^{(-c)}}, t_-(r_2), \rho_2), & \eta_1^{(-c)} < \log \rho_1^{-1}, \eta_1^{(+c)} \geq \theta_1^{(+c)} \\ (\rho_1, \min(e^{-\theta_1^{(+c)}}, t_+(r_2)), & \eta_1^{(-c)} \geq \log \rho_1^{-1}, \eta_1^{(+c)} \geq \theta_1^{(+c)} \\ & \text{and } \dfrac{\lambda_2+\delta}{\mu_2}\rho_1 < \dfrac{\lambda_1+\delta}{\mu_1}\rho_2 \\ (\min(e^{-\theta_1^{(-c)}}, t_-(r_2), \rho_2), & \eta_1^{(-c)} \geq \log \rho_1^{-1}, \eta_1^{(+c)} \geq \theta_1^{(+c)} \\ & \text{and } \dfrac{\lambda_2+\delta}{\mu_2}\rho_1 \geq \dfrac{\lambda_1+\delta}{\mu_1}\rho_2, \end{cases} \tag{1.50}$$

where $t_+ = \max\{z;\ (1.32) \text{ for } \xi = \log r_2^{-1}\}$ and $r_2 = \dfrac{(\lambda_2+\delta)}{\mu_2}\rho_1$.

(h2c) (1.44) implies with $r_2 = \dfrac{\lambda_1+\delta}{\mu_1}\rho_2$ that

$$(r_-, r_+) = \begin{cases} (e^{-\theta_1^{(-c)}}, \rho_2), & \theta_2^{(-c)} < \log r_2^{-1} \\ (t_-(r_2), \rho_2), & \theta_2^{(-c)} \geq \log r_2^{-1}. \end{cases} \tag{1.51}$$

Furthermore, the decay rates are exactly geometric unless either $r_- = e^{-\theta_1^{(-c)}}$ with $\theta_1^{(-c)} = \theta_1^{-\max}$ or $r_+ = e^{-\theta_1^{(+c)}}$ with $\theta_1^{(+c)} = \theta_1^{+\max}$.

Remark 1.4. In this corollary, the case that (1.41) holds is not considered. However, this case can be easily obtained by interchanging the roles of queues 1 and 2 in case (h2).

1.6 Remarks on Existence Results

We remark how our results include the existence results. The exactly geometric rate $r_2 = \rho^2$ is obtained under the conditions (1.39) and (1.42) in [3], [6]. Our results cover all the possible cases although the decay rates are generally of the weak sense. We also note that there are some errors in Theorem 3.2 of [3]. They can be corrected by Corollary 1.2. Namely, the additional conditions (3.16) and (3.18) there are not sufficient to get the decay rates. They are used for all the terms in the sums of (3.15) and (3.17) to be positive. However, this is different from the corresponding eigenvectors to be positive. The right conditions are $\theta_2^{(+c)} \geq \log \rho^{-2}$ and $\theta_2^{(-c)} \geq \log \rho^{-2}$, respectively, where $\theta_2^{(-c)}$ and $\theta_2^{(+c)}$ are given in (1.31) and (1.34), respectively.

1.7 Conclusions

In this paper, we completely characterized the weak tail decay rates in terms of the transition probabilities for the stationary distribution of the two sided DQBD process (Theorems 1.3). For the exactly geometric decay, we find sufficient conditions, which are close to necessary conditions (Theorem 1.4). We then apply those results to the generalized join shortest queue with two waiting lines, whose decay rate problem has been only solved under some special conditions such as the weakly and strongly pooled conditions in the literature. We completely answer to this problem by finding the weak decay rates of the stationary distributions of the minimum of the two queues and their difference for all cases (Theorem 1.5 and Corollary 1.2). It is notable that the strongly and weakly pooled conditions still play the important role for finding the decay rate for the minimum of two queues. That is, the decay rate crucially changes according to whether or not those two conditions are satisfied.

Acknowledgments I am grateful to Yiqiang Zhao for his careful reading the original manuscript of this chapter and many invaluable comments. I also think Mr. Hiroyuki Yamakata for computing some numerical values. This research is supported in part by JSPS under grant No. 18510135.

Appendix 1

We prove Theorem 1.2. Let $\mathbf{x} = (\ldots, x_{-1}, x_0, x_1, \ldots)$ be the right positive invariant vector of $A_*^{(2)}(z)$. Then, we have

$$\begin{aligned}
x_n &= p_{1*}^{(-)}(z)x_{n-1} + p_{0*}^{(-)}(z)x_n + p_{(-1)*}^{(-)}(z)x_{n+1}, \qquad n \le -2,\\
x_{-1} &= p_{1*}^{(-)}(z)x_{-2} + p_{0*}^{(-)}(z)x_{-1} + p_{(-1)*}^{(2)}(z)x_0,\\
x_0 &= p_{1*}^{(-)}(z)x_{-1} + p_{0*}^{(2)}(z)x_0 + p_{(-1)*}^{(+)}(z)x_1,\\
x_1 &= p_{1*}^{(2)}(z)x_0 + p_{0*}^{(+)}(z)x_1 + p_{(-1)*}^{(+)}(z)x_2,\\
x_n &= p_{1*}^{(+)}(z)x_{n-1} + p_{0*}^{(+)}(z)x_n + p_{(-1)*}^{(+)}(z)x_{n+1}, \qquad n \ge 2.
\end{aligned} \tag{1.52}$$

For $s = \pm$, let $w_1^{(s)}$ and $w_2^{(s)}$ be the solutions of the following quadratic equation:

$$p_{(-1)*}^{(s)}(z)w^2 - (1 - p_{0*}^{(s)}(z))w + p_{1*}^{(s)}(z) = 0. \tag{1.53}$$

Then $\mathbf{x}$ must have the following forms:

$$x_n = \begin{cases} x_{-1}(w_1^{(-)})^{n+1} + (x_{-2} - x_{-1}(w_1^{(-)})^{-1}) \displaystyle\sum_{\ell=-n+2}^{0} (w_1^{(-)})^{-\ell}(w_2^{(-)})^{n+2+\ell}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad n \le -2 \\ x_1(w_1^{(+)})^{n-1} + (x_2 - x_1 w_1^{(+)}) \displaystyle\sum_{\ell=0}^{n-2} (w_1^{(+)})^{\ell}(w_2^{(+)})^{n-2-\ell}, \qquad n \ge 2. \end{cases} \tag{1.54}$$

By the irreducibility assumption in (i), $p_{1*}^{(s)}(z) > 0$ and $p_{(-1)*}^{(s)}(z) > 0$. Furthermore, the positivity of x_n, $w_1^{(s)}, w_2^{(s)}$ must be real numbers. Hence, from the fact

$$w_1^{(s)} w_2^{(s)} = \frac{p_{1*}^{(s)}(z)}{p_{(-1)*}^{(s)}(z)} > 0, \tag{1.55}$$

$w_1^{(s)}$ and $w_2^{(s)}$ must be positive. This implies that $\mathbf{x}$ is nonnegative if and only if

$$x_{-2} w_1^{(-)} \ge x_{-1}, \qquad x_2 \ge x_1 w_1^{(+)}. \tag{1.56}$$

From (1.52), we have

$$x_{-2} = \frac{1}{p_{1*}^{(-)}(z)} \left(x_{-1} - p_{0*}^{(-)}(z) x_{-1} - p_{(-1)*}^{(2)}(z) x_0 \right), \tag{1.57}$$

$$x_2 = \frac{1}{p_{(-1)*}^{(+)}(z)} \left(x_1 - p_{1*}^{(2)}(z) x_0 - p_{0*}^{(+)}(z) x_1 \right). \tag{1.58}$$

Substituting these x_{-2} and x_2 into (1.56) yields

$$\left((1 - p_{0*}^{(-)}(z)) w_1^{(-)} - p_{1*}^{(-)}(z) \right) x_{-1} - p_{(-1)*}^{(2)}(z) w_1^{(-)} x_0 \ge 0,$$
$$\left((1 - p_{0*}^{(+)}(z)) - p_{(-1)*}^{(+)}(z) w_1^{(+)} \right) x_1 - p_{1*}^{(2)}(z) x_0 \ge 0.$$

Since $w_1^{(s)}$ satisfies (1.53), we have

$$p_{(-1)*}^{(-)}(z) w_1^{(-)} x_{-1} - p_{(-1)*}^{(2)}(z) x_0 \ge 0,$$
$$p_{1*}^{(+)}(z) x_1 - p_{1*}^{(2)}(z) w_1^{(+)} x_0 \ge 0.$$

Using (1.55), this is equivalent to

$$p_{1*}^{(-)}(z) x_{-1} - p_{(-1)*}^{(2)}(z) w_2^{(-)} x_0 \ge 0, \tag{1.59}$$

$$p_{(-1)*}^{(+)}(z) x_1 - p_{1*}^{(2)}(z) (w_2^{(+)})^{-1} x_0 \ge 0. \tag{1.60}$$

Hence, letting

$$\eta_2 = \log z, \qquad \eta_1^{(s)} = -\log w_2^{(s)},$$

we have (1.14), (1.15) and (1.16).

We next show that these conditions are also sufficient. Suppose that there are $\eta_2 \geq 0$ and $\eta_1^{(s)}$ with $s = \pm 1$ satisfying (1.14), (1.15) and (1.16). Then, we can find $u^{(s)}$ with $s = \pm 1$ such that

$$u^{(-)} + \varphi_0^{(2)}(\eta_2) + u^{(+)} = 1, \quad u^{(-)} \geq \varphi_{-1}^{(2)}(\eta_2) e^{-\eta_1^{(-)}}, \quad u^{(+)} \geq \varphi_1^{(2)}(\eta_2) e^{\eta_1^{(+)}}.$$

Let $x_0 = 1$, and define x_{-1} and x_1 as $x_{-1} = \frac{u^{(-)}}{p_{1*}^{(-)}(z)}$, $x_1 = \frac{u^{(+)}}{p_{(-1)*}^{(+)}(z)}$. Hence, letting $z = e^{\eta_2}$ and $w_2^{(s)} = e^{-\eta_1^{(s)}}$ with $s = \pm 1$, we have (1.59) and (1.60). Then, defining x_{-2}, x_2 and x_n by (1.57), (1.58) and (1.54), respectively, we revive (1.52). Hence, we indeed find the positive left eigenvector $\mathbf{x}$ of $A_*^{(2)}(z)$. This proves the first part of the theorem. The remaining parts are obvious from (1.54) and Lemma 1.2. □

Appendix 2

We prove Lemma 1.8. Define the following functions on $\mathbb{R}_+^3$, where $\mathbb{R}_+ = (0, \infty)$,

$$\begin{aligned}
&f(z_1, z_2, \xi) = \xi, \\
&g_1(z_1, z_2, \xi) = (\lambda_1 \xi + \mu_2) z_1^2 + \mu_1 \xi + (\lambda_2 + \delta) \xi^2 - z_1 \xi, \\
&g_2(z_1, z_2, \xi) = (\lambda_2 \xi + \mu_1) z_2^2 + \mu_2 \xi + (\lambda_1 + \delta) \xi^2 - z_2 \xi, \\
&h(z_1, z_2, \xi) = \left(\lambda_1 + \frac{\delta}{2}\right) z_1 \xi + \mu_2 z_1 + \mu_1 z_2 + \left(\lambda_2 + \frac{\delta}{2}\right) z_2 \xi - \xi.
\end{aligned}$$

Obviously, all the functions are convex. Then, Lemma 1.8 is obtained by the following optimization problem. In particular, $\eta_2^{(c)}$ is obtained as the logarithm of the maximum value of f.

$$\begin{aligned}
&\text{miximinze } f(z_1, z_2, \xi), \\
&\text{subject to} \\
&g_1(z_1, z_2, \xi) = 0, \qquad g_2(z_1, z_2, \xi) = 0, \qquad h(z_1, z_2, \xi) \leq 0, \qquad (1.61) \\
&z_1 > 0, \qquad z_2 > 0, \qquad \xi \geq 1. \qquad (1.62)
\end{aligned}$$

This is a convex optimization problem, and (1.61) is satisfied with equality only if $(z_1, z_2, \xi) = (1, 1, 1)$ or $(\rho^{-1}, \rho^{-1}, \rho^{-2})$ (see Lemma 3.2 of [3]). By D, we denote the set of all feasible solutions satisfying the constraints (1.61) and (1.62). Clearly, D is closed and bounded in $\mathbb{R}_+^3$. For convenience, let

$$D_i = \{z_i; (z_1, z_2, \xi) \in D\}, \qquad i = 1, 2.$$

Since $\{(z_i,\xi) \in \mathbb{R}_+^2; g_i(z_1,z_2,\xi) \le 0\}$ is a convex set, $g_i(z_1,z_2,\xi) = 0$ have two solutions counting multiplicity for each ξ and each $i = 1,2$ if the solution exists. Hence, there exist at most four points $(z_1,z_2,\xi) \in D$ for each ξ.

We show that D is a connected curve with end points $(1,1,1)$ and $(\rho^{-1},\rho^{-1},\rho^{-2})$ if D has three points at least. Suppose that this is not true. Let $(z_1^\circ, z_2^\circ, \xi^\circ) \in D$ be the third point other than the above end points. Then, we must have $h(z_1^\circ, z_2^\circ, \xi^\circ) < 0$. This implies that the point (z_1°, z_2°) is in the interior of the set

$$\{(z_1,z_2) \in \mathbb{R}_+^2; h(z_1,z_2,\xi) \le 0\},$$

for $\xi = \xi^\circ$, which is a polyhedral for each ξ and its region is continuously increased as ξ is increased. Hence, there exists a connected curve which passes through $(z_1^\circ, z_2^\circ, \xi^\circ)$ as an inner point. This curve must have $(1,1,1)$ and $(\rho^{-1},\rho^{-1},\rho^{-2})$ as its end points since otherwise we arrive at the contradiction that there is a point other than those points such that $h = 0$ holds.

Let us consider the cases for (1.39) and (1.40) separately. Here, we do not consider the case for (1.41) since it is symmetric to the case for (1.40). Denote the solutions of $g_i(z_1,z_2,\xi) = 0$ for each ξ by $\underline{z}_i(\xi)$ and $\overline{z}_i(\xi)$, where $\underline{z}_i(\xi) \le \overline{z}_i(\xi)$. First, assume that (1.39) holds. Then $(\underline{z}_1(\rho^{-2}), \underline{z}_2(\rho^2), \rho^{-2}) = (\rho^1,\rho^{-1},\rho^{-2}) \in D$ and $\overline{z}_i(\rho^{-2}) \notin D_i$ for $i = 1,2$. Hence, f is maximized at $(\rho^{-1},\rho^{-1},\rho^{-2})$. We next assume (1.40). Then, we have $(\overline{z}_1(\rho^{-2}), \underline{z}_2(\rho^{-2}), \rho^{-2}) = (\rho^{-1},\rho^{-1},\rho^{-2}) \in D$ and $(\underline{z}_1(\rho^{-2}), \underline{z}_2(\rho^{-2}), \rho^{-2}) \in D$ since $\underline{z}_1(\rho^{-2}) \le \overline{z}_1(\rho^{-2})$. If $\underline{z}_1(\rho^{-2}) = \overline{z}_1(\rho^{-2})$, we can reduce the problem to the case for (1.39). Otherwise, D has three points at least, so it is a connected curve with end points $(1,1,1)$ and $(\rho^{-1},\rho^{-1},\rho^{-2})$ as shown above. This concludes that f is maximized at $(\underline{z}_1(\xi^*), \underline{z}_2(\xi^*), \xi^*)$ such that $\underline{z}_1(\xi^*) = \overline{z}_1(\xi^*)$. Since ξ^* must be the maximum value of ξ satisfying $g_1(z_1,z_2,\xi) = 0$, $\eta_2^{(c)} = \theta_2^{-\max}$. This completes the proof. □

It may be notable that we can also solve the optimization problem by applying Karush-Kuhn-Tucker necessary conditions (e.g., see Sect. 4.3.7 of [11]). However, the present solution is more informative since the feasible region D is identified to be a connected curve.

References

1. M. Miyazawa, "Tail decay rates in a doubly QBD process," Submitted for publication.
2. M. Miyazawa, "Doubly QBD process and a solution to the tail decay rate problem," in *Proc. the Second Asia-Pacific Symposium on Queueing Theory and Network Applications*, pp. 33-42, 2007.
3. H. Li, M. Miyazawa, and Y. Q. Zhao, "Geometric decay in a QBD process with countable background states with applications to shortest queues," *Stochastic Models*, vol. 23, pp. 413-438, 2007.
4. G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Philadelphia: American Statistical Association and the Society for Industrial and Applied Mathematics, 1999.

5. M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models*. Baltimore: Johns Hopkins University Press, 1981.
6. R. D. Foley and D. R. McDonald, "Join the shortest queue: stability and exact asymptotics," *The Annals of Applied Probability*, vol. 11, pp. 569-607, 2001.
7. A. A. Puhalskii and A. A. Vladimirov, "A large deviation principle for join the shortest queue," *Mathematics of Operations Research*, vol. 32, pp. 700-710, 2007.
8. M. Miyazawa and Y. Q. Zhao, "The stationary tail asymptotics in the GI/G/1 type queue with countably many background states," *Advances in Applied Probability*, vol. 36, pp. 1231-1251, 2004.
9. E. Seneta, *Nonnegative Matrices and Markov Chains*, Second Edition. New York: Springer-Verlag, 1981.
10. R. D. Foley and D. R. McDonald, "Bridges and networks: Exact asymptotics," *The Annals of Applied Probability*, vol. 15, pp. 542-586, 2005.
11. M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*. New York: Wiley, 1993.

Chapter 2
Analytical Model of On-Demand Streaming Services Based on Renewal Reward Theory

Hiroshi Toyoizumi

Abstract We propose an analytical model based on renewal reward theory to investigate the dynamics of an on-demand streaming service. At the same time, we also propose a simple method combining a method of multicasts and method of unicasts that can reduce the download rate from the streaming server without causing delay. By modeling the requests as a Poisson arrival and using renewal reward theory, we study the dynamics of this streaming service and derive the optimal combination of unicast and multicast methods. We even show how to estimate the fluctuation of download rates of a streaming service.

2.1 Introduction

Streaming services have become increasingly popular in recent years. However, establishing an efficient large-scale streaming service is still a great challenge because they demand an enormous amount of bandwidth for servers delivering contents. Thus, it is quite important to find an efficient and reliable way to establish a large-scale streaming service over the Internet. There is much research going on to find a better streaming service. For example, [1], [2] proposed a streaming service based on the sophisticated data fragment technique, whereas [3] discusses the possibility of popularity-based delivery and [4] seeks the dynamic structure of a contents delivery network, both aiming to reduce bandwidth. Because there is a wide variety of methods, it is also quite important to evaluate and compare the proposed methods and find the optimal strategy [5]. In most cases, the evaluation is based on the study of arbitrary selected simulations. Only [6], [7] discuss theoretical analysis of reduction of bandwidth of streaming service, but they only succeeded in giving theoretical bounds. In order to understand the dynamics of streaming service, we need an analytically tractable model.

H. Toyoizumi
Graduate School of Accoutancy and Department of Applied Mathematics, Waseda University, Tokyo 169-8050, Japan
e-mail: toyoizumi@waseda.jp

W. Yue et al. (eds.), *Advances in Queueing Theory and Network Applications,*

In this chapter, we propose a simple method combining unicasts and multicasts to reduce the download rate of streaming service. Assuming the arrival of requests is Poisson arrival, we use the technique called renewal reward theory to investigate the dynamics of streaming service. By this analysis, we show we can reduce the download rate by the order of $\sqrt{\rho}$, where ρ is the average download rate required if we use the standard unicast streaming service. Renewal reward theory is one of the fundamental and powerful tools to investigate stochastic processes (see [8], [9], for example). We can derive not only the average overall download rate but also its distribution. This method can be used to design the link speed of the streaming service.

Consider setting up a streaming service (Fig. 2.1). If we use a unicast from the streaming server on each request, users will not experience delay, because the unicast delivers the data on a one-to-one basis. However, using unicasts will result in the waste of bandwidth if users request the same content at the same time. Multicast streaming is realized by copying the data at multicast nodes in a content delivery network so as to reduce the bandwidth. Unlike unicast, multicast is one-to-many, and multicast can deliver the same data to all the users efficiently when sending the same content. However, there is a side effect in multicast streaming. Those who requested later than the start of multicast miss the initial part of the stream. Thus, we propose a simple method using both the unicast and multicast reducing download rate without causing delay. The objective of our method is to reduce the bandwidth required for the streaming server.

Assume there is only one content on the streaming server, for simplicity. We may extend our model to the heterogeneous contents environment, by modeling virtual streaming servers for each content, and treat them separately. A user (or a leaf node of a content delivery network) submits a request for the content to the designated streaming server. The server has two possible options:

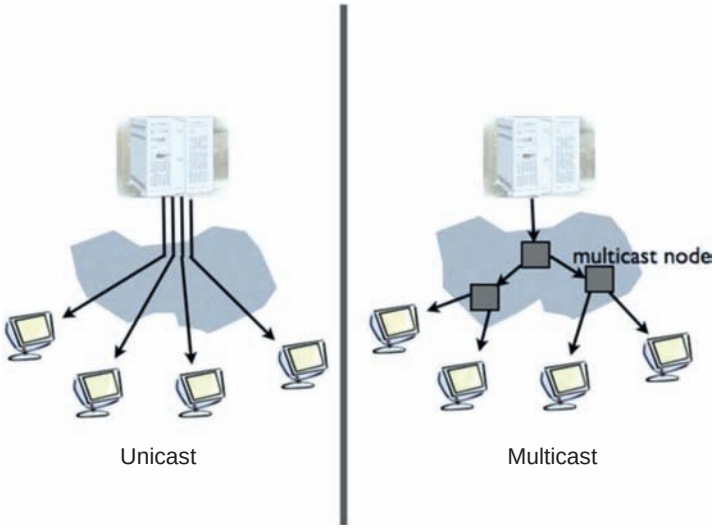


Fig. 2.1 Streaming service on network of unicast and multicast.

(1) Use a multicast so that other users can listen to this content simulataneously, or
(2) Use a unicast that can be listened to by this user exclusively.

Let us see some details on how our method will work. At the time when a user submits a request, if there is no multicast stream, the server has no choice but initiates a new multicast stream. If the server has already started a multicast stream, the server can use a unicast to reduce bandwidth. However, in some cases, the server may save some bandwidth by starting a new multicast even if there is another multicast stream. Figure 2.2 shows an example of how the requests may be handled. Each upward arrow indicates the arrival of requests at the streaming server. The request C1 arrives at the server when there is no stream. Thus, there is no choice, and the server automatically starts a multicast stream (real line). The next request C2, on the other hand, starts listening to the C1 multicast stream, as well as the unicast (dashed line) that corresponds to the top part to which she missed listening (Figure 2.3). The unicast C2 will be terminated when it catches up to the part that has been already stored by listening to the C1 multicast stream. In this way, the request C2 will not see the delay, while saving the bandwidth. At the request C3, the sever selects a new multicast even though there is a C1 multicast. This is because even if C3 started listening to the multicast C1, which has already been started quite some time ago,

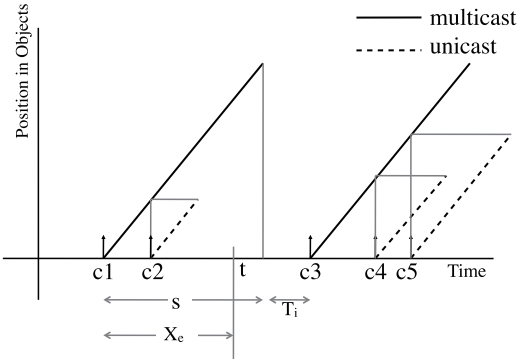


Fig. 2.2 Streaming with unicasts and multicasts. Arrows are the arrivals of request. The vertical line shows the amount the user has listened.

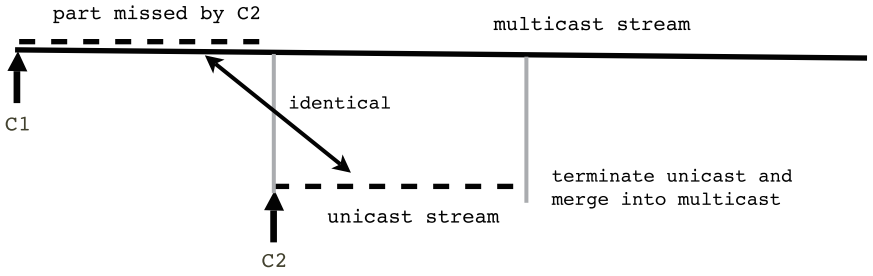


Fig. 2.3 Relationship of C1 multicast and C2 unicast.

C3 has to listen to almost the entire contents by her own unicast. So, there is almost no gain. Thus, instead of listening to the existing multicast, starting a new multicast will reduce the download rate required for future requests, such as C4 and C5. In this chapter, we assume we cannot use the information of future arrival times. So, we need to make a decision under uncertainty.

Note that our method may be applied to a content delivery network on a Peer-to-Peer (P2P) network [4], [10], as well as normal full-scale multicast platforms.

This chapter is organized as follows. In Sect. 2.2, we propose an analytical model to study the optimal strategy for this streaming service, using renewal reward theory. In Sect. 2.3, we present the mean download rate and the optimal strategy. In Sect. 2.4, we derive the download rate distribution. We give some conclusions and remarks in Sect. 2.5.

2.2 Streaming Services and Renewal Model

Assume the server has only one content of the length s, and its download rate of each stream is 1. The arrival of requests is assumed to be a Poisson process with the rate λ. Although this assumption is a mathematical convention, there is research that we can observe a Poisson arrival at the multimedia server in some cases [11].

Suppose that a request arrives at the server at time 0, and the server starts a multicast for this request. Let us assume that all requests arrived during $(0,x]$ are regarded as children of the parent multicast, and the server starts a unicast for each child request. Obviously, x should be no more than the contents length s. Those designated as child requests should listen to the parent multicast and her own unicast simultaneously. The first request arrived after x becomes a new parent and the server starts a new parent multicast. We call x the merging limit time. Our primary goal is to find the optimal x minimizing the total download rate required, using the renewal reward process argument.

Let $N(t)$ be the number of requests arrived during $(0,t]$, and T_n be the arrival time of the nth request ($T_0 = 0$). We evaluate R which is the volume downloaded from the server for the parent and his $N(x)$ child requests; that is,

$$R = \sum_{i=1}^{N(x)} T_i + s, \tag{2.1}$$

because the server has to send the part T_i, which the child request C_i missed listening to in the parent multicast. By conditioning on $N(x)$, we have the expectation of R as

$$\begin{aligned} \mathrm{E}[R] &= s + \mathrm{E}\left[\sum_{i=1}^{N(x)} T_i\right] \\ &= s + \mathrm{E}\left[\mathrm{E}\left[\sum_{i=1}^{N(x)} T_i \middle| N(x)\right]\right]. \end{aligned} \tag{2.2}$$

The arrival is a Poisson process, conditioning on $N(x) = n$, thus the sequence of arrivals $T_1, T_2, \ldots, T_n$ is known to be equivalent to the ordered statistics of $U_1, U_2, \ldots, U_n$ which is a series of independent and identical random variables uniformly distributed on $(0, x]$ (e.g., see [8, Theorem 2.3.1]). Thus,

$$\begin{aligned} \mathrm{E}\left[\sum_{i=1}^{N(x)} T_i \middle| N(x) = n\right] &= \mathrm{E}\left[\sum_{i=1}^{n} U_i\right] \\ &= \frac{nx}{2}. \end{aligned}$$

Using this in (2.2), we have

$$\mathrm{E}[R] = s + \mathrm{E}\left[\frac{N(x)x}{2}\right] = s + \frac{\lambda x^2}{2}. \tag{2.3}$$

Now, let X_m be the interarrival time of the mth parent multicast. Because the interarrival time of the Poisson process is exponentially distributed and memoryless, the time length to the next request after the merging time limit is again exponentially distributed with its mean $1/\lambda$. Hence, X_m are independent and have the form of

$$X_m = x + T_i, \tag{2.4}$$

where T_i is an exponential random variable with the mean $1/\lambda$. Also, let R_m be the volume downloaded by the mth parent multicast and its child unicasts. Because the arrival is a Poisson process, the sequence of the pair of random variables $(X_m, R_m)_{m=1,2,\ldots}$ is independent and identically distributed. Let $S(t)$ be the total accumulated volume demanded by requests whose parent arrived before the time t; in other words,

$$S(t) = \sum_{m=1}^{M(t)} R_m, \tag{2.5}$$

where $M(t)$ is the number of parent multicasts in $[0, t)$. Taking R_m as the reward, the process $S(t)$ is a renewal reward process (see e.g., [8], [9]). This renewal reward representation is used in the following section to derive the average download rate.

2.3 Mean Download Rate and Optimal Strategy

We now find the optimal merging limit time x_0 that minimizes the average download rate from the streaming server. Let $b(x)$ be the average download rate given the merging limit time x, or

$$b(x) = \lim_{t \to \infty} \frac{S(t)}{t}. \tag{2.6}$$

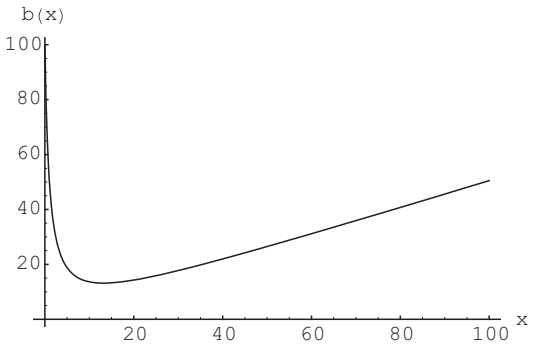


Fig. 2.4 The download rate $b(x)$ and the merging limit time x: the request arrival rate $\lambda = 1$, and the content size $s = 100$.

Theorem 2.1 (Optimal Merging Limit Time). *Assume the requests to the content of size s is a Poisson process with the rate λ. Given the merging limit time x, the average download rate is obtained by*

$$b(x) = \frac{2\lambda s + \lambda^2 x^2}{2(\lambda x + 1)}. \tag{2.7}$$

The function $b(x)$ is indeed a convex function (see Fig. 2.4), so we have x_0 which minimizes $b(x)$ as

$$x_0 = \frac{(1+2\lambda s)^{1/2} - 1}{\lambda}. \tag{2.8}$$

Furthermore, we can substitute (2.8) into (2.7); then we have the optimal download rate,

$$b(x_0) = (1+2\rho)^{1/2} - 1, \tag{2.9}$$

where $\rho = \lambda s$ corresponds to the scale of this streaming service.

Proof. We know that $S(t)$ is a renewal reward process from Sect. 2.2. Renewal reward theory [8, Theorem 3.6.1] is an extension of the strong law of large numbers to the renewal process. By the strong law of large numbers and (2.5), with probability 1, we have

$$\frac{S(t)}{t} = \frac{\sum_{m=1}^{M(t)} R_m}{M(t)} \frac{M(t)}{t} \to \frac{\mathrm{E}[R]}{\mathrm{E}[X]} = \frac{\mathrm{E}[R]}{x + 1/\lambda} \quad \text{as } t \to \infty, \tag{2.10}$$

where X is the interarrival time of the parent multicasts. It is easy to get (2.7) by substituting (2.3) in (2.10).

Figure 2.4 shows the graph of the average download rate $b(x)$. For a smaller merging limit time x, more requests are treated as multicast, which results in a waste of download rate. On the contrary, for a larger x, we may miss the opportunity of saving download rate by merging the future streams. Thus, we can see a fine balance here. In this case the optimal merging limit time is $x_0 = 9.04988$, well below the content size $s = 100$.

Let us study in some detail the optimal merging limit. Take the optimal merging limit time as a function of the request arrival rate λ in (2.8). Letting $\lambda \to 0$, we have

$$x_0(\lambda) \to s,$$

which means for a smaller request rate we cannot count on the following requests, so "be a child whenever you can" is the best strategy. On the contrary, for large λ, we have

$$x_0(\lambda) \to 0, \quad \text{as } \lambda \to \infty.$$

For a larger request rate, you can always expect the following requests. In this case your strategy would be "be a parent and help the following children."

If we use unicast only, instead of the combination of unicast and multicast, the average streaming rate is ρ. The download rate (2.9) obtained by our method has the order of $\sqrt{\rho}$, which gives us a significant saving of download rate, especially when the size of the streaming service is large (see Fig. 2.5). Theoretically, we could improve (2.9) when we exploit the information of future requests. The theoretical lower bound of the download rate given future information was obtained by [6] as

$$b_0 = \log(1+\rho), \tag{2.11}$$

which is also shown in Fig. 2.5. We see that our method cannot achieve this theoretical limit but still it achieves significant saving.

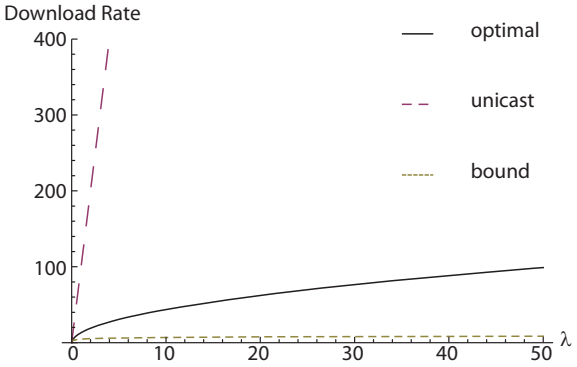


Fig. 2.5 Comparison of download rate: The line unicast is the scheme that uses only unicasts, and the bound is the theoretical lower bound [6]. The contents size is set to be $s = 100$.

2.4 Download Rate Distribution

Using the renewal argument further, we can evaluate the download rate distribution of our merging method. For simplicity, in this section we set the merging limit time x to be the size of the content s. In this case, all requests are treated as child streams whenever they can.

Because we set the download rate of each stream to be 1, the only thing we need to know is the number of active streams. Let L be the number of active streams including both parent and child streams in the steady state.

Theorem 2.2. *The z-transform of the number of active streams L is obtained as*

$$\mathrm{E}\left[z^L\right] = \frac{1}{1+\rho}\left[ze^{(\rho-1)(z-1)/2}\int_{e^{-\rho}}^{1} e^{(z-1)y/2}\frac{dy}{y} + \frac{2}{z+1}\left\{e^{-\rho(1-z)/2} - e^{-\rho}\right\} + e^{-\rho}\right], \tag{2.12}$$

where we set $\rho = \lambda s$.

Proof. Let $L(t)$ be the number of active streams at the time t, and let Y_e be the length to the arrival of the previous parent request from an arbitrary time t (Fig. 2.6). Because Y_e is the forward recurrent time of the renewal interval $s+T_i$, where T_i is exponentially distributed with the mean $1/\lambda$, we have

$$P\{Y_e \le u\} = \frac{1}{s+1/\lambda}\int_0^u (1 - P\{s+T \le y\})dy. \tag{2.13}$$

Thus, we obtain the probability distribution of Y_e as

$$(1+\rho)P\{Y_e \le u\} = \begin{cases} \lambda u & \text{if } u \le s \\ 1+\rho-e^{-\lambda(u-s)} & \text{if } u > s, \end{cases} \tag{2.14}$$

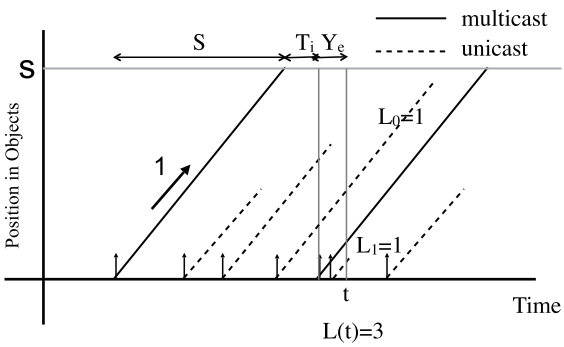


Fig. 2.6 Sample path of streaming service: The fourth child unicast from the previous renewal interval remains active at the time t.

and its density as

$$(1+\rho)\frac{dP\{Y_e \le u\}}{du} = \begin{cases} \lambda & \text{if } u \le s \\ \lambda e^{-\lambda(u-s)} & \text{if } u > s. \end{cases} \tag{2.15}$$

Let L_1 be the number of active child streams at time t that arrived in the renewal interval containing the time t. Furthermore, there is a chance that the child streams started prior to the renewal time still exist after time t (see Fig. 2.6). Let L_0 be the number of those active streams that arrived in the previous renewal interval.

Then, taking into account the parent multicast of this renewal interval, we have

$$L(t) = 1_{(Y_e \le s)} + L_0 + L_1. \tag{2.16}$$

Consider conditioning on $Y_e = u$. In the case when $u \le s$, L_0 and L_1 are independent, and both are Poisson random variables with the mean $\lambda s P_0(u)$ and $\lambda u/2$, respectively, where $P_0(u)$ is the probability that a child stream started in the previous interval still exists at time t. Indeed, the arrival of child streams is Poisson with rate λ, and given the number of arrivals, the survival of each child stream is independent of other streams.

Suppose a child stream arrives U_1 later than the parent multicast that started the current renewal interval. The child stream should exist to cover the missing part of the length U_1, and it is alive up to $2U_1$ from the start of the parent multicast. The child stream exists at time t only when $2U_1 > u$. Because U is uniformly distributed on $[0,u]$, the probability that a child stream exists at time t is $P\{U_1 \ge u/2\} = 1/2$. Thus, L_1, the number of active child streams at time t that arrived in $[t-u,t]$, is a Poisson random variable with the mean $\lambda u/2$. Similarly, suppose a child stream in the previous renewal interval arrives U_0 after the previous parent multicast. Then, the child stream remains active at time t only when $2U_0 > s+T+u$. Thus,

$$\begin{aligned} P_0(u) &= P\{2U_0 > s+T+u\} \\ &= \{\lambda(s-u) - (1-e^{-\lambda(s-u)})\}/(2\rho), \end{aligned} \tag{2.17}$$

because U_0 is a uniform random variable on $[0,s]$. Thus L_0 is a Poisson random variable with the mean $\lambda s P_0(u)$. Using this information we have

$$\begin{aligned} \int_0^s \mathrm{E}\left[z^{L(t)} \middle| Y_e = u \right] dP\{Y_e \le u\} &= \int_0^s z e^{\lambda s P_0(u)(z-1)} e^{\lambda u(z-1)/2} dP\{Y_e \le u\} \\ &= \frac{\lambda z e^{(\rho-1)(z-1)/2}}{1+\rho} \int_0^s e^{(z-1)e^{-\lambda(s-u)}/2} du \\ &= \frac{z e^{(\rho-1)(z-1)/2}}{1+\rho} \int_{e^{-\rho}}^1 e^{(z-1)y/2} \frac{dy}{y}. \end{aligned} \tag{2.18}$$

On the other hand, when $s < u \le 2s$, it is easy to see that no child streams from the previous interval exist at time t. Thus, $L_0 = 0$ and L_1 is a Poisson random variable with its mean $\lambda(s-u/2)$. Hence we have

$$\int_s^{2s} \mathrm{E}\left[z^{L(t)} \middle| Y_e = u\right] dP\{Y_e \le u\} = \int_s^{2s} e^{\lambda(s-u/2)(z-1)} dP\{Y_e \le u\}$$
$$= \frac{2}{(1+\rho)(z+1)} \left\{ e^{-\rho(1-z)/2} - e^{-\rho} \right\}. \quad (2.19)$$

Lastly, when $u > 2s$, $L(t) = 0$. Hence, we have

$$\int_{2s}^{\infty} \mathrm{E}\left[z^{L(t)} \middle| Y_e = u\right] dP\{Y_e \le u\} = \int_{2s}^{\infty} dP\{Y_e \le u\}$$
$$= \frac{1}{(1+\rho)} e^{-\rho}. \quad (2.20)$$

By using all these results and by separating integral intervals appropriately, we can get (2.12).

Corollary 2.1. *The mean and variance of L are given by*

$$\mathrm{E}[L] = \frac{2\rho + \rho^2}{2(1+\rho)} < \rho, \quad (2.21)$$
$$V[L] = \{4\rho^3 - 4\rho^2 + 11\rho + 9 - 4(\rho^2 + 3\rho + 2)e^{-\rho}$$
$$- (\rho + 1)e^{-2\rho}\} / \{8(1+\rho)^2\}. \quad (2.22)$$

Here we give a numerical example. If we use only unicasts for requests, L is nothing but a simple M/D/∞ queueing system. Thus, L is a Poisson random variable with its mean $\rho = \lambda s$. In Fig. 2.7, we compare the variance of L of the proposed merging method with the M/D/∞ queue. We already know that we can save the average download rate using our proposed method. In Fig. 2.7, we also see the reduction of the download rate fluctuation, which is another superiority of our method.

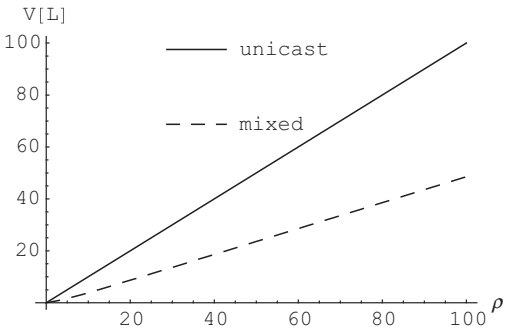


Fig. 2.7 Variance of L in the unicast scheme and the proposed method (mixed).

2.5 Conclusions

In this chapter, we proposed a simple method realizing bandwidth reduction without delay. By using renewal reward theory, we succeed in estimating the download rate, not only the average but also the variance. By using the evaluation, we find that in an optimal case we can reduce the download rate of streaming service by $\sqrt{\rho}$, the squareroot of the streaming service size. Furthermore, we see that our proposed method can also reduce the fluctuation of the download rate. The technique used in this chapter can be adopted to design the bandwidth requirement for general streaming services.

References

1. K. A. Hua and S. Sheu, Skyscraper broadcasting: A new broadcasting scheme for metropolitan video-on-demand systems, in *Proc. SIGCOMM*, pp. 89–100, 1997. [Online]. Available: citeseer.nj.nec.com/hua97skyscraper.html.
2. J. W. Byers, M. Luby, M. Mitzenmacher, and A. Rege, A digital fountain approach to reliable distribution of bulk data, in *Proc. SIGCOMM*, pp. 56–67, 1998. [Online]. Available: citeseer.nj.nec.com/byers98digital.html.
3. K. Thirumalai, J. F. Paris, and D. D. E. Long, Tabbycat: an inexpensive scalable server for video-on-demand, in *Proc. IEEE International Conference on Communications, ICC'03*, vol. 2, pp. 896–900, 2003.
4. M. Tran and W. Tavanapong, Peers-assisted dynamic content distribution networks, in *Proc. The IEEE Conference on Local Computer Networks*, pp. 123–131, 2005.
5. A. Mahanti, D. Eager, M. Vernon, and D. Sundaram-Stukel, Scalable on-demand media streaming with packet loss recovery, in *Proc. SIGCOMM'2001*, p. 12, 2001. [Online]. Available: citeseer.nj.nec.com/mahanti01scalable.html.
6. D. L. Eager, M. K. Vernon, and J. Zahorjan, Minimizing bandwidth requirements for on-demand data delivery, *Knowledge and Data Engineering*, vol. 13, no. 5, pp. 742–757, 2001. [Online]. Available: http://citeseer.nj.nec.com/eager99minimizing.html.
7. D. L. Eager, M. K. Vernon, and J. Zahorjan, Optimal and efficient merging schedules for video-on-demand servers, in *Proc. ACM Multimedia (1)*, pp. 199–202, 1999. [Online]. Available: citeseer.nj.nec.com/eager99optimal.html.
8. S. M. Ross, *Stochastic Processes*. New York: John Wiley and Sons, 1996.
9. R. Wolff, *Stochastic Modeling and the Theory of Queues*. New Jersey: Prentice Hall, 1989.
10. Y. Guo, K. Suh, J. Kurose, and D. Towsley, A peer-to-peer on-demand streaming service and its performance evaluation, in *Proc. International Conference on Multimedia and Expo*, vol. 2, pp. II-649–52, 2003.
11. J. M. Almeida, J. Krueger, D. L. Eager, and M. K. Vernon, Analysis of educational media server workloads, in *Proc. 11th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV '01)*, pp. 21–30, 2001.

Chapter 3
A Pure Decrement Service Geom/G/1 Queue with Multiple Adaptive Vacations

Zhanyou Ma, Wuyi Yue, and Naishuo Tian

Abstract In this chapter, a Geom/G/1 queue model with a pure decrement service policy and multiple adaptive vacations is analyzed. The Probability Generation Function (P.G.F.) of the queue length is obtained by using an embedded Markov chain method. The P.G.F. of the waiting time is then derived based on the independence between the arrival process and the waiting time. The probabilities for the system being in various states of busy, vacation, or idle, respectively, are also derived. Finally, some special cases for the Geom/G/1 queue model with a pure decrement service policy and multiple adaptive vacations are given to demonstrate the general properties of the queue models.

3.1 Introduction

Tian [1] introduced a multiple adaptive vacation policy, and studied a multiple adaptive vacation M/G/1 queue model with an exhaustive service rule, and through this, queue models with multiple vacations and single vacation were extended. Zhang and Tian studied the discrete time queue model with multiple adaptive vacations, and obtained the P.G.F. of the queue length and waiting time in [2].

However, they only researched the queue models with the exhaustive service polity. Many researchers have studied discrete time queue models with some

Z. Ma
College of Science, Yanshan University, Qinhuangdao 066004, China
e-mail: mzhy55@ysu.edu.cn

W. Yue
Department of Intelligence and Informatics, Konan University, Kobe 658-8501, Japan
e-mail: yue@konan-u.ac.jp

N. Tian
College of Science, Yanshan University, Qinhuangdao 066004, China
e-mail: tiannsh@ysu.edu.cn

W. Yue et al. (eds.), *Advances in Queueing Theory and Network Applications,*

vacation policy. For example, M/G/1 queues with multiple types of feedback and gated vacations were studied, and some important results were derived in [3]. A number of discrete time queue models were studied in [4] and [5]. Also, Wu and Takagi studied the queue model with working vacations in [6], and extended the general vacation polity.

The new queue model enriched the theory of the queue with vacations, and involved many queue models as special cases. A discrete time GI/Geo/1 queue model with multiple vacations was studied in [7]. A discrete time queue model with timed vacations was analyzed in [8]–[10].

Bischof studied the queue model with vacations under six different service disciplines in [11], which expanded the research of the nonexhaustive service disciplines. Performance evaluations of SVC-Based IP-Over-ATM networks were given using discrete time queueing theory in [12]. However, these papers did not integrate multiple adaptive vacations with nonexhaustive service disciplines. The authors' purpose for studying a new queueing model was to promote this integration.

In this chapter, we analyze a general Geom/G/1 queueing model with a pure decrement service strategy and multiple adaptive vacations. We show that the pure decrement service systems analyzed in [5] and [13] are special cases of our model presented in this chapter. Furthermore, we compare the system performance for pure decrement service strategies with multiple vacations and single vacation.

The chapter is organized as follows. Section 3.2 describes the analysis model in detail. Section 3.3 presents analysis of system performance. Some special cases are presented in Sect. 3.4. In Sect. 3.5, we discuss some numerical results. Concluding remarks are given in Sect. 3.6.

3.2 Model Description

Based on the classical Geom/G/1 queueing model, we introduce the strategies of a pure decrement service and the multiple adaptive vacations [5], [13].

A pure decrement service strategy can be described as follows. Once the service period starts, the server will keep on working until the number of customers in the system is one less than the number of customers at the start instant of the service period. The server will then enter a new vacation period. If there are some customers waiting at a vacation completion instant, the server will complete the vacation period and start a new service period. Otherwise, the server will take some vacations consecutively according to the assistant workload completed at that time.

The maximum number of vacations during a vacation period is denoted by H. H is a positive integer random variable with the probability distribution h_j and the P.G.F. $H(z)$ as follows:

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

Let V_k $(k = 1, 2, \ldots, H)$ be the time length for the kth vacation. V_k is an independently identically distributed (i.i.d.) random variable. If there is no customer in the system at the Hth vacation completion instant, the system will enter an idle period and wait for a new customer to arrive. If a customer arrives during the idle period, the server will enter a service period immediately, and continue until there are waiting customers in the system, before taking a vacation again at the completion instant of the service. The system will continually repeat the above processes.

Specifically, (1) if $H \to \infty$, the model corresponds to a pure decrement service Geom/G/1 queueing model with multiple vacations [5], [10], [13]. (2) If $H = 1$, the model corresponds to a pure decrement service Geom/G/1 queueing model with a single vacation [5], [13]. (3) If H follows another distribution, the model corresponds to another special queueing model.

The basic assumptions of the new model presented in this chapter are given as follows.

(1) In order to describe the system states in the nth discrete time instants, we assume that customer arrivals can only occur at discrete time instants $t = n^-$, $n = 0, 1, \ldots$, the service starts and ends can only occur at discrete time instants $t = n^+, n = 1, 2, \ldots$. The model is called a late arrival system. The interarrival time, denoted by T, is supposed to be an i.i.d. discrete random variable following a geometric distribution with parameter p $(0 < p < 1)$. We can write the probability distribution of T as follows:

$$P(T = j) = p\bar{p}^{j-1}, \qquad j = 1, 2, \ldots,$$

where $\bar{p} = 1 - p$. We denote by C_n the number of customers arriving during the interval $[0, n]$; then C_n follows a binomial distribution,

$$P(C_n = j) = \binom{n}{j} p^j \bar{p}^{n-j}, \qquad j = 0, 1, \ldots, n.$$

(2) The service time S of a customer is supposed to be an i.i.d. discrete random variable with a general distribution; the probability distribution s_j and the P.G.F. $S(z)$ of S are given as follows:

$$P(S_i = j) = s_j, \quad j \geq 1, \quad S(z) = \sum_{j=1}^{\infty} s_j z^j.$$

Let $\mathrm{E}[S]$ and $\mathrm{E}[S(S-1)]$ be the mean and the second factorial moment of S; then we have

$$\frac{1}{\mu} = \mathrm{E}[S] = \sum_{i=0}^{\infty} i s_i, \quad \mathrm{E}[S(S-1)] = \left.\frac{d^2 S(z)}{dz^2}\right|_{z=1}.$$

(3) The time length V of a vacation is a nonnegative i.i.d. discrete random variable with general probability distribution v_j and the P.G.F. $V(z)$ given by

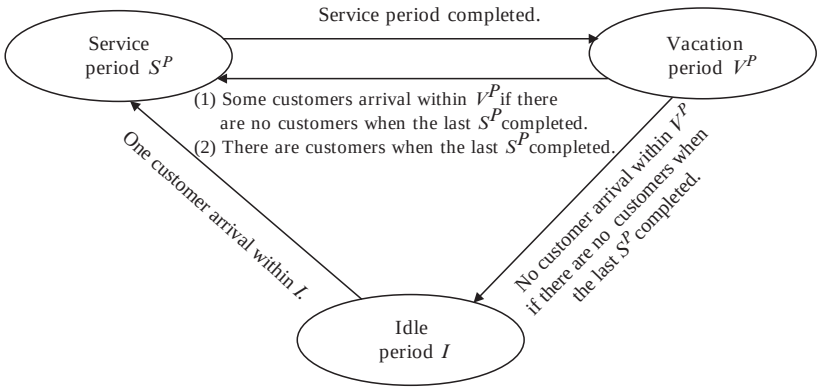


Fig. 3.1 State transition diagram of the model.

$$P(V = j) = v_j, \quad j \geq 1, \quad V(z) = \sum_{j=1}^{\infty} v_j z^j,$$

where the mean $\mathrm{E}[V]$ and the second factorial moment $\mathrm{E}[V(V-1)]$ of V exist.

Suppose that there is a single server in this system, and its buffer capacity is infinite. The interarrival time, the service time, and the time length of a vacation are mutually independent. The service order is First-Come First-Served (FCFS). The model is denoted by Geom/G/1 (PD, MAVs), where PD and MAVs represent the Pure Decrement and the Multiple Adaptive Vacations, respectively. Let S^P represent the service period, V^P represent the vacation period, and I represent the idle period, respectively. The state transition diagram of the model is shown in Fig. 3.1.

Let L_v represent the stationary queue length at the departure instant of a customer, and let $Q_b^{(n)}$ represent the number of customers in the system at the nth vacation completion instant, the P.G.F. of $Q_b^{(n)}$ is denoted by $Q_b^{(n)}(z)$. L_v is supposed to follow an identical distribution for the new model in the service orders of FCFS or Last-Come First-Served (LCFS). For simplification, we assume that the service of the model presented in this chapter follows a LCFS strategy.

3.3 Analysis of System Performance Measures

3.3.1 Number of Customers at the Beginning of a Service Period

Let J be the number of consecutive vacations taken by the server after the end of a service period when the system is empty. J is a random variable, and we have

$$J = \min\{H, k : V_1 + \cdots + V_{k-1} < T < V_1 + \cdots + V_k\}.$$

We define two events as follows:

$$A_I = \{\text{ Service period starts with the end of an idle period if there are no customers at the end of the last } S^P\};$$
$$A_v = \{\text{ Service period starts with the end of a vacation period if there are no customers at the end of the last } S^P\}.$$

Then, we have $P(A_I)$, the probability of A_I and $P(A_v)$, the probability of A_v as

$$P(A_I) = Q_b^{(n)}(0)H(V(\bar{p})), \qquad P(A_v) = Q_b^{(n)}(0)(1-H(V(\bar{p}))).$$

According to a pure decrement service order, if a service period with zero duration is allowed, $Q_b^{(n)}$ is the number of customers in the system at the next start instant of the service period. If $Q_b^{(n)}$ is greater than zero, the service period starts immediately and keeps on working until the number of customers in the system is one less than the number of customers at the start instant of the service period. Then the server will take a vacation. $Q_b^{(n+1)}$ is equal to the sum of $Q_b^{(n)}-1$ plus the number of customers arriving during the vacation. If $Q_b^{(n)}=0$, there are two cases as follows:

(1) If there are customer arrivals during the kth $(1 \le k \le H)$ vacation, a service period starts at the instant where the kth $(1 \le k \le H)$ vacation completes. The number of customers in the system at the start instant of the service period $Q_b^{(n+1)}$ is equal to the number of customers arriving during the vacation.
(2) If no customers arrive during the Hth vacation, an idle period will begin at the end of the vacation and continue until a new customer arrives. In this case, $Q_b^{(n+1)}$ is equal to 1. Therefore,

$$Q_b^{(n+1)}(z) = Q_b^{(n)}(0)H(V(\bar{p}))z + \frac{Q_b^{(n)}(z)-Q_b^{(n)}(0)}{z}V(1-p(1-z)) + Q_b^{(n)}(0)(1-H(V(\bar{p})))V(1-p(1-z)). \tag{3.1}$$

If the system is in a steady state, the P.G.F. $Q_b(z)$ of $Q_b^{(n+1)}$ does not depend on n in (3.1). If we let $\lim_{n\to\infty} Q_b^{(n+1)}(z) = Q_b(z)$, we can obtain $Q_b(z)$ as follows:

$$Q_b(z) = \frac{Q_b(0)\Big((1-z(1-H(V(\bar{p}))))\times V(1-p(1-z)) - H(V(\bar{p}))z^2\Big)}{V(1-p(1-z))-z}. \tag{3.2}$$

Because $Q_b(1)=1$, we have that

$$Q_b(0) = \frac{1-p\mathrm{E}[V]}{1+H(V(\bar{p}))(1-p\mathrm{E}[V])}. \tag{3.3}$$

According to the Foster rule (see Tian and Zhang [13]), we can prove that if $\rho = p/\mu < 1$ and $p\mathrm{E}[V] < 1$, the system can reach a steady state.

3.3.2 Stationary Queue Length and Waiting Time

Theorem 3.1. *If $\rho = p/\mu < 1$ and $pE[V] < 1$, the stationary queue length L_v in the Geom/G/1 (PD, MAVs) queue can be decomposed into three independent random variables:*

$$L_v = L + L_d + L_r,$$

where L is the stationary queue length in a classical Geom/G/1 queue [5], [13]. The P.G.F. $L(z)$ of L is

$$L(z) = \frac{(1-\rho)(1-z)S(1-p(1-z))}{S(1-p(1-z))-z}. \tag{3.4}$$

The additional queue length L_d is the number of customers arriving during a vacation or is equal to zero, and the additional queue length L_r is the number of customers in the system at the start instant of a vacation. P.G.F.s $L_d(z)$ and $L_r(z)$ of additional queue lengths L_d and L_r are given by

$$\begin{aligned} L_d(z) &= \frac{1-V(1-p(1-z))+H(V(\bar{p}))V(1-p(1-z))-H(V(\bar{p}))z}{(H(V(\bar{p}))+p\mathrm{E}[V](1-H(V(\bar{p}))))(1-z)}, \\ L_r(z) &= \frac{(1-p\mathrm{E}[V])(1-z)}{V(1-p(1-z))-z}. \end{aligned} \tag{3.5}$$

Proof. Q_b is the number of customers in the system at the start instant of a service. In the pure decrement service rule and LCFS order, a nonzero service period in the system is exactly the same as a standard busy period in a Geom/G/1 queue. So there are two kinds of customers in the system at a departure instant as follows:

(1) If $Q_b > 0$, only the customer who initiates the new service period can be served, and the residual $Q_b - 1$ customers wait to be served during the next service period. The P.G.F. of the number of these customers is given by

$$\frac{Q_b(z) - Q_b(0)}{(1-Q_b(0))z}. \tag{3.6}$$

(2) The number of customers (sub generation) who arrive during the service period and cannot be served is equivalent to the number of customers in a classical Geom/G/1 queue; the P.G.F. is given by

$$\frac{(1-\rho)(1-z)S(1-p(1-z))}{S(1-p(1-z))-z}. \tag{3.7}$$

Because the two kinds of customers are mutually independent, we have that

$$L_v(z) = \frac{(1-\rho)(1-z)S(1-p(1-z))}{S(1-p(1-z))-z} \times \frac{Q_b(z)-Q_b(0)}{(1-Q_b(0))z}. \tag{3.8}$$

Substituting (3.2) and (3.3) into (3.8), we have that

$$\begin{aligned} L_v(z) &= \frac{(1-\rho)(1-z)S(1-p(1-z))}{S(1-p(1-z))-z} \\ &\quad \times \frac{1-V(1-p(1-z))+H(V(\bar{p}))V(1-p(1-z))-H(V(\bar{p}))z}{(H(V(\bar{p}))+p\mathrm{E}[V](1-H(V(\bar{p}))))(1-z)} \\ &\quad \times \frac{(1-p\mathrm{E}[V])(1-z)}{V(1-p(1-z))-z} \\ &= L(z)L_d(z)L_r(z). \end{aligned} \tag{3.9}$$

Therefore, $L_v(z)$ is also the P.G.F. of the system queue length in the FCFS service strategy. □

Simplifying $L_d(z)$ in Theorem 3.1, we have that

$$\begin{aligned} L_d(z) &= \frac{H(V(\bar{p}))}{H(V(\bar{p}))+p\mathrm{E}[V](1-H(V(\bar{p})))} \\ &\quad + \frac{p\mathrm{E}[V](1-H(V(\bar{p})))}{H(V(\bar{p}))+p\mathrm{E}[V](1-H(V(\bar{p})))} \times \frac{1-V(1-p(1-z))}{p\mathrm{E}[V](1-z)}. \end{aligned} \tag{3.10}$$

Therefore, the additional queue length L_d is equal to zero with the following probability,

$$\frac{H(V(\bar{p}))}{H(V(\bar{p}))+p\mathrm{E}[V](1-H(V(\bar{p})))}$$

and is equal to the number of customers arriving before an arbitrary time instant during a vacation with the following probability,

$$\frac{p\mathrm{E}[V](1-H(V(\bar{p})))}{H(V(\bar{p}))+p\mathrm{E}[V](1-H(V(\bar{p})))}.$$

Differentiating the two sides of (3.9) and using L'Hospital's rule, we can obtain the mean $\mathrm{E}[L_v]$ of the number of customers at steady state for a Geom/G/1 (PD, MAVs) queue system as follows:

$$\begin{aligned} \mathrm{E}[L_v] &= \rho + \frac{p^2\mathrm{E}[S(S-1)]}{2(1-\rho)} + \frac{p^2\mathrm{E}[V(V-1)](1-H(V(\bar{p})))}{2(H(V(\bar{p}))+p\mathrm{E}[V](1-H(V(\bar{p}))))} \\ &\quad + \frac{p^2\mathrm{E}[V(V-1)]}{2(1-p\mathrm{E}[V])}. \end{aligned} \tag{3.11}$$

Theorem 3.2. *If $\rho = p/\mu < 1$ and $pE[V] < 1$, the stationary waiting time W_v of a customer can be decomposed into three independent random variables in a Geom/G/1 (PD, MAVs) queue as follows:*

$$W_v = W + W_d + W_r,$$

where W is the stationary waiting time in a classical Geom/G/1 queue [5], [13]. The P.G.F. $W(z)$ of the stationary waiting time W is given by

$$W(z) = \frac{(1-\rho)(1-z)}{(1-z)-p(1-S(z))}. \tag{3.12}$$

The additional delay W_d is a vacation or is equal to zero, and the additional delay W_r is the time delay caused by the existing customers at the start instant of a vacation. P.G.F.s $W_d(z)$ and $W_r(z)$ of additional delays W_d and W_r are given by

$$\begin{aligned} W_d(z) &= \frac{p(1-H(V(\bar{p})))(1-V(z))+H(V(\bar{p}))(1-z)}{(H(V(\bar{p}))+p\mathrm{E}[V](1-H(V(\bar{p}))))(1-z)}, \\ W_r(z) &= \frac{(1-p\mathrm{E}[V])(1-z)}{(1-z)-p(1-V(z))}. \end{aligned} \tag{3.13}$$

Proof. In a Geom/G/1 (PD, MAVs) queue, the waiting time is independent of the customers' inputting process after the arrival instant of the customers in the FCFS service strategy. The queue length of the system at a customer's service completion instant is composed of the number of other customers arriving during the waiting time and the service time of the customer. Therefore, we have that

$$L_v(z) = W_v(1-p(1-z))S(1-p(1-z)). \tag{3.14}$$

Substituting the result of Theorem 3.1 into (3.14), we have that

$$\begin{aligned} W_v(z) &= \frac{(1-\rho)(1-z)}{(1-z)-p(1-S(z))} \\ &\quad \times \frac{p(1-H(V(\bar{p})))(1-V(z))+H(V(\bar{p}))(1-z)}{(H(V(\bar{p}))+p\mathrm{E}[V](1-H(V(\bar{p}))))(1-z)} \\ &\quad \times \frac{(1-p\mathrm{E}[V])(1-z)}{(1-z)-p(1-V(z))} \\ &= W(z)W_d(z)W_r(z). \end{aligned} \tag{3.15}$$

□

From Theorem 3.2, we can obtain the P.G.F. $W_d(z)$ of the stationary waiting time W_d as follows:

$$\begin{aligned} W_d(z) &= \frac{H(V(\bar{p}))}{H(V(\bar{p}))+p\mathrm{E}[V](1-H(V(\bar{p})))} \\ &\quad + \frac{p\mathrm{E}[V](1-H(V(\bar{p})))}{H(V(\bar{p}))+p\mathrm{E}[V](1-H(V(\bar{p})))} \times \frac{1-V(z)}{\mathrm{E}[V](1-z)}. \end{aligned} \tag{3.16}$$

Therefore, the additional delay W_d is equal to zero with the following probability,

$$\frac{H(V(\bar{p}))}{H(V(\bar{p}))+p\mathrm{E}[V](1-H(V(\bar{p})))}$$

and is equal to a vacation time with the following probability,

$$\frac{p\mathrm{E}[V](1-H(V(\bar{p})))}{H(V(\bar{p}))+p\mathrm{E}[V](1-H(V(\bar{p})))}.$$

Differentiating the two sides of (3.15) and using L'Hospital's rule, we can get the mean waiting time $\mathrm{E}[W_v]$ of a customer during a steady state for a Geom/G/1 (PD, MAVs) queue system as follows:

$$\mathrm{E}[W_v]=\frac{p\mathrm{E}[S(S-1)]}{2(1-\rho)}+\frac{p\mathrm{E}[V(V-1)](1-H(V(\bar{p})))}{2(H(V(\bar{p}))+p\mathrm{E}[V](1-H(V(\bar{p}))))}+\frac{p\mathrm{E}[V(V-1)]}{2(1-p\mathrm{E}[V])}.$$

3.3.3 Analysis of Service Cycle

According to the number J of consecutive vacations [1], [13], we have

$$\begin{aligned}&P(J\geq 1)=1,\\&P(J\geq j)=P(H\geq j)P(V_1+\cdots+V_{j-1}<T)=(V(\bar{p}))^{j-1}\sum_{k=j}^{\infty}h_k,\quad j\geq 2;\end{aligned}\tag{3.17}$$

thus the P.G.F. $J(z)$ of J can be given as

$$J(z)=1-\frac{1-z}{1-V(\bar{p})z}(1-H(V(\bar{p})z)).\tag{3.18}$$

Vacation time lengths in the following two cases are: (1) if a customer is present at a vacation start instant, the total time length is the time of a vacation; (2) if there are no customers present at a vacation start instant, the total vacation time length is the sum of the time lengths of a random number of vacations. Concluding from the two cases above, we can get P.G.F. $V_G(z)$ of the total time length V_G for consecutive vacations as follows:

$$\begin{aligned}V_G(z)=&\frac{1-(1-H(V(\bar{p})))(1-p\mathrm{E}[V])}{1+H(V(\bar{p}))(1-p\mathrm{E}[V])}V(z)+\frac{1-p\mathrm{E}[V]}{1+H(V(\bar{p}))(1-p\mathrm{E}[V])}\\&\times\left(1-\frac{1-V(z)}{1-V(\bar{p})V(z)}(1-H(V(\bar{p})V(z)))\right).\end{aligned}\tag{3.19}$$

Therefore, the mean total time length of a vacation can be obtained as

$$\begin{aligned}\mathrm{E}[V_G]=&\frac{1-(1-H(V(\bar{p})))(1-p\mathrm{E}[V])}{1+H(V(\bar{p}))(1-p\mathrm{E}[V])}\mathrm{E}[V]\\&+\frac{1-p\mathrm{E}[V]}{1+H(V(\bar{p}))(1-p\mathrm{E}[V])}\times\frac{1-H(V(\bar{p}))}{1-V(\bar{p})}\mathrm{E}[V].\end{aligned}\tag{3.20}$$

In a Geom/G/1 (PD, MAVs) queue model, the server is usually in an idle state. If there are customers in the system at the start instant of the vacation, the idle period will be zero after the completion of a vacation. If there are no customers

in the system at the start instant of the vacation, and there are still no customers at the Jth vacation completion instant, the time length I_v of the server's idle period is the inter-arrival time following a nonnegative exponential distribution. We can give the mean $\mathrm{E}[I_v]$ of I_v as

$$\mathrm{E}[I_v] = \frac{1}{p}\frac{(1-p\mathrm{E}[V])H(V(\bar{p}))}{1+H(V(\bar{p}))(1-p\mathrm{E}[V])}. \tag{3.21}$$

According to a pure decrement service strategy, we know that the service period in the new model presented in this chapter is identical to the busy period in a classical Geom/G/1 queue system. This means the P.G.F. $S_p(z)$ of the service period S_p in the queue models of [5] and [13] satisfies the following and the mean length of the service period is given by

$$S_p(z) = S(zS_p((1-p(1-z)))), \qquad \mathrm{E}[S_p] = \frac{1}{\mu - p}.$$

We call the intermediate time between two continuous start instants of the service a service cycle, denoted by C. The mean of the service cycle $\mathrm{E}[C]$ can thus be obtained as follows:

$$\begin{aligned}\mathrm{E}[C] &= \mathrm{E}[S_p] + \mathrm{E}[V_G] + \mathrm{E}[I_v] \\ &= \frac{1-V(\bar{p})+V(\bar{p})(1-p\mathrm{E}[V])(1-H(V(\bar{p})))}{(1+H(V(\bar{p})))(1-p\mathrm{E}[V])(1-V(\bar{p}))}\mathrm{E}[V] \\ &\quad + \frac{1}{p}\frac{(1-p\mathrm{E}[V])H(V(\bar{p}))}{1+H(V(\bar{p}))(1-p\mathrm{E}[V])} + \frac{1}{\mu-p}.\end{aligned} \tag{3.22}$$

Let p_B, p_V, and p_I be the probabilities that the server is in a busy, vacation, or idle state, respectively. We can give that

$$\begin{aligned}p_B &= \frac{\mathrm{E}[S_p]}{\mathrm{E}[C]} = \frac{1}{\mathrm{E}[C](\mu-p)}, \\ p_V &= \frac{\mathrm{E}[V](1-V(\bar{p})+V(\bar{p})(1-p\mathrm{E}[V])(1-H(V(\bar{p}))))}{\mathrm{E}[C](1+H(V(\bar{p})))(1-p\mathrm{E}[V])(1-V(\bar{p}))}, \\ p_I &= \frac{(1-p\mathrm{E}[V])H(V(\bar{p}))}{p\mathrm{E}[C](1+H(V(\bar{p}))(1-p\mathrm{E}[V]))}.\end{aligned} \tag{3.23}$$

3.4 Special Cases

If the random variable H is supposed to have different probability distributions, we can derive some vacation queueing systems with a pure decrement service as special cases of the model presented in this chapter as follows:

Example 3.1. Pure decrement service Geom/G/1 queue with multiple vacations—Geom/G/1 (PD, MV).

If $H \to \infty$, the queue turns into a pure decrement service Geom/G/1 queue with multiple vacations. There is no idle state in the system, and $H(z) = 0$. Then the P.G.F.s of the additional queue lengths L_d, L_r and the additional delays W_d, W_r are respectively given by

$$L_d(z) = \frac{1 - V(1 - p(1-z))}{p\mathrm{E}[V](1-z)}, \quad L_r(z) = \frac{(1 - p\mathrm{E}[V])(1-z)}{V(1-p(1-z)) - z},$$
$$W_d(z) = \frac{1 - V(z)}{p\mathrm{E}[V](1-z)}, \quad W_r(z) = \frac{(1 - p\mathrm{E}[V])(1-z)}{(1-z) - p(1-V(z))}. \tag{3.24}$$

Equation (3.24) corresponds with the results given in [5], [10] and [13].

Example 3.2. Pure decrement service Geom/G/1 queue with single vacation—Geom/ G/1 (PD, SV).

If $H = 1$, the system turns into a pure decrement service Geom/G/1 queue with a single vacation. There is an idle state in the system, and $H(z) = z$. Then the P.G.Fs. of the additional queue lengths L_d, L_r and the additional delays W_d, W_r are respectively given by

$$L_d(z) = \frac{1 - V(\bar{p})z - (1 - V(\bar{p}))V(1 - p(1-z))}{(V(\bar{p}) + p\mathrm{E}[V](1 - V(\bar{p})))(1-z)},$$
$$L_r(z) = \frac{(1 - p\mathrm{E}[V])(1-z)}{V(1-p(1-z)) - z},$$
$$W_d(z) = \frac{p(1 - V(\bar{p}))(1 - V(z)) + V(\bar{p})(1-z)}{(V(\bar{p}) + p\mathrm{E}[V](1 - V(\bar{p})))(1-z)},$$
$$W_r(z) = \frac{(1 - p\mathrm{E}[V])(1-z)}{(1-z) - p(1-V(z))}. \tag{3.25}$$

Equation (3.25) corresponds with the results given in [5] and [13].

Example 3.3. The number of vacations H follows a Poisson distribution in a Geom/G/1 queue with a pure decrement service strategy—Geom/G/1 (PD, PV).

If the number of vacations follows a Poisson distribution with a parameter λ, namely $P(H = i) = (\lambda^i / i!)e^{-\lambda}, \lambda > 0, i = 0, 1, 2, \ldots$, then $H(z) = e^{\lambda(z-1)}$. Substituting $H(V(\bar{p})) = e^{\lambda(V(\bar{p})-1)}$ into (3.5) and (3.13), the P.G.Fs. of the additional queue lengths L_d, L_r and the additional delays W_d, W_r are given by

$$L_d(z) = \frac{1 - V(1 - p(1-z)) + e^{\lambda(V(\bar{p})-1)}V(1 - p(1-z)) - e^{\lambda(V(\bar{p})-1)}z}{(e^{\lambda(V(\bar{p})-1)} + p\mathrm{E}[V](1 - e^{\lambda(V(\bar{p})-1)}))(1-z)},$$
$$L_r(z) = \frac{(1 - p\mathrm{E}[V])(1-z)}{V(1-p(1-z)) - z},$$
$$W_d(z) = \frac{p(1 - e^{\lambda(V(\bar{p})-1)})(1 - V(z)) + e^{\lambda(V(\bar{p})-1)}(1-z)}{(e^{\lambda(V(\bar{p})-1)} + p\mathrm{E}[V](1 - e^{\lambda(V(\bar{p})-1)}))(1-z)},$$
$$W_r(z) = \frac{(1 - p\mathrm{E}[V])(1-z)}{(1-z) - p(1-V(z))}. \tag{3.26}$$

The special cases mentioned above correspond to different probability distributions of H, and we can obtain different pure decrement service queue models with vacations. From these analyses, we can conclude that the model presented in this chapter is a general model including many special queue models.

3.5 Numerical Results

In this section, we present some numerical results that provide insight into the system behavior. Using the equations presented in Sect. 3.3, we can numerically compare the performance measures of the systems for three different Geom/G/1 (PD, MAVs) queue models: the pure decrement service Geom/G/1 queue with multiple vacations, the pure decrement service Geom/G/1 queue with single vacation and the model where the number of vacations H follows a Poisson distribution in Geom/G/1 queue with a pure decrement service strategy.

Here we assume that the service time S and the time length V of a vacation follow geometric distributions; that is, S follows a geometric distribution with mean $1/\mu = 10$. V follows a geometric distribution with mean $\mathrm{E}[V] = 10$. As we presented in Sect. 3.2, if $H \to \infty$, the model corresponds to a Geom/G/1 (PD, MV) queue. If $H = 1$, the model corresponds to Geom/G/1 (PD, SV) queue. If H follows a Poisson distribution, the model corresponds to a Geom/G/1 (PD, PV) queue. Parameter $\lambda = 0.1$, traffic intensity ρ range from 0.1 to 0.8.

Figure 3.2 shows the mean queue length $\mathrm{E}[L_v]$ as a function of the the traffic intensity ρ with three cases of H; that is, $H \to \infty$ for a Geom/G/1 (PD, MV) queue, $H = 1$ for a Geom/G/1 (PD, SV) queue, and H follows a Poisson distribution for a Geom/G/1 (PD, PV) queue. We can find that when ρ increases, $\mathrm{E}[L_v]$ increases to a

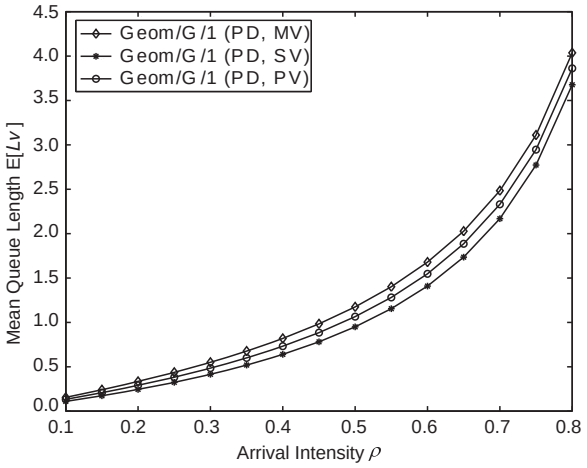


Fig. 3.2 Mean queue length $\mathrm{E}[L_v]$ versus traffic intensity ρ.

high level for all the cases. This is because the larger ρ is, the higher the possibility that there will be customers arriving during the server cycle C. We also note that the mean queue length $\mathrm{E}[L_v]$ of Geom/G/1 (PD, MV) is larger than that of Geom/G/1 (PD, SV) and Geom/G/1 (PD, PV). This is because the longer the vacation times are, the larger the mean queue length $\mathrm{E}[L_v]$ will be.

Figure 3.3 shows how the mean waiting time $\mathrm{E}[W_v]$ changes with the traffic intensity ρ for the three different cases of H; that is, $H \to \infty$ for a Geom/G/1 (PD, MV) queue, $H = 1$ for a Geom/G/1 (PD, SV) queue, and H follows a Poisson distribution for a Geom/G/1 (PD, PV) queue. We can find that when ρ increases, $\mathrm{E}[W_v]$ increases to a high level. This is because the greater ρ is, the higher the possibility that there will be customers arriving during the server cycle C; then the mean waiting time will be greater. We also note that the mean waiting time $\mathrm{E}[W_v]$ of Geom/G/1 (PD, MV) is longer than that of Geom/G/1 (PD, SV) and Geom/G/1 (PD, PV). This is because the longer the vacation time lengths are, the greater the mean waiting time $\mathrm{E}[W_v]$ will be.

In Fig. 3.4, we can observe that, for the Geom/G/1 (PD, MV) queue, when ρ increases, the mean service cycle $\mathrm{E}[C]$ of Geom/G/1 (PD, MV) increases, too. It can also be noted that the curves of the mean service cycle $\mathrm{E}[C]$ for the Geom/G/1 (PD, SV) queue and Geom/G/1 (PD, PV) queue follow two stages. In the first stage, the heavier the traffic intensity ρ is, the lower the mean service cycle $\mathrm{E}[C]$ will be. In the second stage, the heavier the traffic intensity ρ is, the higher the mean service cycle $\mathrm{E}[C]$ will be.

In Fig. 3.5, we plot the probability for the system being at the various states as a function of the traffic intensity ρ in Geom/G/1 (PD, PV). It can be observed that when ρ increases, the probability for the system being either in a busy or vacation state increases, whereas the probability of the system being in an idle state decreases and limits to zero. This is because the greater ρ is, the more customers will arrive,

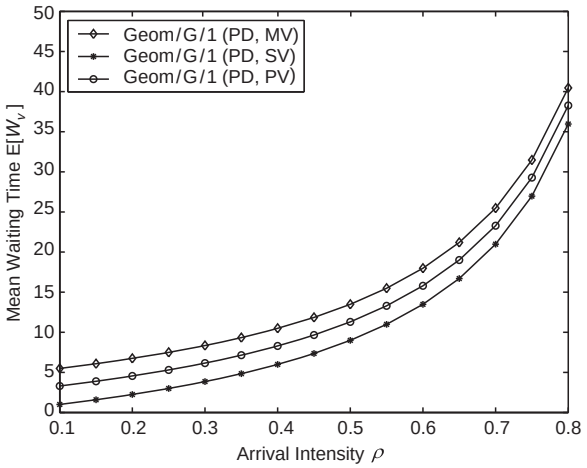


Fig. 3.3 Mean waiting time $\mathrm{E}[W_v]$ versus traffic intensity ρ.

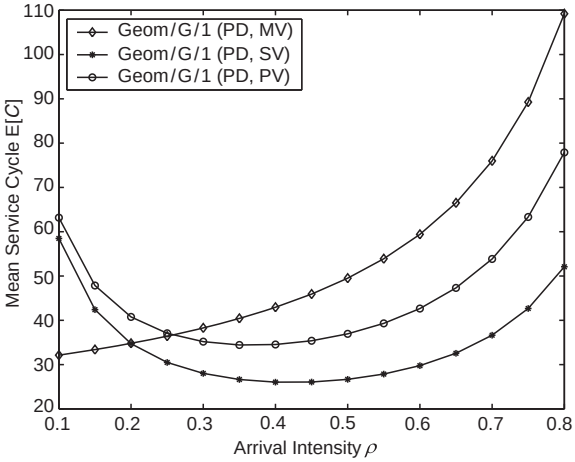


Fig. 3.4 Mean server cycle E$[C]$ versus traffic intensity ρ.

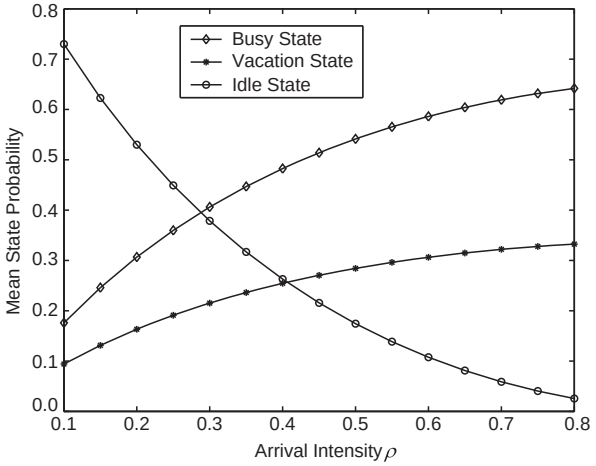


Fig. 3.5 Mean state probability versus traffic intensity ρ.

and so the probability of the system being in a busy or a vacation state will increase, whereas the probability for the system being in an idle state will be smaller.

3.6 Conclusions

In this chapter, we presented a detailed description of a Geom/G/1 queue model with a pure decrement service strategy and multiple adaptive vacations. By using the method of an embedded Markov chain, we derived the P.G.F.s of the queue length and the customers' waiting time. Furthermore, we presented the stochastic

decompositions for the additional queue length and the additional delay. Lastly, we obtained the probabilities of the server being in the various states of busy, vacation, or idle, respectively. The model is an extension for many special multiple adaptive vacation queue models with a pure decrement service strategy. When applying to communication networks, it is especially useful for solving problems associated with network flow.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China (No. 10671170) and Natural Science Foundation of Hebei Province (No. F2008000864), and was supported in part by GRANT-IN-AID FOR SCIENTIFIC RESEARCH (No. 19500070) and MEXT.ORC (2004-2008), Japan.

References

1. N. Tian, Multi-stage adaptive vacation policies in an M/G/1 queueing system, *Applied Mathematics*, vol. 5, no. 4, pp. 12–18, 1992 (in Chinese).
2. G. Zhang and N. Tian, Discrete time Geo/G/1 queue with multiple adaptive vacations, *Queueing Systems*, vol. 38, no. 4, pp. 419–429, 2001.
3. O. Boxma and U. Yechiali, An M/G/1 queue with multiple types of feedback and gated vacations, *Journal of Applied Probability*, vol. 34, no. 3, pp. 773–784, 1997.
4. H. Takagi, Mean message waiting time in a symmetric polling system, in *Proc. Performance'84, E. Gelenbe (Editor)*, 1985.
5. H. Takagi, *Queueing Analysis, Volume 3: Discrete-Time Systems*. Amsterdam: Elsevier Science, 1993.
6. D. Wu and H. Takagi, M/G/1 queue with multiple working vacations, *Performance Evaluation*, vol. 63, no. 7, pp. 654–681, 2006.
7. N. Tian and G. Zhang, A discrete-time GI/Geo/1 queue with multiple vacations, *Queueing Systems*, vol. 40, no. 3, pp. 283–294, 2002.
8. D. Fiems and H. Bruneel, Analysis of a discrete time queueing system with timed vacations, *Queueing Systems*, vol. 42, no. 3, pp. 243–254, 2002.
9. J. Cohen, *The Single Server Queue*. Amsterdam: North-Holland, 1982.
10. D. Gross and C. M. Harris, *Fundamentals of Queueing Theory (Second edition)*. New York: John Wiley & Sons, 1985.
11. W. Bischof, Analysis of M/G/1 queues with setup times and vacations under six different service disciplines, *Queueing Systems*, vol. 39, no. 4, pp. 265–301, 2001.
12. Z. Niu, Y. Takahashi, and N. Endo, Performance evaluation of SVC-based IP-over-ATM network, *IEICE Transactions on Communications*, vol. E81-B, pp. 948–957, 1998.
13. N. Tian and G. Zhang, *Vacation Queueing Models-Theory and Applications*. New York: Springer-Verlag, 2006.

Chapter 4
Performance Analysis of an M/M/1 Working Vacation Queue with Setup Times

Xiuli Xu and Naishuo Tian

Abstract We investigate an M/M/1 working vacation queue with setup times, using a quasi birth-and-death process and a matrix-geometric solution method to derive the distributions for the stationary queue length and the waiting time of a customer in the system. Furthermore, we get stochastic decomposition structures of stationary indices, and obtain the distributions of the additional queue length and additional delay. Finally, numerical examples are presented.

4.1 Introduction

The vacation queue models have been investigated extensively in view of their application in computer systems, communication networks, and production managing. In a classical vacation queue, the server completely stops serving customers and may do some additional work or maintain servers during a vacation. Various vacation policies provide more flexibility for optimal design and operating control of the system. The details can be seen in the monographs of Takagi [1], Tian and Zhang [2].

Servi and Finn [3] introduced a class of semi-vacation policies: the server works at a lower rate rather than completely stopping service during a vacation. Such a vacation is called a working vacation (WV). Part of the service ability keeps the system operating at a lower speed during a vacation. If service speed degenerates to zero in a working vacation, the working vacation queue becomes a classical vacation queue model. Therefore, the working vacation queue is the generalization of the classical vacation queue and the analysis of this kind of model is more complicated

X. Xu
College of Sciences, Yanshan University, Qinhuangdao 066004, China
e-mail: xxl-ysu@163.com

N. Tian
College of Sciences, Yanshan University, Qinhuangdao 066004, China
e-mail: tiannsh@ysu.edu.cn

W. Yue et al. (eds.), *Advances in Queueing Theory and Network Applications,*

than the previous work. In view of partly utilizing service ability during a vacation, the working vacation policy of the queue with a single server is similar to the partial servers' vacation policy in a multiserver queue; for the details see Tian and Zhang [4] and Xu and Zhang [5].

Servi and Finn [3] studied an M/M/1 queue with multiple working vacations, and obtained the probability generating function (P.G.F.) of the number of customers in the system and the Laplace–Stieltjes transform (LST) of the waiting time distribution, and applied the results to performance analysis of a gateway router in fiber communication networks.

On the basis of [3], Liu, Xu, and Tian [6] gave simple explicit expressions of distributions for the stationary queue length and waiting time that have an intuitive probability interpretation. Furthermore, the authors got stochastic decomposition structures of stationary indices, derived an expected regular busy period, and an expected busy cycle. Moreover, Kim, Choi, and Chae [7] and Wu and Takagi [8] generalized the work of [3] to an M/G/1 queue with multiple working vacations. Li and Tian [9] examined a discrete time GI/Geom/1 queue with working vacation and service interruption. Baba [10] discussed a GI/M/1 queue with multiple working vacations. Banik, Gupta, and Pathak [11] studied a GI/M/1/N working vacation queue with limited waiting space.

In this chapter, we investigate an M/M/1 queue with single working vacation and setup times. If the setup time equals zero, our model becomes an M/M/1 queue with a single working vacation. Furthermore, if the working vacation time equals zero at the same time, this model boils down to a classical M/M/1 queue. Therefore, our model has a more comprehensive application background.

The rest of this chapter is organized as follows. In Sect. 4.2 we describe the quasi birth-and-death process model of the system and get the explicit expression of the rate matrix which assures that various system indices have an analytic solution. Section 4.3 and Sect. 4.4, respectively, discuss the number of customers in the system and the waiting time of a customer, give their stochastic decomposition structures, and obtain the distributions of additional queue length and additional delay. Furthermore, Sect. 4.5 includes some numerical examples in order to give changing curves of performance indices with the change of system parameters. Concluding remarks are given in Sect. 4.6.

4.2 Model Description and Preliminary

A policy of single working vacation and setup times is introduced into a classical M/M/1 queue with arrival rate λ and service rate μ_b. The server begins a working vacation of random length at the instants when the queue becomes empty, and vacation duration V follows an exponential distribution with parameter θ. During a working vacation arriving customers are served at a rate of μ_v ($\mu_v < \mu_b$) according to arrival order. When a vacation ends, if there are customers in the queue, the server changes the service rate from μ_v to μ_b, and a regular busy period starts.

Otherwise, the server begins a closed-down period. During a closed-down period, an arriving customer cannot be served immediately and has to experience a period of setup time, setup duration S follows an exponential distribution with parameter α, a regular busy period starts after a setup period.

Working vacation V is an operating period at a lower speed. When the number of customers in the system is relatively few, we set a lower speed operating period and a closed-down period in order to economize operating cost together with serving customers; these two periods have essential differences because customers can be served in the former period and not in the latter period. Therefore, this working vacation and setup time policy has practical significance in the optimal design of the system. Similar to previous denotation, this queue is referred to as an M/M/1 queue with single working vacation and setup times.

We assume that interarrival times, service times, working vacation times, and setup times are mutually independent. In addition, the service order is first-in-first out (FIFO).

Let $Q(t)$ be the number of customers in the system at time t and

$$J(t) = \begin{cases} 0, \text{ the system is in a working vacation period at time } t \\ 1, \text{ the system is in a setup period or closed-down period at time } t \\ 2, \text{ the system is in a regular busy period at time } t, \end{cases}$$

then $\{Q(t), J(t)\}$ is a Markov process with the state space

$$\Omega = \Big\{(0,0),(0,1)\Big\} \bigcup \Big\{(k,j) : k \geq 1,\, j = 0,1,2\Big\}.$$

Using the lexicographical sequence for the states, the infinitesimal generator can be written as follows:

$$\widetilde{\mathbf{Q}} = \begin{bmatrix} \mathbf{A}_{00} & \mathbf{A}_{01} & & & \\ \mathbf{B}_{10} & \mathbf{A} & \mathbf{C} & & \\ & \mathbf{B} & \mathbf{A} & \mathbf{C} & \\ & & \mathbf{B} & \mathbf{A} & \mathbf{C} \\ & & & \vdots & \vdots & \vdots \end{bmatrix},$$

where

$$\mathbf{A}_{00} = \begin{bmatrix} -(\lambda+\theta) & \theta \\ 0 & -\lambda \end{bmatrix}, \qquad \mathbf{A}_{01} = \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \end{bmatrix}, \qquad \mathbf{B}_{10} = \begin{bmatrix} \mu_v & 0 \\ 0 & 0 \\ \mu_b & 0 \end{bmatrix},$$

and

$$\mathbf{A} = \begin{bmatrix} -(\lambda+\theta+\mu_v) & 0 & \theta \\ 0 & -(\lambda+\alpha) & \alpha \\ 0 & 0 & -(\lambda+\mu_b) \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} \mu_v & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \mu_b \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}.$$

The matrix $\widetilde{\mathbf{Q}}$ has a block-tridiagonal structure which indicates that $\{Q(t), J(t)\}$ is a quasi birth-and-death (QBD) process, see Neuts [12] or Latouche and Ramaswami [13]. To analyze this QBD process, it is necessary to solve for the minimal nonnegative solution of the matrix quadratic equation as follows:

$$\mathbf{R}^2\mathbf{B} + \mathbf{R}\mathbf{A} + \mathbf{C} = 0 \tag{4.1}$$

and this solution is called the rate matrix and denoted by $\mathbf{R}$. Obviously, we have the following lemma.

Lemma 4.1. *If $\rho = \lambda(\mu_b)^{-1} < 1$, the matrix equation (4.1) has the minimal nonnegative solution*

$$\mathbf{R} = \begin{bmatrix} r & 0 & \dfrac{\theta\, r}{\mu_b(1-r)} \\ 0 & \dfrac{\lambda}{\lambda+\alpha} & \rho \\ 0 & 0 & \rho \end{bmatrix}, \tag{4.2}$$

where

$$r = \frac{1}{2\mu_v}\left(\lambda + \theta + \mu_v - \sqrt{(\lambda+\theta+\mu_v)^2 - 4\lambda\mu_v}\right)$$

and $0 < r < 1$.

Because r satisfies the equation

$$\mu_v r^2 - (\lambda + \theta + \mu_v)r + \lambda = 0,$$

dividing both sides of this equation by r, we get

$$\lambda + \theta + \mu_v(1-r) = \frac{\lambda}{r}.$$

Equivalently, we have

$$\frac{\theta}{1-r} + \mu_v = \frac{\lambda}{r}. \tag{4.3}$$

Lemma 4.2. *The QBD process $\{Q(t), J(t)\}$ is positive recurrent if and only if $\rho < 1$.*

Proof. Based on Theorem 3.1.1 of Neuts [12], the QBD process $\{Q(t), J(t)\}$ is positive recurrent if and only if the spectral radius $\mathrm{SP}(\mathbf{R})$ of the rate matrix $\mathbf{R}$ is less than 1, and set of equations $(x_0, x_1, x_2, x_3, x_4, x_5)B[\mathbf{R}] = 0$ has a positive solution, where

$$B[\mathbf{R}] = \begin{bmatrix} \mathbf{A}_{00} & \mathbf{A}_{01} \\ \mathbf{B}_{10} & \mathbf{RB}+\mathbf{A} \end{bmatrix}$$

$$= \begin{bmatrix} -(\lambda+\theta) & \theta & \lambda & 0 & 0 \\ 0 & -\lambda & 0 & \lambda & 0 \\ \mu_v & 0 & -\dfrac{\lambda}{r} & 0 & \dfrac{\theta}{1-r} \\ 0 & 0 & 0 & -(\lambda+\alpha) & \lambda+\alpha \\ \mu_b & 0 & 0 & 0 & -\mu_b \end{bmatrix}. \tag{4.4}$$

$B[\mathbf{R}]$ is an irreducible and aperiodic generator with finite state. Therefore, $(x_0, x_1, x_2, x_3, x_4, x_5)B[\mathbf{R}] = 0$ has a positive solution (e.g., the equilibrium probability vector of $B[\mathbf{R}]$ is a positive solution). Thus, process $\{Q(t), J(t)\}$ is positive recurrent if and only if

$$\mathrm{SP}(\mathbf{R}) = \max\Big(r, \frac{\lambda}{\lambda+\alpha}, \rho\Big) < 1.$$

Note that $0 < r < 1$ and $0 < \lambda/(\lambda+\alpha) < 1$, the above relation means that $\rho < 1$.

4.3 Queue Length Distribution

If $\rho < 1$, let (Q, J) be the stationary limit of the QBD process $\{Q(t), J(t)\}$. Let

$$\pi_k = \begin{cases} (\pi_{00}, \pi_{01}), & k = 0 \\ (\pi_{k0}, \pi_{k1}, \pi_{k2}), & k \geq 1, \end{cases}$$

$$\pi_{kj} = P\{Q = k, J = j\} = \lim_{t\to\infty} P\{Q(t) = k, J(t) = j\}, \qquad (k, j) \in \Omega.$$

Theorem 4.1. *If $\rho < 1$, the stationary probability distribution of (Q, J) is*

$$\begin{cases} \pi_{k0} = Kr^k, & k \geq 0 \\ \pi_{k1} = K\dfrac{\theta}{\lambda}\Big(\dfrac{\lambda}{\lambda+\alpha}\Big)^k, & k \geq 0 \\ \pi_{k2} = K\Bigg[\dfrac{\theta r}{\mu_b(1-r)}\displaystyle\sum_{j=1}^{k-1} r^j\rho^{k-1-j} + \dfrac{\theta}{\lambda+\alpha}\sum_{j=1}^{k-1}\rho^j\Big(\dfrac{\lambda}{\lambda+\alpha}\Big)^{k-1-j} \\ \qquad\quad + \dfrac{\theta}{\mu_b(1-r)}\rho^{k-1}\Bigg], \qquad k \geq 1, \end{cases} \tag{4.5}$$

where we assume that the null sum is equal to zero and

$$K = \left[\frac{1}{1-r} + \frac{\theta}{\lambda} + \frac{\theta}{\alpha(1-\rho)} + \frac{\theta(1-r+r^2)}{\mu_b(1-r)^2(1-\rho)}\right]^{-1}.$$

Proof. With the matrix-geometric solution method (see Neuts [12]), we have

$$(\pi_{k0}, \pi_{k1}, \pi_{k2}) = (\pi_{10}, \pi_{11}, \pi_{12})\mathbf{R}^{k-1}, \qquad k \geq 1 \tag{4.6}$$

and $(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}, \pi_{12})$ satisfies the set of equations as follows:

$$(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}, \pi_{12})B[\mathbf{R}] = 0.$$

Substituting $B[\mathbf{R}]$ into the above equation, we obtain the set of equations as

$$\begin{cases} -(\lambda+\theta)\pi_{00} + \mu_v\pi_{10} + \mu_b\pi_{12} = 0 \\ \theta\pi_{0,0} - \lambda\pi_{01} = 0 \\ \lambda\pi_{0,0} - \dfrac{\lambda}{r}\pi_{10} = 0 \\ \lambda\pi_{01} - (\lambda+\alpha)\pi_{11} = 0 \\ \dfrac{\theta}{1-r}\pi_{10} + (\lambda+\alpha)\pi_{1,1} - \mu_b\pi_{12} = 0. \end{cases}$$

Taking $\pi_{00} = K$, we get

$$(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}, \pi_{12}) = K\left(1, \frac{\theta}{\lambda}, r, \frac{\theta}{\lambda+\alpha}, \frac{\theta}{\mu_b(1-r)}\right).$$

From (4.2), utilizing the rule of matrix multiplication and iterative method, we can easily get the expression of $\mathbf{R}^k$ as

$$\mathbf{R}^k = \begin{bmatrix} r^k & 0 & \dfrac{\theta}{\mu_b(1-r)}\displaystyle\sum_{j=1}^{k} r^j\rho^{k-j} \\ 0 & \left(\dfrac{\lambda}{\lambda+\alpha}\right)^k & \displaystyle\sum_{j=1}^{k}\rho^j\left(\dfrac{\lambda}{\lambda+\alpha}\right)^{k-j} \\ 0 & 0 & \rho^k \end{bmatrix}, \quad k \geq 1.$$

Furthermore, substituting $(\pi_{10}, \pi_{11}, \pi_{12})$ and $\mathbf{R}^{k-1}$ into (4.6), we obtain (4.5). Finally, the constant factor K can be determined by the normalization condition.

With (4.5), the probabilities that the system is in a working vacation period, a closed-down period, a setup period, and a regular busy period in steady-state are as follows, respectively.

$$\begin{aligned} &P\{J=0\} = \sum_{k=0}^{\infty}\pi_{k0} = K\frac{1}{1-r}, \\ &P\{\text{the server is in a closed-down period}\} = \pi_{01} = K\frac{\theta}{\lambda}, \\ &P\{\text{the server is in a setup period}\} = \sum_{k=1}^{\infty}\pi_{k1} = K\frac{\theta}{\alpha}, \\ &P\{J=2\} = \sum_{k=1}^{\infty}\pi_{k2} = K\left[\frac{\theta\rho}{\alpha(1-\rho)} + \frac{\theta(1-r+r^2)}{\mu_b(1-r)^2(1-\rho)}\right]. \end{aligned} \tag{4.7}$$

Theorem 4.2. *If $\rho < 1$ and $\mu_b > \mu_v$, the number of customers Q in the system can be decomposed into the sum of two independent random variables: $Q = Q_0 + Q_d$, where Q_0 is the number of customers of a classical M/M/1 queue in steady state and follows a geometric distribution with parameter $1-\rho$. An additional number of customers Q_d has a modified geometric distribution as follows:*

$$P\{Q_d = k\} = \begin{cases} K^*\delta_1, & k = 0 \\ K^*\delta_2, & k = 1 \\ K^*\delta_3(1-r)r^{k-1} + K^*\delta_4(1-\beta)\beta^{k-1}, & k \geq 2, \end{cases} \tag{4.8}$$

where

$$\begin{aligned}
\beta &= \frac{\lambda}{\lambda+\alpha}, \\
\delta_1 &= (1-r)(1-\beta)\frac{\lambda+\theta}{\lambda}, \\
\delta_2 &= \Big[\frac{\theta}{\mu_b} + (r-\rho)(1-r) + \frac{\theta(1-r)(\beta-\rho)}{\lambda}\Big](1-\beta), \\
\delta_3 &= \Big[(r-\rho) + \frac{\theta r}{\mu_b(1-r)}\Big](1-\beta), \\
\delta_4 &= \frac{\theta(1-r)}{\lambda+\alpha}, \\
K^* &= \Big\{(1-\beta)(1-\rho) + \frac{\theta(1-\rho)(1-r)(1-\beta)}{\lambda} + \frac{\theta(1-r)}{\lambda+\alpha} \\
&\quad + \frac{\theta(1-\beta)(1-r+r^2)}{\mu_b(1-r)}\Big\}^{-1}.
\end{aligned}$$

Proof. Denote $\lambda/(\lambda+\alpha) = \beta$ and with (4.5), the Probability Generation Function (P.G.F.) of Q can be written as follows:

$$\begin{aligned}
Q(z) &= \sum_{k=0}^{\infty} z^k \pi_{k0} + \sum_{k=0}^{\infty} z^k \pi_{k1} + \sum_{k=1}^{\infty} z^k \pi_{k2} \\
&= K\Big\{\frac{1}{1-rz} + \frac{\theta}{\lambda}\frac{1}{1-\beta z} + \frac{\theta r^2}{\mu_b(1-r)(\rho-r)}\Big[\frac{z}{1-\rho z} - \frac{z}{1-rz}\Big] \\
&\quad + \frac{\theta\rho}{\lambda-(\lambda+\alpha)\rho}\Big[\frac{z}{1-\beta z} - \frac{z}{1-\rho z}\Big] + \frac{\theta}{\mu_b(1-r)}\frac{z}{1-\rho z}\Big\} \\
&= \frac{1-\rho}{1-\rho z}K^*\Big\{(1-\beta)\frac{1-r}{1-rz}(1-\rho z) + \frac{\theta(1-r)}{\lambda}\frac{1-\beta}{1-\beta z}(1-\rho z) \\
&\quad + \frac{\theta}{\mu_b}(1-\beta)z + \frac{\theta r(1-\beta)}{\mu_b(1-r)}\frac{r(1-r)z^2}{1-rz} + \frac{\theta\rho(1-r)}{\lambda}\frac{\beta(1-\beta)z^2}{1-\beta z}\Big\}.
\end{aligned}$$

Note that

$$\frac{1-r}{1-rz}(1-\rho z) = (1-r) + (r-\rho)\frac{(1-r)z}{1-rz},$$
$$\frac{1-\beta}{1-\beta z}(1-\rho z) = (1-\beta) + (\beta-\rho)\frac{(1-\beta)z}{1-\beta z}.$$

$Q(z)$ can be rewritten as

$$\begin{aligned} Q(z) &= \frac{1-\rho}{1-\rho z}K^*\left\{\delta_1 + \delta_2 z + \delta_3\frac{r(1-r)z^2}{1-rz} + \delta_4\frac{\beta(1-\beta)z^2}{1-\beta z}\right\} \\ &= \frac{1-\rho}{1-\rho z}Q_d(z) \end{aligned}$$

and we can prove that $\delta_1 + \delta_2 + r\delta_3 + \beta\delta_4 = (K^*)^{-1}$. Therefore, $Q_d(z)$ is a P.G.F. Expanding $Q_d(z)$ into the power series of z, we get the distribution of an additional number of customers Q_d. Thus, we can get (4.8).

With the stochastic decomposition structure in Theorem 4.2, we can easily get the means as follows:

$$\mathrm{E}[Q_d] = K^*\left[\delta_2 + \frac{2r-r^2}{1-r}\delta_3 + \frac{2\beta-\beta^2}{1-\beta}\delta_4\right], \qquad \mathrm{E}[Q] = \frac{\rho}{1-\rho} + \mathrm{E}[Q_d].$$

4.4 Waiting Time Analysis

Denoting the waiting time of a customer in the system by W, we have the following stochastic decomposition results.

Theorem 4.3. *If $\rho < 1$ and $\mu_b > \mu_v$, the waiting time W can be decomposed into the sum of two independent random variables: $W = W_0 + W_d$ where W_0 is the waiting time of a customer in a corresponding classical M/M/1 queue and has an exponential distribution with parameter $\mu_b(1-\rho)$. Additional delay W_d has the modified exponential distribution and LST as follows:*

$$W_d^*(s) = K^*\left[\sigma_1 + \sigma_2\frac{\gamma}{\gamma+s} + \sigma_3\frac{\alpha}{\alpha+s}\right], \tag{4.9}$$

where

$$\gamma = \frac{\lambda(1-r)}{r}, \qquad \sigma_1 = \delta_1 + \delta_2 - \frac{1-r^2}{r}\delta_3 - \frac{(1-\beta)(2\lambda+\alpha)}{\lambda}\delta_4,$$
$$\sigma_2 = \frac{1}{r}\delta_3, \qquad \sigma_3 = \frac{1}{\beta}\delta_4.$$

Proof. The classical relationship between the P.G.F. of Q and the LST of waiting time W (see [6]) is

$$Q(z) = W^*\big(\lambda(1-z)\big).$$

From Theorem 4.2, the P.G.F. of the number of customers Q is

$$Q(z) = \frac{1-\rho}{1-\rho z} K^* \left\{ \delta_1 + \delta_2 z + \delta_3 \frac{r(1-r)z^2}{1-rz} + \delta_4 \frac{\beta(1-\beta)z^2}{1-\beta z} \right\}. \tag{4.10}$$

Taking $z = 1 - s/\lambda$ in (4.10) and denoting $\lambda(1-r)/r = \gamma$, we have

$$\begin{aligned} W^*(s) &= \frac{\mu_b(1-\rho)}{\mu_b(1-\rho)+s} K^* \left\{ \delta_1 + \delta_2\Big(1 - \frac{s}{\lambda}\Big) + \delta_3 \frac{1-r}{\lambda} \left[\frac{\big(\frac{\lambda}{r}\big)^2}{\gamma+s} - \gamma + s \right] \right. \\ &\quad \left. + \delta_4 \frac{1-\beta}{\lambda} \left[\frac{(\lambda+\alpha)^2}{\alpha+s} - (2\lambda+\alpha) + s \right] \right\} \\ &= \frac{\mu_b(1-\rho)}{\mu_b(1-\rho)+s} K^* \left\{ \sigma_1 + \sigma_2 \frac{\gamma}{\gamma+s} + \sigma_3 \frac{\alpha}{\alpha+s} \right\} \\ &= \frac{\mu_b(1-\rho)}{\mu_b(1-\rho)+s} W_d^*(s). \end{aligned}$$

It is easy to verify that $\sigma_1 + \sigma_2 + \sigma_3 = \delta_1 + \delta_2 + r\delta_3 + \beta\delta_4 = (K^*)^{-1}$. Therefore, $W_d^*(s)$ is a LST.

We can easily get means as follows:

$$\mathrm{E}[W_d] = K^* \left(\sigma_2 \frac{1}{\gamma} + \sigma_3 \frac{1}{\alpha} \right), \qquad \mathrm{E}[W] = \frac{1}{\mu_b(1-\rho)} + \mathrm{E}[W_d].$$

4.5 Numerical Results

Consider an asynchronous transfer mode (ATM) network, where cell arrivals in a switched virtual channel (SVC) form a Poisson process with parameter λ; cell transmission time is an exponential distributed random variable with rate μ_b. When a SVC finishes cell transmission and becomes empty, we set a period of working vacation, during which arriving cells can be transmitted at a lower rate μ_v ($\mu_v < \mu_b$) immediately. If there are no cells in the SVC after a working vacation, we close down the SVC in order to save the operating cost and need to rebuild a SVC when a cell arrives. The policy of working vacation both takes over cell transmission and saves switching cost. Therefore, our model is fitter for modeling practical situations than [6].

Now, we illustrate the results obtained above numerically and discuss the effect of system parameters on system performance indices. We assume that the service rate μ_b in a regular busy period equals 0.5 and arrival rate λ equals 0.3; at the same time, we assume that setup time is an exponential distributed random variable with

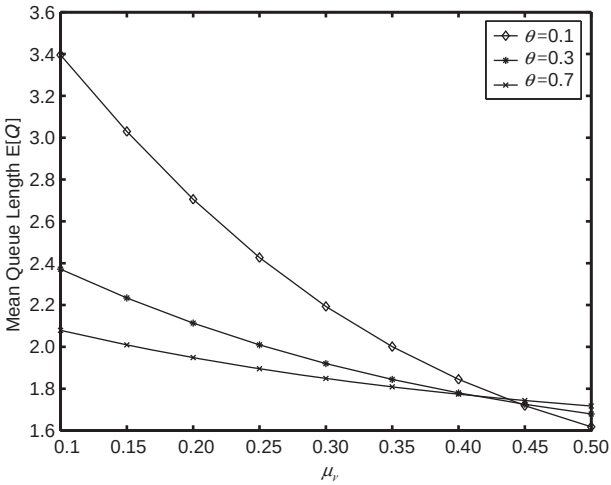


Fig. 4.1 Mean queue length E[Q] versus service rate μ_v in working vacation period.

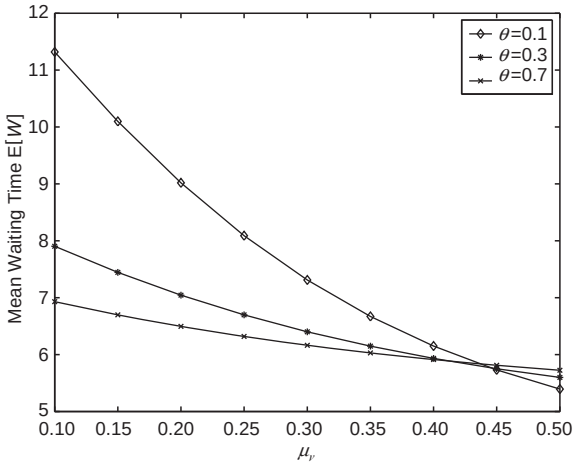


Fig. 4.2 Mean waiting time E[W] versus service rate μ_v in working vacation period.

mean $\alpha = 0.8$. We respectively plot the values of mean queue length E[Q] and mean waiting time E[W] by changing the service rate μ_v in a vacation period, meanwhile, in order to investigate the influence of the mean length $1/\theta$ of a vacation, we show the results for three values of θ. For comparison, we have Figs. 4.1 and 4.2.

On the other hand, we assume that the service times in a service period and in a vacation are exponentially distributed with rate $\mu_b = 0.7$ and $\mu_v = 0.5$, respectively. Moreover, we assume that arrival rate λ is equal to 0.3. We respectively plot the values of mean queue length E[Q] and mean waiting time E[W] by changing setup

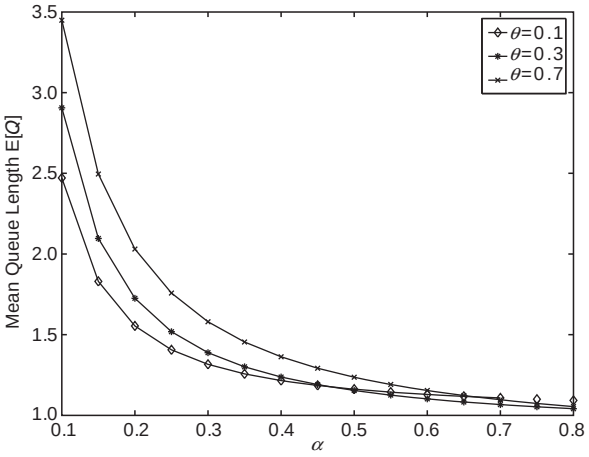


Fig. 4.3 Mean queue length $\mathrm{E}[Q]$ versus setup rate α.

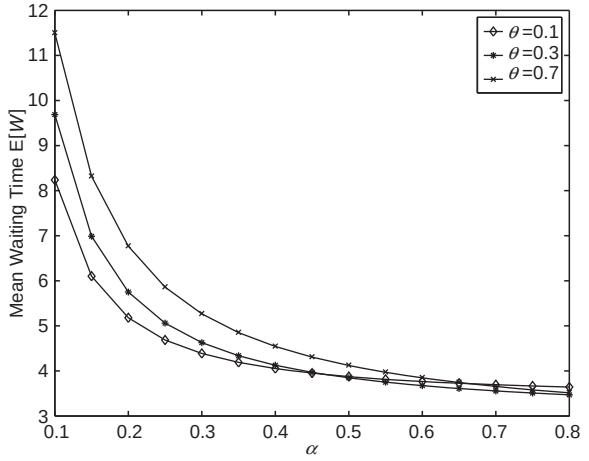


Fig. 4.4 Mean waiting time $\mathrm{E}[W]$ versus setup rate α.

rate α; meanwhile, in order to investigate the influence of the mean length $1/\theta$. of a vacation, we show the results for three values of θ. For comparison, we have Figs. 4.3 and 4.4.

4.6 Conclusions

We proposed a new queueing model with setup times and single working vacation in this chapter, using a quasi birth-and-death process and matrix-geometric solution method to derive the distributions for the stationary queue length and waiting time of

a customer in the system. Furthermore, we got stochastic decomposition structures of stationary indices, and obtained the distributions of the additional queue length and additional delay. The numerical results were presented.

Acknowledgments This work was supported by National Natural Science Foundation of China (No.10671170).

References

1. H. Takagi, *Queueing Analysis, Vol. 1*. Amsterdam: Elsevier Science, 1991.
2. N. Tian and Z. G. Zhang, *Vacation Queueing Models-Theory and Applications*. New York: Springer-Verlag, 2006.
3. L. Servi and S. Finn, M/M/1 queue with working vacations (M/M/1/WV), *Performance Evaluation*, vol. 50, no. 1, pp. 41–52, 2002.
4. N. Tian and Z. G. Zhang, A two threshold vacation policy in multiserver queueing systems, *European Journal Operational Research*, vol. 168, no. 1, pp. 153–163, 2006.
5. X. Xu and Z. G. Zhang, Analysis of multi-server queue with a single vacation (e, d)-policy, *Performance Evaluation*, vol. 63, no. 7, pp. 625–638, 2006.
6. W. Liu, X. Xu, and N. Tian, Some results on the M/M/1 queue with working vacations, *Operations Research Letters*, vol. 35, no. 5, pp. 595–600, 2007.
7. J. Kim, D. Choi, and K. Chae, Analysis of queue-length distribution of the M/G/1 queue with working vacations, in *Proc. Hawaii International Conference on Statistics and Related Fields*, pp. 23–37, 2003.
8. D. Wu and H. Takagi, M/G/1 queue with multiple working vacations, *Performance Evaluation*, vol. 63, no. 7, pp. 654–681, 2006.
9. J. Li and N. Tian, The discrete time GI/Geom/1 queue with working vacation and service interruption, *Applied Mathematics and Computation*, vol. 185, no. 1, pp. 1–10, 2007.
10. Y. Baba, Analysis of a GI/M/1 queue with multiple working vacations, *Operations Research Letters*, vol. 33, no. 2, pp. 201–209, 2005.
11. A. Banik, U. Gupta, and S. Pathak, On the GI/M/1/N queue with multiple working vacations-analytic analysis and computation, *Applied Mathematical Modelling*, vol. 31, no. 9, pp. 1701–1710, 2007.
12. M. Neuts, *Matrix-Geometric Solutions in Stochastic Models*. Baltimore: Johns Hopkins University Press, 1981.
13. G. Latouche and V. Ramaswami, *Introduction to Matrix Analysis Methods in Statistics Modeling*. ASA-SIAM Series on Applied Probability, 1999.

Chapter 5
Modeling of Production System with Nonrenewal Batch Input, Early Setup, and Extra Jobs

Ho Woo Lee, No Ik Park, Se Won Lee, and Jung Woo Baek

Abstract In this chapter, we model and solve a very general single-machine production system with early setup, bilevel threshold control, and extra job operations. The first threshold is used to control the setup starting time and the second threshold is used to control the production starting time. The system is modeled by the BMAP/G/1 queue and the manufacturing lead time is analyzed. The factorization principle is used to derive the distribution of the manufacturing lead time and the mean value. A numerical example is provided.

5.1 Introduction

Industrial engineers have long been interested in analyzing the trade-offs between the system setup and work-in-process (WIP) inventory in order to provide the conditions under which the system operates most economically in the long run. Usually the system setup increases the work-in-process inventory which results in a higher holding cost. But when the system setup cost is very high, this increased holding cost may offset the setup cost because the setup increases the manufacturing cycle

H.W. Lee
Department of Systems Management Engineering, Sungkyunkwan University, Suwon, Korea
e-mail: hwlee@skku.edu

N.I. Park
BcN BcN Research Division, ETRI, Daejon, Korea
e-mail: nipark@etri.re.kr

S.W. Lee
Department of Industrial Engineering, Graduate School, Sungkyunkwan University, Suwon, Korea
e-mail: swlee94@skku.edu

J.W. Baek
Department of Systems Management Engineering, Sungkyunkwan University, Suwon, Korea
e-mail: rainbeak@skku.edu

W. Yue et al. (eds.), *Advances in Queueing Theory and Network Applications,*

time which will result in reduced long-run operating cost per unit time. Hence the system setup and WIP inventory are the two most important factors in the cost-effective operation of a production system. Queueing models have played important roles in their analytical efforts along this line.

In most studies on production systems, it has been assumed that the feed process into the production system follows the Poisson process, mainly due to its analytical tractability. But in many real production settings, the interarrival times of the raw materials are correlated, and independently identically distributed (i.i.d.) exponential interarrival times are rarely found. Also, in many production systems, setup operations take several days and are very costly. One way to reduce the setup cost per unit time is to delay the production until some number of raw materials accumulates and this is the well-known N-policy in a queueing context. The N-policy results in a longer cycle length which means fewer cycles per unit time. But at the same time, the average WIP inventory level becomes larger. Thus, in real production settings, the N-policy is used to reduce the overall average cost per unit time when the setup cost is extremely high compared to the WIP holding cost.

In this chapter, we model and solve a very general single-machine production system with early setup, bilevel threshold control, and extra job operations. The first threshold is used to control the setup starting time and the second threshold is used to control the production starting time. The system can be modeled by the BMAP/G/1 queue with bilevel thresholds, setup time, and multiple vacations. We are especially interested in the manufacturing lead time (MLT), which is defined as the time from the arrival of an order till the time the ordered production is finished. The MLT is an important measure of the performance of the production system because whether the manufacturer can meet the due date of an order is one of the most important success indicators of the production system.

Because the MLT corresponds to the system sojourn time (waiting time + processing time) of a queueing system, our objective is to derive the waiting time distribution of the BMAP/G/1 queueing system under the above-mentioned mixed control policy. The idea and basic methods that are employed in this chapter can be applied to many exhaustive BMAP/G/1 systems with more variability.

The N-policy system was first studied by Yadin and Naor [1]. For other works on N-policy queues, see Hersh and Brosh [2], Hofri [3], Kella [4], Lee and Srinivasan [5], Takagi [6], Lee, and chae [7], and Lee and Ahn [8], to list a few.

Lee and Park [9] showed that the double threshold (α, N)-policy is better than the single threshold N-policy when the setup cost is extremely high compared to the WIP holding cost. We note Lee, Park, and Jeon [10] applied the factorization property of the queue length to the analysis of the WIP inventory of a production system with maintenance, setup, and thresholds.

The chapter is organized as follows. In Sects. 5.2 and 5.3, the system model is described and some notation definitions are given. In Sects. 5.4 and 5.5, the waiting time distribution and the mean waiting time are derived. Numerical examples are shown in Sect. 5.6 and conclusions are drawn in Sect. 5.7.

5.2 System Model

Our queueing system operates as follows (see Fig. 5.1). As soon as the system empties, the server leaves for a vacation of random length V with distribution function (DF) $V(x)$ and the Laplace–Stieltjes transform (LST) $V^*(\theta)$ (the server attends to extra jobs during the vacation). After it returns from the vacation, if it finds α or more customers, it immediately starts a setup of random length H with DF $H(x)$ and the LST $H^*(\theta)$. Otherwise, it takes repeated i.i.d. vacations until it finds α or more customers to start a setup. After the setup is finished, if the total number of customers in the system (queue length) is greater than or equal to N, the server immediately begins to serve the customers. If not, the server waits in the system until the queue length reaches or exceeds N.

In our system, customers arrive according to a BMAP (Batch Markovian Arrival Process) with parameter matrices $(D_0, D_1, D_2, \ldots)$ with $D(z) = \sum_{n=0}^{\infty} D_n z^n$ as the matrix generating function (GF) where $D = D(1) = \sum_{n=0}^{\infty} D_n$ is the infinitesimal generator of the underlying Markov chain (UMC). We assume that the service times are i.i.d. random variables with DF $S(x)$ and the LST $S^*(\theta)$. We also assume that the service times, the vacation times, the setup time, and the arrival process are independent of each other.

An excellent treatment of the BMAP and BMAP/G/1 queues can be found in Lucantoni [11], [12]. For computational algorithms concerning BMAP queues, see Lucantoni [11], [12], Ramaswami [13], and Latouche and Ramaswami [14].

Chang, Takine, and Chae et al. [15] studied the factorization property for a BMAP/G/1 queue with generalized vacations. Lee, Park, and Jeon [16] applied the factorization property to the Park, and Jeon BMAP/G/1 queue with early setup and bilevel threshold policy.

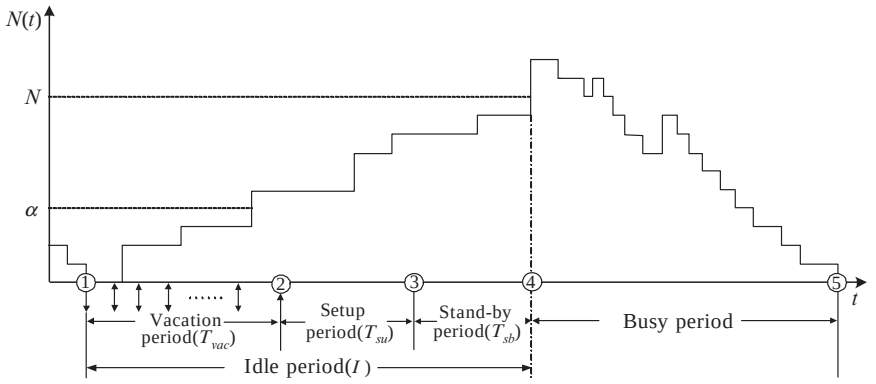


Fig. 5.1 The system.

5.3 Preliminaries

Let π be the stationary vector of the UMC. Then, π can be obtained from

$$\pi D = 0, \quad \pi e = 1,$$

where e is the column vector of 1s with appropriate dimension.

Let λ_g be the group arrival rate. Then, we have

$$\lambda_g = \pi \sum_{n=1}^{\infty} D_n e = \pi(D - D_0)e = -\pi D_0 e.$$

The total customer arrival rate λ becomes

$$\lambda = \pi \sum_{n=1}^{\infty} n D_n e.$$

Let Γ be the size of an arbitrary arrival group with $\gamma_k = Pr(\Gamma = k)$. Then, we have

$$\gamma_k = \frac{\pi D_k e}{\pi \sum_{n=1}^{\infty} D_n e} = \frac{\pi D_k e}{\lambda_g} \tag{5.1}$$

and

$$\mathrm{E}[\Gamma] = \lambda / \lambda_g. \tag{5.2}$$

Let δ_k be the probability that the test customer belongs to a group of size k. From the theory of discrete-time renewal theory, we have, after using (5.1) and (5.2),

$$\delta_k = \frac{k \cdot \gamma_k}{\mathrm{E}[\Gamma]} = \frac{k \pi D_k e}{\lambda}.$$

Now, let us consider a "virtual customer" who arrives at an arbitrary point of time during the busy period and sees the system state (n, i) where n is the queue length (i.e., the number of customers including the one in service) and i is the phase of the UMC at the arrival instance. Let the time-average probability of this state be $y_{busy,n,i}$ with vector $y_{busy,n} = (y_{busy,n,1}, \ldots, y_{busy,n,m})$ and the vector GF $Y_{busy}(z) = \sum_{n=1}^{\infty} y_{busy,n} z^n$. Now, let us consider an arbitrary "actual customer" who arrives during the busy period. If he belongs to a group of size k (with probability δ_k), and is ith within his group (with probability $1/k$), he has $(i-1)$ customers preceding him in his group. Thus, the vector GF $Y^+_{busy}(z)$ of the number of customers just after his arrival becomes

$$Y^+_{busy}(z) = \sum_{k=1}^{\infty} \sum_{i=1}^{k} \delta_k \frac{1}{k} z^{i-1} Y_{busy}(z) \frac{D_k}{\pi D_k e} = Y_{busy}(z) \frac{D - D(z)}{\lambda(1-z)}, \tag{5.3}$$

where $D_k / \pi D_k e$ is multiplied to convert the virtual joint probability of the queue length and the UMC phase to the actual joint probability (note that our test customer belongs to a group of size k). Equation (5.3) was already stated in Lucantoni [11], [12].

5.4 Waiting Time Distribution

In order to obtain the vector Laplace–Stieltjes transform (LST) $w_A^*(\theta)$ of the waiting time of an actual test customer, the first step is to find the vector LSTs $w_{vac,V}^*(\theta)$, $w_{su,V}^*(\theta)$, $w_{sb,V}^*(\theta)$, and $w_{busy,V}^*(\theta)$ of the waiting time of the virtual customer who arrives at an arbitrary time in each period. Once we obtain these quantities, we can obtain the vector LSTs $w_{vac,A}^*(\theta)$, $w_{su,A}^*(\theta)$, $w_{sb,A}^*(\theta)$, and $w_{busy,A}^*(\theta)$ of the waiting time of an actual test customer by postmultiplying appropriate quantities to convert the virtual probabilities to actual probabilities.

To obtain $w_{busy,V}^*(\theta)$, we need $Y_{busy}^*(z,\theta)$ which is the joint transform of the queue length and the remaining service time at the arrival instance of the virtual customer. Then we get

$$w_{busy,V}^*(\theta) = \left[\frac{Y_{busy}^*(z,\theta)}{z}\right]_{z=S^*(\theta)} = \frac{Y_{busy}^*[S^*(\theta),\theta]}{S^*(\theta)}.$$

Then, in the analogous manner as in (5.3), we get

$$w_{busy,A}^*(\theta) = \frac{Y_{busy}^*[S^*(\theta),\theta]}{S^*(\theta)} \frac{D - D(S^*(\theta))}{\lambda(1 - S^*(\theta))}. \tag{5.4}$$

Now, if we let $Y_{idle}(z)$ be the vector GF of the queue length at an arbitrary idle time in a BMAP/G/1 queue with generalized vacations, it is proven by Chang et al. [15] that $Y_{busy}^*(z,\theta)$ is given by

$$Y_{busy}^*(z,\theta)[\theta I + D(z)] = (1-\rho)Y_{idle}(z)zD(z)[A(z) - S^*(\theta)I][zI - A(z)], \tag{5.5}$$

where $\rho = \lambda \mathrm{E}[S]$ is the server utilization and $A(z)$ is the matrix GF of the number of customers that arrive during the service time which is given by $A(z) = \int_0^\infty e^{D(z)x} dS(x)$ (Lucantoni [12]). Thus, our temporary objective is to obtain $Y_{idle}(z)$.

5.4.1 Obtaining $Y_{idle}(z)$

In this subsection, we derive the vector GF $Y_{idle}(z)$ of the queue length at an arbitrary idle time. To this end, we first find p_{vac}, p_{su}, and p_{sb} which are time-average probabilities that the system is in a vacation period, in a setup period, and in a standby period, respectively, under the condition that the system is idle (see Fig. 5.1). Let $\mathrm{E}[T_{vac}]$, $\mathrm{E}[H]$, and $\mathrm{E}[T_{sb}]$ be the mean length of each period. Then, we get

$$\mathrm{E}[I] = \mathrm{E}[T_{vac}] + \mathrm{E}[H] + \mathrm{E}[T_{sb}]$$

and

$$p_{vac} = \frac{\mathrm{E}[T_{vac}]}{\mathrm{E}[I]}, \quad p_{su} = \frac{\mathrm{E}[H]}{\mathrm{E}[I]}, \quad p_{sb} = \frac{\mathrm{E}[T_{sb}]}{\mathrm{E}[I]}. \tag{5.6}$$

In the sequel, we denote $(F)_{ij}$ as the (i,j)-element of a matrix F.

We first derive $\mathrm{E}[T_{vac}]$. Let us define a grand vacation process as in Lee et al. [16]. A grand vacation (GV) is the sum of i.i.d. individual vacations until there is a change in queue length upon a return from a vacation. The first grand vacation (GV) G_1 starts from ① (see Fig. 5.1) and lasts until the queue length differs from 0 upon a return from a vacation. At this point, if the queue length is less than α, the second GV G_2 starts and lasts until there is a change in the queue length upon a return from a vacation. The GV process continues in this manner until the queue length upon return from a vacation is greater than or equal to α.

We note that a GV is equivalent to the vacation period in the simple BMAP/G/1 queue with multiple vacations. Let $(R_n)_{ij}$ be the probability that the GV process visits level (queue length) n and the UMC phase is j just after the visit given that the UMC phase is i at ①. It was proven in Lee et al. [16] that R_n can be computed from the following recursion,

$$R_0 = I, \quad R_n = \sum_{i=1}^{n} R_{n-i}(I - V_0)^{-1} V_i, \quad (n \geq 1),$$

where V_i is the matrix probability that i customers arrive during a vacation.

Because $[(I - V_0)^{-1}]_{ij}$ is the mean number of vacations (within a GV) that starts with phase j under the condition that the GV started with phase i, we have

$$\mathrm{E}[T_{vac}] = \left[\kappa \sum_{n=0}^{\alpha-1} R_n (I - V_0)^{-1} e \right] \mathrm{E}[V], \tag{5.7}$$

where κ is the phase probability vector at ①. Obtaining κ is discussed later.

To derive $\mathrm{E}[T_{sb}]$, let us define $(\Phi_k^{sb})_{ij}, (\alpha \leq k \leq N-1)$ as follows:

$(\Phi_k^{sb})_{ij}$ = *Pr* (the stand-by process visits level k and the phase of UMC is j just after the visit | UMC phase is i at ①).

Noting that (i,j)-element of the matrix $(-D_0)^{-1}$ is the mean time the UMC stays in phase j until the next arrival given that the current phase is in i (see, e.g., Latouche and Ramaswami [14]), we have

$$\mathrm{E}[T_{sb}] = \kappa \sum_{k=\alpha}^{N-1} \Phi_k^{sb} (-D_0)^{-1} e. \tag{5.8}$$

Thus, the mean length of an arbitrary idle period is given by

$$\mathrm{E}[I] = \kappa \left[\sum_{n=0}^{\alpha-1} R_n (I - V_0)^{-1} \mathrm{E}[V] + \mathrm{E}[H] I + \sum_{k=\alpha}^{N-1} \Phi_k^{sb} (-D_0)^{-1} \right] e. \tag{5.9}$$

Then p_{vac}, p_{su}, and p_{sb} can be obtained from (5.6)–(5.9).

Computation of κ and $\{\Phi_k^{sb}, (\alpha \leq k \leq N-1)\}$ is discussed later.

Let $p_{vac}(z)$, $p_{su}(z)$, and $p_{sb}(z)$ be the vector GFs of the queue length at an arbitrary epoch in each period under the condition that the system is idle. We first obtain

$p_{vac}(z)$. Consider an arbitrary time point t^* during the vacation period. At the start of the vacation that contains t^*, the queue length is n and the UMC phase is j with probability

$$\frac{[\kappa R_n(I-V_0)^{-1}]_j}{\kappa\sum_{n=0}^{\alpha-1} R_n(I-V_0)^{-1}e},$$

where the denominator is the mean number of individual vacations during the vacation period. Now, the matrix GF $V^*(z)$ of the number of customers that arrive during the elapsed vacation is given by

$$V^*(z)=\int_0^\infty e^{D(z)x}\left[\frac{1-V(x)}{\mathrm{E}[V]}\right]dx=\frac{[V(z)-I]}{\mathrm{E}[V]}D(z)^{-1},$$

where $V(z)$ is the GF of $\{V_i\}$. Thus, we get

$$p_{vac}(z)=p_{vac}\frac{\kappa\sum_{n=0}^{\alpha-1} R_n[I-V_0]^{-1}z^n}{\kappa\sum_{n=0}^{\alpha-1} R_n[I-V_0]^{-1}e}\frac{[V(z)-I]}{\mathrm{E}[V]}D(z)^{-1}. \tag{5.10}$$

Now, to derive $p_{su}(z)$, let us define $H_\alpha^-(z)=\sum_{k=\alpha}^\infty H_{k(\alpha)}^- z^k$ as the GF of the matrix probability $H_{k(\alpha)}^-$ that there are k customers at the start of the setup period (point ②). Noticing that $H_\alpha^-(z)$ is equivalent to the queue length GF at the start of the busy period in the simple BMAP/G/1 queue with α-policy and multiple vacation, we have from Lee et al. [16],

$$H_\alpha^-(z)=I+\sum_{j=0}^{\alpha-1} R_j[I-V_0]^{-1}z^j[V(z)-I]. \tag{5.11}$$

Then, we get

$$p_{su}(z)=p_{su}\cdot\kappa H_\alpha^-(z)H^*(z), \tag{5.12}$$

where

$$H^*(z)=\frac{[H(z)-I]}{\mathrm{E}[H]}D(z)^{-1}$$

is the GF of the number of customers that arrive during the elapsed setup time in which $H(z)$ is the matrix GF of the number of customers that arrive during a setup time.

Under the condition that the system is in a stand-by period, the queue length is k and the UMC phase is j with probability

$$\frac{\left(\kappa \Phi_k^{sb}(-D_0)^{-}1\right)_j}{\kappa \sum_{n=\alpha}^{N-1} \Phi_n^{sb}(-D_0)^{-1} e}.$$

Thus we get

$$p_{sb}(z) = p_{sb} \cdot \frac{\kappa \sum_{k=\alpha}^{N-1} \Phi_k^{sb}(-D_0)^{-1} z^k}{\kappa \sum_{n=\alpha}^{N-1} \Phi_n^{sb}(-D_0)^{-1} e}. \tag{5.13}$$

Combining (5.10), (5.12), and (5.13), we get

$$\begin{aligned} Y_{idle}(z) &= p_{vac}(z) + p_{su}(z) + p_{sb}(z) \\ &= \frac{\kappa}{\mathrm{E}[I]} \left\{ \sum_{n=0}^{\alpha-1} R_n [I - V_0]^{-1} z^n [V(z) - I] D(z)^{-1} \right. \\ &\quad \left. + H_\alpha^-(z)[H(z) - I]D(z)^{-1} + \sum_{n=\alpha}^{N-1} \Phi_n^{sb}(-D_0)^{-1} z^n \right\}. \end{aligned} \tag{5.14}$$

Now, we need to devise a scheme to compute the probability $\Phi_k^{sb}, (\alpha \le k \le N-1)$ that the stand-by process visits level k. This depends on the queue length probability at ③. By conditioning on the queue length at ②, the probability $H_{k(\alpha)}^+$ at the end of the setup period becomes

$$H_{k(\alpha)}^+ = \sum_{i=\alpha}^{k} H_{i(\alpha)}^- H_{k-i} \tag{5.15}$$

and

$$\Phi_k^{sb} = \sum_{i=0}^{k} H_{i(\alpha)}^+ D_{k-i}^*, \quad (\alpha \le k \le N-1),$$

where D_n^* is the probability matrix that the idle period process of the BMAP/G/1/α-policy queueing system (without vacations and setup) visits level n and H_k is the probability that k customers arrive during a setup time. We note, by conditioning on the level visited prior to level n, that we have a recursion,

$$D_0^* = I, \;\; D_n^* = \sum_{l=0}^{n-1} D_l^* (-D_0)^{-1} D_{n-l}.$$

Now, κ can be computed from

$$\kappa K = \kappa, \quad \kappa e = 1,$$

where K is the phase transition probability between ① and ⑤ and can be obtained from

$$K = K(z)|_{z=1},$$

in which $K(z)$ is the matrix GF of the mean number of customers that are served between ① and ⑤. To obtain $K(z)$, we need the GF $Q_{(\alpha,N)}(z)$ of the queue length at the start of the busy period (④). We can show that (see Appendix 3):

$$Q_{(\alpha,N)}(z) = H_{\alpha}^{-}(z)H(z) + \left[\sum_{n=\alpha}^{N-1} \Phi_n^{sb}(-D_0)^{-1}z^n\right] D(z), \tag{5.16}$$

where $H(z)$ is the matrix GF of the number of customers that arrive during the setup time. Using (5.11) in (5.16), we get

$$\begin{aligned} K(z) = Q_{(\alpha,N)}(z)|_{z=G(z)} = &\sum_{n=0}^{\alpha-1} R_n[I-V_0]^{-1}[G(z)]^n[V(G(z))-I]H(G(z)) \\ &+ H(G(z)) + \sum_{n=\alpha}^{N-1} \Phi_n^{sb}(-D_0)^{-1}[G(z)]^n D(G(z)). \end{aligned}$$

Thus we have

$$\begin{aligned} K = K(z)|_{z=1} = &\sum_{n=0}^{\alpha-1} R_n[I-V_0]^{-1}G^n[V(G)-I]H(G) \\ &+ H(G) + \sum_{n=\alpha}^{N-1} \Phi_n^{sb}(-D_0)^{-1}G^n D(G). \end{aligned}$$

Using (5.14) in (5.5), we get

$$\begin{aligned} Y_{busy}^*(z,\theta)[\theta I + D(z)] = &\left\{\sum_{n=0}^{\alpha-1} R_n[I-V_0]^{-1}z^n[V(z)-I]H(z)\right. \\ &\left. + \sum_{n=\alpha}^{N-1} \Phi_n^{sb}(-D_0)^{-1}z^n D(z) + H(z) - I\right\} \\ &\cdot \left\{[z-S^*(\theta)]A(z)[zI-A(z)]^{-1} - S^*(\theta)I\right\}. \end{aligned}$$

Then, we can obtain $w_{busy,A}^*(\theta)$ from (5.4).

5.4.2 *Obtaining the LST of the Waiting Time of the Customer Who Arrives During the Idle Period*

Now to find the vector LST $w_{vac,A}^*(\theta)$ of the waiting time of the actual test customer that arrives during a vacation, we first need to know the number of customers that arrive during the time period from the end of the current vacation to the start of the

setup period because this determines the remaining vacation period and thereby the remaining idle period. For this purpose, let us define the notation as follows:

$T^v_{\alpha-k}$: The remaining time until the setup starts from the end of the current vacation at which there are k customers
$A(T^v_{\alpha-k})$: The number of customers that arrive during $T^v_{\alpha-k}$
J_1: The UMC phase at the end of the current vacation
J_2: The UMC phase at the start of the setup time

Let us define the (i,j)-element of the matrix transform $T^{V*}_{\alpha-k}(\theta,n)$ as follows:

$$\left[T^{V*}_{\alpha-k}(\theta,n)\right]_{ij} = \int_0^\infty e^{-\theta t} Pr(t < T^v_{\alpha-k} \le t+dt,\, A(T^v_{\alpha-k}) = n,\, J_2 = j | J_1 = i).$$

Then, we have

$$T^{V*}_{\alpha-k}(0,n) = H^-_{n(\alpha-k)}, \quad (n \ge \alpha - k).$$

If the test customer who arrives during a vacation belongs to a group of size j and stands ith in her group, she first has to wait that:

(i) The service times of the customers at the start of the current vacation
(ii) The service times of the customers that arrive during the elapsed vacation time
(iii) The time until the end of the current vacation
(iv) The service times of those $(i-1)$ customers who precede her in her group
(v) The remaining vacation period (from the end of the current vacation)
(vi) The time until the busy period starts.

These quantities are dependent on each other. Let us define ψ^V_n as

$$\psi^V_n = \frac{\kappa R_n [I - V_0]^{-1}}{\kappa \sum_{k=0}^{\alpha-1} R_k [I - V_0]^{-1} e},$$

which is the vector probability that the queue length at the start of the current vacation is n. Then the LST of the waiting time above ((ii)–(v)) contribution is as follows:

$$\psi^V_n [S^*(\theta)]^n \Omega^*_V(a,j,b,\theta) [S^*(\theta)]^a [S^*(\theta)]^{i-1},$$

where $\Omega^*_V(a,j,b,\theta)$ is given in (5.34) in Appendix 1 and represents the remaining vacation time including the probability that a customers arrive during the elapsed vacation time; the test customer belongs to a group of size j (the virtual phase is converted to the actual phase at this point. See (5.25) in Appendix 1. See also Kasahara et al. [17]), and b customers arrive during the remaining vacation time.

Now, additional waiting time depends on the situation at the end of the current vacation. Consider the group G^* to which the test customer belongs. Let us define the following quantities:

$Q^-(G^*)$: The number of customers just before G^* arrives
$Q^+(G^*)$: The number of customers just after G^* arrives
Q_V^+: The number of customers at the end of the current vacation
Q_H^-: The number of customers at the start of the setup period
Q_H^+: The number of customers at the end of the setup period

Then we have different cases as follows:

(Case 1) $Q^-(G^*) < \alpha,\ Q^+(G^*) \le \alpha$
(case 1-1) $Q^-(G^*) < \alpha,\ Q^+(G^*) \le \alpha,\ Q_V^+ \le \alpha,\ Q_H^- \le N,\ Q_H^+ \le N,$
(case 1-2) $Q^-(G^*) < \alpha,\ Q^+(G^*) \le \alpha,\ Q_V^+ \le \alpha,\ Q_H^- \le N,\ Q_H^+ > N,$
(case 1-3) $Q^-(G^*) < \alpha,\ Q^+(G^*) \le \alpha,\ Q_V^+ \le \alpha,\ Q_H^- > N,$
(case 1-4) $Q^-(G^*) < \alpha,\ Q^+(G^*) \le \alpha,\ \alpha < Q_V^+ \le N,\ Q_H^+ \le N,$
(case 1-5) $Q^-(G^*) < \alpha,\ Q^+(G^*) \le \alpha,\ \alpha < Q_V^+ \le N,\ Q_H^+ > N,$
(case 1-6) $Q^-(G^*) < \alpha,\ Q^+(G^*) \le \alpha,\ Q_V^+ > N.$

(Case 2) $Q^-(G^*) < \alpha,\ \alpha < Q^+(G^*) \le N$
(case 2-1) $Q^-(G^*) < \alpha,\ \alpha < Q^+(G^*) \le N,\ \ < Q_V^+ \le N,\ Q_H^+ \le N,$
(case 2-2) $Q^-(G^*) < \alpha,\ \alpha < Q^+(G^*) \le N,\ \ < Q_V^+ \le N,\ Q_H^+ > N,$
(case 2-3) $Q^-(G^*) < \alpha,\ \alpha < Q^+(G^*) \le N,\ \ < Q_V^+ > N.$

(Case 3) $Q^-(G^*) < \alpha,\ Q^+(G^*) > N.$

(Case 4) $\alpha < Q^-(G^*) < N$
(case 4-1) $\alpha < Q^-(G^*) < N,\ Q^+(G^*) \le N,\ Q_H^- \le N,\ \ < Q_H^+ \le N,$
(case 4-2) $\alpha < Q^-(G^*) < N,\ Q^+(G^*) \le N,\ Q_H^- \le N,\ \ < Q_H^+ > N,$
(case 4-3) $\alpha < Q^-(G^*) < N,\ Q^+(G^*) \le N,\ Q_H^- > N,$
(case 4-4) $\alpha < Q^-(G^*) < N,\ Q^+(G^*) > N.$

(Case 5) $\alpha < Q^-(G^*) \ge N.$

Now, the waiting times in (case 1-1) and (case 1-2) are as follows:

$$B_1 = \sum_{n=0}^{\alpha-1} \psi_n^V [S^*(\theta)]^n \sum_{a=0}^{\alpha-n-1} \sum_{j=1}^{\alpha-n-a} \sum_{b=0}^{\alpha-n-a-j} \Omega_V^*(a,j,b,\theta)[S^*(\theta)]^a \frac{1}{j} \sum_{i=1}^{j} [S^*(\theta)]^{i-1}$$
$$\cdot \sum_{c=\alpha-n-a-j-b}^{N-n-a-j-b} T^{V*}_{\alpha-n-a-j-b}(\theta,c)$$
$$\cdot \left[\sum_{k=0}^{N-n-a-j-b-c} H_k^*(\theta) T^*_{N-n-a-j-b-c-k}(\theta) + \sum_{k=N-n-a-j-b-c+1}^{\infty} H_k^*(\theta) \right],$$

where $H_k^*(\theta)$ is the matrix LST of the length of the setup time including the probability that k customers arrive during the setup, and $T_n^*(\theta)$ is the matrix LST of the idle period in the single-threshold BMAP/G/1 queue under n-policy (without vacations and setup) which becomes, conditioning on the first group size,

$$T_n^*(\theta) = [\theta I - D_0]^{-1}\left[\sum_{k=1}^{n-1} D_k T_{n-k}^*(\theta) + \sum_{k=n}^{\infty} D_k\right]$$
$$= [\theta I - D_0]^{-1}\left[\sum_{k=1}^{n-1} D_k [T_{n-k}^*(\theta) - I] + D - D_0\right]$$

with $T_0^*(0) = I$. For the remaining cases, we have

(Case 1-3)

$$B_2 = \sum_{n=0}^{\alpha-1} \psi_n^V [S^*(\theta)]^n \sum_{a=0}^{\alpha-n-1} \sum_{j=1}^{\alpha-n-a} \sum_{b=0}^{\alpha-n-a-j} \Omega_V^*(a,j,b,\theta)[S^*(\theta)]^a \frac{1}{j} \sum_{i=1}^{j} [S^*(\theta)]^{i-1}$$
$$\cdot \sum_{c=N-n-a-j-b+1}^{\infty} T_{\alpha-n-a-j-b}^{V*}(\theta,c) H^*(\theta).$$

(Case 1-4) and (Case 1-5)

$$B_3 = \sum_{n=0}^{\alpha-1} \psi_n^V [S^*(\theta)]^n \sum_{a=0}^{\alpha-n-1} \sum_{j=1}^{\alpha-n-a} \sum_{b=\alpha-n-a-j+1}^{N-n-a-j} \Omega_V^*(a,j,b,\theta)$$
$$\cdot [S^*(\theta)]^a \frac{1}{j} \sum_{i=1}^{j} [S^*(\theta)]^{i-1}$$
$$\cdot \left[\sum_{k=0}^{N-n-a-j-b} H_k^*(\theta) T_{N-n-a-j-b-k}^*(\theta) + \sum_{k=N-n-a-j-b+1}^{\infty} H_k^*(\theta)\right].$$

(Case 1-6)

$$B_4 = \sum_{n=0}^{\alpha-1} \psi_n^V [S^*(\theta)]^n \sum_{a=0}^{\alpha-n-1} \sum_{j=1}^{\alpha-n-a} \sum_{b=N-n-a-j+1}^{\infty} \Omega_V^*(a,j,b,\theta)[S^*(\theta)]^a \frac{1}{j}$$
$$\cdot \sum_{i=1}^{j} [S^*(\theta)]^{i-1} H^*(\theta).$$

(Case 2-1) and (Case 2-2)

$$B_5 = \sum_{n=0}^{\alpha-1} \psi_n^V [S^*(\theta)]^n \sum_{a=0}^{\alpha-n-1} \sum_{j=\alpha-n-a+1}^{N-n-a} \sum_{b=0}^{N-n-a-j} \Omega_V^*(a,j,b,\theta)$$
$$\cdot [S^*(\theta)]^a \frac{1}{j} \sum_{i=1}^{j} [S^*(\theta)]^{i-1}$$
$$\cdot \left[\sum_{k=0}^{N-n-a-j-b} H_k^*(\theta) T_{N-n-a-j-b-k}^*(\theta) + \sum_{k=N-n-a-j-b+1}^{\infty} H_k^*(\theta)\right].$$

(Case 2-3)

$$B_6 = \sum_{n=0}^{\alpha-1} \psi_n^V [S^*(\theta)]^n \sum_{a=0}^{\alpha-n-1} \sum_{j=\alpha-n-a+1}^{N-n-a} \sum_{b=N-n-a-j+1}^{\infty} \Omega_V^*(a,j,b,\theta)[S^*(\theta)]^a \frac{1}{j} \cdot \sum_{i=1}^{j} [S^*(\theta)]^{i-1} H^*(\theta).$$

(Case 3)

$$B_7 = \sum_{n=0}^{\alpha-1} \psi_n^V [S^*(\theta)]^n \sum_{a=0}^{\alpha-n-1} \sum_{j=\alpha-n-a+1}^{\infty} \sum_{b=0}^{\infty} \Omega_V^*(a,j,b,\theta)[S^*(\theta)]^a \frac{1}{j} \cdot \sum_{i=1}^{j} [S^*(\theta)]^{i-1} H^*(\theta).$$

(Case 4-1) and (Case 4-2)

$$B_8 = \sum_{n=0}^{\alpha-1} \psi_n^V [S^*(\theta)]^n \sum_{a=\alpha-n}^{N-n-1} \sum_{j=1}^{N-n-a} \sum_{b=0}^{N-n-a-j} \Omega_V^*(a,j,b,\theta)[S^*(\theta)]^a \frac{1}{j} \sum_{i=1}^{j} [S^*(\theta)]^{i-1} \cdot \left[\sum_{k=0}^{N-n-a-j-b} H_k^*(\theta) T^*_{N-n-a-j-b-k}(\theta) + \sum_{k=N-n-a-j-b+1}^{\infty} H_k^*(\theta) \right].$$

(Case 4-3)

$$B_9 = \sum_{n=0}^{\alpha-1} \psi_n^V [S^*(\theta)]^n \sum_{a=\alpha-n}^{N-n-1} \sum_{j=1}^{N-n-a} \sum_{b=N-n-a-j+1}^{\infty} \Omega_V^*(a,j,b,\theta)[S^*(\theta)]^a \frac{1}{j} \cdot \sum_{i=1}^{j} [S^*(\theta)]^{i-1} H^*(\theta).$$

(Case 4-4)

$$B_{10} = \sum_{n=0}^{\alpha-1} \psi_n^V [S^*(\theta)]^n \sum_{a=\alpha-n}^{N-n-1} \sum_{j=N-n-a+1}^{\infty} \sum_{b=0}^{\infty} \Omega_V^*(a,j,b,\theta)[S^*(\theta)]^a \frac{1}{j} \cdot \sum_{i=1}^{j} [S^*(\theta)]^{i-1} H^*(\theta).$$

(Case 5)

$$B_9 = \sum_{n=0}^{\alpha-1} \psi_n^V [S^*(\theta)]^n \sum_{a=N-n}^{\infty} \sum_{j=1}^{\infty} \sum_{b=0}^{\infty} \Omega_V^*(a,j,b,\theta)[S^*(\theta)]^a \frac{1}{j} \sum_{i=1}^{j} [S^*(\theta)]^{i-1} H^*(\theta).$$

Now, combining all these, we get

$$w^*_{vac,A}(\theta) = (1-\rho)p_{vac}\sum_{n=1}^{11} B_n. \tag{5.17}$$

In the similar way, we can obtain the waiting time of the actual customer who arrives during the setup time and we get

$$\begin{aligned} & w^*_{su,A}(\theta) \\ & = (1-\rho)p_{su}\kappa\left[\sum_{n=\alpha}^{N-1} H^-_{n(\alpha)}[S^*(\theta)]^n\left\{\sum_{a=0}^{N-n-1}\sum_{j=1}^{N-n-a}\sum_{b=0}^{N-n-a-j}\Omega^*_V(a,j,b,\theta)\right.\right. \\ & \quad \cdot [S^*(\theta)]^a\frac{1}{j}\sum_{i=1}^{j}[S^*(\theta)]^{i-1}\left[T^*_{N-n-a-j-b}(\theta)-I\right] \\ & \quad \left.\left. + \sum_{a=0}^{\infty}\sum_{j=1}^{\infty}\sum_{b=0}^{\infty}\Omega^*_V(a,j,b,\theta)[S^*(\theta)]^a\frac{1}{j}\sum_{i=1}^{j}[S^*(\theta)]^{i-1}\right\}\right. \\ & \quad \left. + \sum_{n=N}^{\infty} H^-_{n(\alpha)}[S^*(\theta)]^n\sum_{a=0}^{\infty}\sum_{j=1}^{\infty}\sum_{b=0}^{\infty}\Omega^*_V(a,j,b,\theta)[S^*(\theta)]^a\frac{1}{j}\sum_{i=1}^{j}[S^*(\theta)]^{i-1}\right]. \end{aligned} \tag{5.18}$$

For the actual customer who arrives during the standby period, we get

$$\begin{aligned} w^*_{sb,A}(\theta) &= (1-\rho)p_{sb}\sum_{n=\alpha}^{N-1}\psi^{sb}_n[S^*(\theta)]^n \\ &\quad \cdot\left\{\sum_{j=1}^{N-k}\frac{D_j}{\lambda}\sum_{i=1}^{j}[S^*(\theta)]^{i-1}(T^*_{N-k-j}(\theta)-I)+\frac{D-D(S^*(\theta))}{\lambda[1-S^*(\theta)]}\right\}, \end{aligned} \tag{5.19}$$

where

$$\psi^{sb}_k = \frac{\kappa\Phi^s_k b(-D_0)^{-1}}{\kappa\sum_{n=\alpha}^{N-1}\Phi^{sb}_n(-D_0)^{-1}e}$$

is the vector probability that there are k customers under the condition that system is in a standby period.

Finally the LST of the actual waiting customer can be obtained from (5.17)–(5.19), and we get

$$W^*_q(\theta) = w^*_A(\theta)e = w^*_{vac,A}(\theta)e + w^*_{su,A}(\theta)e + w^*_{sb,A}(\theta)e + w^*_{busy,A}(\theta)e.$$

For the simplicity of the subsequent analysis, let us write the LST of the waiting time of an arbitrary actual waiting customer as

$$W^*_q(\theta) = w^*_N(\theta)e + w^*_1(\theta)e, \tag{5.20}$$

where

$$w_N^*(\theta)e = (1-\rho)\left[p_{vac}\sum_{n=0}^{\alpha-1}\psi_n^V[S^*(\theta)]^n\frac{1-V^*(\theta)}{\mathrm{E}[V]\theta}H^*(\theta) + p_{su}\kappa\frac{1-H^*(\theta)}{\mathrm{E}[H]\theta} + p_{sb}\sum_{n=\alpha}^{N-1}\psi_n^{sb}[S^*(\theta)]^n\right] \cdot \theta[\theta I - D(S^*(\theta))]^{-1}\frac{D-D(S^*(\theta))}{\lambda[1-S^*(\theta)]}e$$

and

$$w_1^*(\theta)e = (1-\rho)p_{vac}\sum_{n=0}^{\alpha-1}\psi_n^V[S^*(\theta)]^n\cdot\sum_{k=1}^{5}C_k e + 1-\rho)p_{su}\kappa\sum_{n=\alpha}^{N-1}H^-_{n(\alpha)}[S^*(\theta)]^n\cdot C_6 e + (1-\rho)p_{sb}\sum_{n=\alpha}^{N-1}\psi_n^{sb}[S^*(\theta)]^n\sum_{j=1}^{N-n}\frac{D_j}{\lambda}\sum_{i=1}^{j}[S^*(\theta)]^{i-1} \cdot [T^*_{N-n-j}(\theta)-I]e,$$

where

$$C_1 = \sum_{a=0}^{\alpha-n-1}\sum_{j=1}^{\alpha-n-a}\sum_{b=0}^{\alpha-n-a-j}\Omega_V^*(a,j,b,\theta)[S^*(\theta)]^a\frac{1}{j}\sum_{i=1}^{j}[S^*(\theta)]^{i-1} \cdot \sum_{c=\alpha-n-a-j-b}^{N-n-a-j-b}T^V_{\alpha-n-a-j-b}*(\theta,c)\sum_{k=0}^{N-n-a-j-b-c}H_k^*(\theta) \cdot [T^*_{N-n-a-j-b-c-k}(\theta)-I],$$

$$C_2 = \sum_{a=0}^{\alpha-n-1}\sum_{j=1}^{\alpha-n-a}\sum_{b=0}^{\alpha-n-a-j}\Omega_V^*(a,j,b,\theta)[S^*(\theta)]^a\frac{1}{j}\sum_{i=1}^{j}[S^*(\theta)]^{i-1} \cdot [T^*_{N-n-a-j-b-c-k}(\theta)-I]H^*(\theta),$$

$$C_3 = \sum_{a=0}^{\alpha-n-1}\sum_{j=1}^{\alpha-n-a}\sum_{b=\alpha-n-a-j+1}^{N-n-a-j}\Omega_V^*(a,j,b,\theta)[S^*(\theta)]^a\frac{1}{j}\sum_{i=1}^{j}[S^*(\theta)]^{i-1} \cdot \sum_{k=0}^{N-n-a-j-b}H_k^*(\theta)[T^*_{N-n-a-j-b-c-k}(\theta)-I],$$

$$C_4 = \sum_{a=0}^{\alpha-n-1} \sum_{j=\alpha-n-a+1}^{N-n-a} \sum_{b=0}^{N-n-a-j} \Omega_V^*(a,j,b,\theta)[S^*(\theta)]^a \frac{1}{j} \sum_{i=1}^{j} [S^*(\theta)]^{i-1} \cdot \sum_{k=0}^{N-n-a-j-b} H_k^*(\theta)[T^*_{N-n-a-j-b-c-k}(\theta) - I],$$

$$C_5 = \sum_{a=\alpha-n}^{\alpha-n-1} \sum_{j=1}^{N-n-a} \sum_{b=0}^{N-n-a-j} \Omega_V^*(a,j,b,\theta)[S^*(\theta)]^a \frac{1}{j} \sum_{i=1}^{j} [S^*(\theta)]^{i-1} \cdot \sum_{k=0}^{N-n-a-j-b} H_k^*(\theta)[T^*_{N-n-a-j-b-c-k}(\theta) - I],$$

$$C_6 = \sum_{a=0}^{N-n-1} \sum_{j=1}^{N-n-a} \sum_{b=0}^{N-n-a-j} \Omega_V^*(a,j,b,\theta)[S^*(\theta)]^a \frac{1}{j} \sum_{i=1}^{j} [S^*(\theta)]^{i-1} \cdot [T^*_{N-n-a-j-b-c-k}(\theta) - I].$$

5.5 Mean Waiting Time

From (5.20), the mean actual waiting time becomes

$$W_q = -W_q^{*(1)}(0) = -w_N^{*(1)}(0)e - w_1^{*(1)}(0)e,$$

where

$$-w_1^{*(1)}(0)e = (1-\rho)p_{vac} \sum_{n=0}^{\alpha-1} \psi_n^V \sum_{k=1}^{5} E_k + 1 - \rho)p_{su}\kappa \sum_{n=\alpha}^{N-1} H^-_{n(\alpha)} \sum_{a=0}^{N-n-1} \sum_{j=1}^{N-n-a} \sum_{b=0}^{N-n-a-j} \Omega_V(a,j,b) \cdot \tau_{N-n-a-j-b} + (1-\rho)p_{sb} \sum_{n=\alpha}^{N-1} \psi_n^{sb} \sum_{j=1}^{N-n} \frac{jD_j}{\lambda} \tau_{N-n-j},$$

where

$$\Omega_V(a,j,b) = \Omega_V^*(a,j,b,\theta)|_{\theta=0},$$

$$\tau_n = -\frac{d}{d\theta} T_n^*(\theta)\Big|_{\theta=0} e = \sum_{k=0}^{n-1} D_k^*(-D_0)^{-1} e,$$

$$E_1 = \sum_{a=0}^{\alpha-n-1} \sum_{j=1}^{\alpha-n-a} \sum_{b=0}^{\alpha-n-a-j} \Omega_V(a,j,b) \sum_{c=\alpha-n-a-j-b}^{N-n-a-j-b} H^{-}_{c(\alpha-n-a-j-b)} \cdot \sum_{k=0}^{N-n-a-j-b-c} H_k \tau_{N-n-a-j-b-c-k},$$

$$E_2 = \sum_{a=0}^{\alpha-n-1} \sum_{j=1}^{\alpha-n-a} \sum_{b=0}^{\alpha-n-a-j} \Omega_V(a,j,b) \tau^V_{\alpha-n-a-j-b},$$

$$E_3 = \sum_{a=0}^{\alpha-n-1} \sum_{j=1}^{\alpha-n-a} \sum_{b=\alpha-n-a-j+1}^{N-n-a-j} \Omega_V(a,j,b) \sum_{k=0}^{N-n-a-j-b-c} H_k \tau_{N-n-a-j-b-c-k},$$

$$E_4 = \sum_{a=0}^{\alpha-n-1} \sum_{j=\alpha-n-a+1}^{N-n-a} \sum_{b=0}^{N-n-a-j} \Omega_V(a,j,b) \sum_{k=0}^{N-n-a-j-b-c} H_k \tau_{N-n-a-j-b-c-k},$$

$$E_5 = \sum_{a=\alpha-n}^{N-n-1} \sum_{j=1}^{N-n-a} \sum_{b=0}^{N-n-a-j} \Omega_V(a,j,b) \sum_{k=0}^{N-n-a-j-b-c} H_k \tau_{N-n-a-j-b-c-k}$$

and

$$H_k = H_k^*(\theta)\big|_{\theta=0}.$$

Now we need to determine $w_N^{*(1)}(0)e$. Let us rewrite $w_N^*(\theta)e$ as

$$w_N^*(\theta)e = z^*(\theta)\frac{D - D(S^*(\theta))}{\lambda[1 - S^*(\theta)]}e, \tag{5.21}$$

where

$$\begin{aligned} z^*(\theta) = (1-\rho)\Bigg[& p_{vac} \sum_{n=0}^{\alpha-1} \psi_n^V [S^*(\theta)]^n \frac{1 - V^*(\theta)}{\mathrm{E}[V]\theta} H^*(\theta) \\ & + p_{su}\kappa \frac{1 - H^*(\theta)}{\mathrm{E}[H]\theta} + p_{sb} \sum_{n=\alpha}^{N-1} \psi_n^{sb} [S^*(\theta)]^n \Bigg] \\ & \cdot \theta[\theta I - D(S^*(\theta))]^{-1}. \end{aligned} \tag{5.22}$$

Taking the derivative of (5.21) with respect to θ we get

$$-w_N^{*(1)}(\theta)\Big|_{\theta=0} e = -\frac{z^{*(1)}(0)D^{(1)}e}{\lambda} + \frac{z^*(0)\mathrm{E}[S]D^{(2)}e}{2\lambda}, \tag{5.23}$$

where

$$D^{(n)} = \frac{d^n}{dz^n} D^{(n)}(z)\Big|_{z=1}.$$

The derivation of (5.23) is given in Appendix 2. Now we can show that

$$z^*(0) = \pi. \tag{5.24}$$

Then from (5.23) and (5.24), the mean actual waiting time becomes

$$\begin{aligned}
W_q = {} & (1-\rho)p_{vac}\sum_{n=0}^{\alpha-1}\psi_n^V\sum_{k=1}^{5}E_k \\
& +1-\rho)p_{su}\kappa\sum_{n=\alpha}^{N-1}H^-_{n(\alpha)}\sum_{a=0}^{N-n-1}\sum_{j=1}^{N-n-a}\sum_{b=0}^{N-n-a-j}\Omega_V(a,j,b)\tau_{N-n-a-j-b} \\
& +(1-\rho)p_{sb}\sum_{n=\alpha}^{N-1}\psi_n^{sb}\sum_{j=1}^{N-n}\frac{jD_j}{\lambda}\tau_{N-n-j} \\
& -\frac{1}{\lambda}\left[p_{vac}\sum_{n=0}^{\alpha-1}\psi_n^V+p_{su}\kappa+p_{sb}\sum_{n=\alpha}^{N-1}\psi_n^{sb}\right](D+e\pi)^{-1}D^{(1)}e \\
& +p_{vac}\sum_{n=0}^{\alpha-1}n\psi_n^V\mathrm{E}[S]e+p_{vac}\mathrm{E}[H]+p_{vac}\frac{\mathrm{E}[V^2]}{2\mathrm{E}[V]} \\
& +p_{sb}\sum_{n=\alpha}^{N-1}n\psi_n^{sb}\mathrm{E}[S]e+p_{su}\frac{\mathrm{E}[H^2]}{2\mathrm{E}[H]}+\frac{\lambda\mathrm{E}[S^2]}{2(1-\rho)}+\frac{\pi\mathrm{E}^2[S]D^{(2)}e}{2\rho(1-\rho)} \\
& +\frac{1}{1-\rho}-\frac{\pi\mathrm{E}[S]D^{(1)}(D+e\pi)^{-1}D^{(1)}e}{\lambda(1-\rho)}.
\end{aligned}$$

5.6 Numerical Example

In this section, we present a numerical example. We consider the parameter matrices as follows:

$$D_0 = \begin{pmatrix} -2.05 & 0.1 & 0.45 \\ 0.4 & -2.65 & 1.05 \\ 0.25 & 0.1 & -1.85 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0.2 & 0.4 & 0.2 \\ 0.5 & 0.3 & 0.2 \\ 0.3 & 0.6 & 0.1 \end{pmatrix},$$

$$D_2 = \begin{pmatrix} 0.15 & 0.1 & 0.15 \\ 0.1 & 0.1 & 0 \\ 0.05 & 0.1 & 0.05 \end{pmatrix}, \quad D_3 = \begin{pmatrix} .0.1 & 0 & 0.2 \\ 0.5 & 0.3 & 0.2 \\ 0.3 & 0.6 & 0.1 \end{pmatrix}.$$

Then, we get

$$D = \sum_{j=0}^{3} D_j = \begin{pmatrix} 0.15 & 0.1 & 0.15 \\ 0.1 & 0.1 & 0 \\ 0.05 & 0.1 & 0.05 \end{pmatrix}.$$

Table 5.1 Comparison of mean performance measures with simulation.

(α, N)	ρ	Measure	Theoretical Value	Simulation	RPE
(3,5)	0.4311	L	5.0550	5.0517	0.065
(3,5)	0.4311	W_q	2.1456	2.1461	−0.023
(3,5)	0.8622	L	11.6486	11.6616	−0.112
(3,5)	0.8622	W_q	5.0047	5.0116	−0.138
(3,7)	0.4311	L	5.2116	5.2051	0.125
(3,7)	0.4311	W_q	2.2182	2.2166	0.298
(3,7)	0.8622	L	11.8049	11.8050	0.000
(3,7)	0.8622	W_q	5.0771	5.0812	−0.081

From $\pi D = 0$, $\pi e = 1$, and $\lambda = \pi \sum_{n=1}^{\infty} nD_n e$, we get

$$\pi = (0.35326, 0.23913, 0.40761), \quad \lambda = 2.1554348.$$

We consider two cases of thresholds: $(\alpha, N) = (3,5)$ and $(\alpha, N) = (3,7)$. For both cases we assume that the setup time and the vacation time follow the exponential distribution with mean 1.0. For each case, we assume two Erlang service times of order 2 with different mean service times: $\mathrm{E}[S] = 0.2$ and $\mathrm{E}[S] = 0.4$. Table 5.1 shows the comparison of the mean waiting times and the mean queue lengths that can be obtained from Little's law $L = \lambda\{W_q + \mathrm{E}[S]\}$ with those obtained from simulation estimates. The relative percentage error (RPE) is defined by

$$\frac{\text{Theoretical value} - \text{Simulation estimate}}{\text{Theoretical value}}.$$

5.7 Conclusions and Summary

In this chapter, we applied the BMAP/G/1 queue with early setup and multiple vacation to the analysis of the manufacturing lead time of a production system with extra jobs and bilevel threshold control. We employed the factorization principle to derive the distribution of the manufacturing lead time and the mean value.

Acknowledgments This work was supported by grant No. R01-2006-000-10906-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

Appendix 1

We define the joint matrix transform by

$$\Omega_V^*(z_1,j,z_2,\theta)=\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}\int_0^{\infty}z_1^a z_2^b e^{-\theta y}d\Omega_V(a,j,b,y).$$

Then, we have

$$\Omega_V^*(z_1,j,z_2,\theta)=\int_0^{\infty}e^{-\theta y}\int_0^{x}e^{D(z_1)(x-y)}\frac{j\pi D_j\mathrm{e}}{\lambda}\frac{D_j}{\pi D_j\mathrm{e}}e^{D(z_2)y}\frac{x\cdot dV(x)}{\mathrm{E}[V]}\frac{1}{x}dy, \quad (5.25)$$

which is equivalent to

$$\Omega_V^*(z_1,j,z_2,\theta)=\int_0^{\infty}\int_0^{x}e^{D(z_1)t}\frac{jD_j}{\lambda}e^{D(z_2)(x-t)}e^{-\theta(x-t)}\frac{dV(x)}{\mathrm{E}[V]}dt.$$

Our temporary objective is to obtain the coefficient matrix $\Omega_V^*(a,j,b,\theta)$ of $\Omega_V^*(z_1,j,z_2,\theta)$ such that

$$\Omega_V^*(z_1,j,z_2,\theta)=\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}z_1^a z_2^b\Omega_V^*(a,j,b,\theta). \quad (5.26)$$

To this end, we apply the well-known uniformization technique. Let us define Θ such that $\Theta=\max_i(-D_0)_{ii}$. First, $e^{D(z_1)t}$ and $e^{D(z_2)(x-t)}$ can be written as

$$e^{D(z_1)t}=e^{-\Theta t}e^{\Theta(I+\Theta^{-1}D(z_1))t}=\sum_{k=0}^{\infty}\frac{e^{-\Theta t}(\Theta t)^k}{k!}(I+\Theta^{-1}D(z_1))^k \quad (5.27)$$

and

$$\begin{aligned}e^{D(z_2)(x-t)}&=e^{-\Theta(x-t)}e^{\Theta(I+\Theta^{-1}D(z_2))(x-t)}\\&=\sum_{k=0}^{\infty}\frac{e^{-\Theta(x-t)}[\Theta(x-t)]^k}{k!}(I+\Theta^{-1}D(z_2))^k. \quad (5.28)\end{aligned}$$

Using (5.27) and (5.28) in (5.26) yields

$$\begin{aligned}\Omega_V^*(z_1,j,z_2,\theta)=\int_0^{\infty}e^{-(\Theta+\theta)x}\int_0^{x}e^{\theta t}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\frac{t^k(x-t)^l}{k!l!}(\Theta I+D(z_1))^k\frac{jD_j}{\lambda}\\\cdot(\Theta I+D(z_2))^l\frac{dV(x)}{\mathrm{E}[V]}dt. \quad (5.29)\end{aligned}$$

In (5.29), only $(\Theta I+D(z_1))^k(jD_j/\lambda)(\Theta I+D(z_2))^l$ contains z_1 and z_2. To evaluate this matrix, we define $F_{k,l}(a,j,b),\ (k,l,a,b=0,1,\ldots;j=1,2,\ldots)$ such that

$$\sum_{a=0}^{\infty}\sum_{b=0}^{\infty} z_1^a z_2^b F_{k,l}(a,j,b) = (\Theta I + D(z_1))^k \frac{jD_j}{\lambda}(\Theta I + D(z_2))^l, \tag{5.30}$$

where $F_{0,0}(0,j,0) = jD_j/\lambda$, and $F_{0,0}(a,j,b) = 0,\ (a \geq 1, b \geq 1)$. $F_{k,l}(a,j,b)$ represents the situation in which a jobs arrive from k Poisson events (with rate Θ) during the elapsed vacation time and b jobs arrive from l Poisson events during the remaining vacation time. Then, $F_{k,l}(a,j,b)$ satisfies the following recursions:

$$F_{k+1,l}(a,j,b) = \begin{cases} (\Theta I + D_0)F_{k,l}(a,j,b), & (a=0) \\ \displaystyle\sum_{i=0}^{a-1} D_{a-i}F_{k,l}(a,j,b) + (\Theta I + D_0)F_{k,l}(a,j,b), & (a \geq 1), \end{cases} \tag{5.31}$$

$$F_{k,l+1}(a,j,b) = \begin{cases} F_{k,l}(a,j,b)(\Theta I + D_0), & (b=0) \\ \displaystyle\sum_{i=0}^{b-1} F_{k,l}(a,j,i)D_{b-i} + F_{k,l}(a,j,b)(\Theta I + D_0), & (b \geq 1). \end{cases} \tag{5.32}$$

Using (5.31) and (5.32) in (5.30), we get

$$\begin{aligned} &\Omega_V^*(z_1,j,z_2,\theta) \\ &\quad = \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} z_1^a z_2^b \int_0^{\infty} e^{-(\Theta+\theta)x} \int_0^x e^{\theta t} \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} \frac{t^k (x-t)^l}{k!l!} F_{k,l}(a,j,b) \frac{dV(x)}{\mathrm{E}[V]} dt. \end{aligned} \tag{5.33}$$

The coefficient matrix of $z_1^a z_2^b$ in (5.33) is given by

$$\Omega_V^*(a,j,b,\theta) = \int_0^{\infty} e^{-(\Theta+\theta)x} \int_0^x e^{\theta t} \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} \frac{t^k (x-t)^l}{k!l!} F_{k,l}(a,j,b) \frac{dV(x)}{\mathrm{E}[V]} dt. \tag{5.34}$$

If we disregard the length of the remaining vacation time, we have

$$\Omega_V(a,j,b) = \Omega_V^*(a,j,b,\theta)\big|_{\theta=0} = \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} f_{k,l} F_{k,l}(a,j,b),$$

where

$$f_{k,l} = \frac{1}{\mathrm{E}[V](k+l+1)!} \int_0^{\infty} x^{k+l+1} e^{-\Theta x} dV(x).$$

Appendix 2: Derivation of (5.23)

Taking a derivative of (5.22) with respect to θ, using $\theta = 0$ and adding $z^*(\theta)e\pi$ to both sides yields

$$z^{*(1)}(0) = z^{*(1)}(0)e\pi(D+e\pi)^{-1} - z^*(0)[I - \mathrm{E}[S]D^{(1)}](D+e\pi)^{-1} + (1-\rho)\left[p_v \sum_{n=0}^{\alpha-1} \psi_n^V + p_{su}\kappa + p_{sb} \sum_{n=\alpha}^{N-1} \psi_n^{sb}\right](D+e\pi)^{-1}. \tag{5.35}$$

Taking the second derivative of (5.22), using $\theta = 0$ and postmultiplying e yields

$$\begin{aligned} & z^{*(1)}(0)[I - \mathrm{E}[S]D^{(1)}]e \\ &= -\ (1-\rho)\left\{p_{vac} \sum_{n=0}^{\alpha-1} n\psi_n^V \mathrm{E}[S] + p_{vac} \sum_{n=0}^{\alpha-1} \psi_n^V \mathrm{E}[H] \right. \\ &\quad \left. +\ p_{vac} \sum_{n=0}^{\alpha-1} \psi_n^V \frac{\mathrm{E}[V^2]}{2\mathrm{E}[V]} + p_{su}\kappa \frac{\mathrm{E}[H^2]}{2\mathrm{E}[H]} + p_{sb} \sum_{n=\alpha}^{N-1} n\psi_n^{sb} \mathrm{E}[S]\right\} e \\ &\quad -\ \frac{\pi}{2}\left[\mathrm{E}[S^2]D^{(1)} + \mathrm{E}^2[S]D^{(2)}\right]e. \end{aligned} \tag{5.36}$$

From (5.36), we get

$$\begin{aligned} z^{*(1)}(0)e &= z^{*(1)}(0)\mathrm{E}[S]D^{(1)}e \\ &= -(1-\rho)\left\{p_{vac} \sum_{n=0}^{\alpha-1} n\psi_n^V \mathrm{E}[S] + p_{vac} \sum_{n=0}^{\alpha-1} \psi_n^V \mathrm{E}[H] \right. \\ &\quad \left. +\ p_{vac} \sum_{n=0}^{\alpha-1} \psi_n^V \frac{\mathrm{E}[V^2]}{2\mathrm{E}[V]} + p_{su}\kappa \frac{\mathrm{E}[H^2]}{2\mathrm{E}[H]} + p_{sb} \sum_{n=\alpha}^{N-1} n\psi_n^{sb} \mathrm{E}[S]\right\} e \\ &\quad -\ \frac{\pi}{2}\left[\mathrm{E}[S^2]D^{(1)} + \mathrm{E}^2[S]D^{(2)}\right]e. \end{aligned} \tag{5.37}$$

Postmultiplying both sides of (5.35) by $D^{(1)}e$, we get

$$z^{*(1)}(0)D^{(1)}e = \lambda z^{*(1)}(0)e - z^*(0)[I - \mathrm{E}[S]D^{(1)}](D+e\pi)^{-1}D^{(1)}e + (1-\rho)\left[p_{vac} \sum_{n=0}^{\alpha-1} \psi_n^V + p_{su}\kappa + p_{sb} \sum_{n=\alpha}^{N-1} \psi_n^{sb}\right](D+e\pi)^{-1}D^{(1)}e. \tag{5.38}$$

Using (5.37) in (5.38), we get

$$\begin{aligned}\frac{z^{*(1)}(0)D^{(1)}e}{\lambda} &= -p_{vac}\sum_{n=0}^{\alpha-1} n\psi_n^V \mathrm{E}[S]e + p_{vac}\mathrm{E}[H] + p_{vac}\frac{\mathrm{E}[V^2]}{2\mathrm{E}[V]} \\ &\quad + p_{su}\frac{\mathrm{E}[H^2]}{2\mathrm{E}[H]} + p_{sb}\sum_{n=\alpha}^{N-1} n\psi_n^{sb}\mathrm{E}[S]e \\ &\quad - \pi 2(1-\rho)[\mathrm{E}[S^2]D^{(1)} + \mathrm{E}^2[S]D^{(2)}]e \\ &\quad + \frac{1}{\lambda}\left[p_{vac}\sum_{n=0}^{\alpha-1}\psi_n^V + p_{su}\kappa + p_{sb}\sum_{n=\alpha}^{N-1}\psi_n^{sb}\right](D+e\pi)^{-1}D^{(1)}e \\ &\quad - \pi\lambda(1-\rho)[I-\mathrm{E}[S]D^{(1)}](D+e\pi)^{-1}D^{(1)}e.\end{aligned}$$

Thus, we get

$$\begin{aligned}-w_N^{*(1)}(\theta)\Big|_{\theta=0}e &= -\frac{z^{*(1)}(0)D^{(1)}e}{\lambda} + \frac{\mathrm{E}[S]z^*(0)D^{(2)}e}{2\lambda} \\ &= -\frac{1}{\lambda}\left[p_{vac}\sum_{n=0}^{\alpha-1}\psi_n^V + p_{su}\kappa + p_{sb}\sum_{n=\alpha}^{N-1}\psi_n^{sb}\right](D+e\pi)^{-1}D^{(1)}e \\ &\quad + p_{vac}\sum_{n=0}^{\alpha-1} n\psi_n^V\mathrm{E}[S]e + p_{vac}\mathrm{E}[H] + p_{vac}\frac{\mathrm{E}[V^2]}{2\mathrm{E}[V]} \\ &\quad + p_{sb}\sum_{n=\alpha}^{N-1} n\psi_n^{sb}\mathrm{E}[S]e + p_{su}\frac{\mathrm{E}[H^2]}{2\mathrm{E}[H]} + \frac{\lambda\mathrm{E}[S^2]}{2(1-\rho)} \\ &\quad + \frac{\pi\mathrm{E}^2[S]D^{(2)}e}{2\rho(1-\rho)} + \frac{1}{1-\rho} - \frac{\pi\mathrm{E}[S]D^{(1)}(D+e\pi)^{-1}D^{(1)}e}{\lambda}(1-\rho).\end{aligned}$$

Appendix 3: Derivation of (5.16)

Let $Q_n^{(\alpha,N)}$ be the matrix probability that there are n customers at the start of the busy period. Then, we have

$$Q_n^{(\alpha,N)} = \sum_{n=N}^{\infty} H_{n(\alpha)}^{+} + \sum_{n=N}^{\infty}\sum_{k=\alpha}^{N-1}\Phi_k^{sb}(-D_0)^{-1}D_{n-k}.$$

Taking GF and using (5.15), we have

$$\begin{aligned}Q_{(\alpha,N)}(z) &= \sum_{n=N}^{\infty} H_{n(\alpha)}^{+}z^n + \sum_{k=\alpha}^{N-1}\Phi_k^{sb}(-D_0)^{-1}z^k D(z) - \sum_{k=\alpha}^{N-1}\Phi_k^{sb}(-D_0)^{-1}z^k \\ &\quad \cdot \sum_{i=0}^{N-k-1} D_i z^i. \qquad (5.39)\end{aligned}$$

The last term in (5.39) becomes

$$\begin{aligned}
&\sum_{k=\alpha}^{N-1} \Phi_k^{sb}(-D_0)^{-1} z^k \sum_{i=0}^{N-k-1} D_i z^i \\
&\quad = \sum_{k=\alpha}^{N-1} \Phi_k^{sb}(-D_0)^{-1} z^k \sum_{i=1}^{N-k-1} D_i z^i + \sum_{k=\alpha}^{N-1} \Phi_k^{sb}(-D_0)^{-1} z^k D_0 \\
&\quad = \sum_{k=\alpha+1}^{N-1} \sum_{i=\alpha}^{k-1} \Phi_i^{sb}(-D_0)^{-1} D_{k-i} z^k - \sum_{k=\alpha}^{N-1} \Phi_k^{sb} z^k,
\end{aligned} \tag{5.40}$$

where we used

$$\sum_{k=\alpha}^{N-1} \Phi_k^{sb}(-D_0)^{-1} z^k \sum_{i=1}^{N-k-1} D_i z^i = \sum_{k=\alpha+1}^{N-1} \sum_{i=\alpha}^{k-1} \Phi_i^{sb}(-D_0)^{-1} D_{k-i} z^k.$$

Using $\Phi_k^{sb} = \sum_{i=\alpha}^{k} H_{i(\alpha)}^{+} D_{k-i}^{*}$ in (5.40), we get

$$\begin{aligned}
&\sum_{k=\alpha+1}^{N-1} \sum_{i=\alpha}^{k-1} \Phi_i^{sb}(-D_0)^{-1} D_{k-i} z^k - \sum_{k=\alpha}^{N-1} \Phi_k^{sb} z^k \\
&\quad = \sum_{k=\alpha+1}^{N-1} \sum_{i=\alpha}^{k-1} \sum_{n=\alpha}^{i} H_{n(\alpha)}^{+} D_{i-n}^{*}(-D_0)^{-1} D_{k-i} z^k - \sum_{k=\alpha}^{N-1} \sum_{i=\alpha}^{k} H_{i(\alpha)}^{+} D_{k-i}^{*} z^k.
\end{aligned} \tag{5.41}$$

Let us simplify (5.41). For convenience, let us define

$$\sum_{k=\alpha+1}^{N-1} \Gamma_k z^k = \sum_{k=\alpha+1}^{N-1} \sum_{i=\alpha}^{k-1} \sum_{n=\alpha}^{i} H_{n(\alpha)}^{+} D_{i-n}^{*}(-D_0)^{-1} D_{k-i} z^k. \tag{5.42}$$

Then, using $D_k^{*} = \sum_{l=0}^{k-1} D_l^{*}(-D_0)^{-1} D_{k-l}$, we get

$$\sum_{k=\alpha+1}^{N-1} \Gamma_k z^k = \sum_{k=\alpha+1}^{N-1} \sum_{i=\alpha}^{k-1} H_{i(\alpha)}^{+} D_{k-i}^{*} z^k. \tag{5.43}$$

Using (5.43) in (5.42), we get

$$\begin{aligned}
&\sum_{k=\alpha+1}^{N-1} \sum_{i=\alpha}^{k-1} \Phi_i^{sb}(-D_0)^{-1} D_{k-i} z^k - \sum_{k=\alpha}^{N-1} \Phi_k^{sb} z^k \\
&\quad = \sum_{k=\alpha+1}^{N-1} \sum_{i=\alpha}^{k-1} H_{i(\alpha)}^{+} D_{k-i}^{*} z^k - \left[\sum_{k=\alpha} +1^{N-1} \sum_{i=\alpha}^{k-1} H_{i(\alpha)}^{+} D_{k-i}^{*} z^k + \sum_{k=\alpha}^{N-1} H_{k(\alpha)}^{+} z^k \right] \\
&\quad = -\sum_{k=\alpha}^{N-1} H_{k(\alpha)}^{+} z^k.
\end{aligned} \tag{5.44}$$

Using (5.44) in (5.40), we have

$$\sum_{k=\alpha}^{N-1} \Phi_k^{sb}(-D_0)^{-1} z^k \sum_{i=0}^{N-k-1} D_i z^i = -\sum_{k=\alpha}^{N-1} H_{k(\alpha)}^{+} z^k.$$

Thus, (5.39) becomes

$$\begin{aligned} Q_{(\alpha,N)}(z) &= \sum_{n=N}^{\infty} H_{n(\alpha)}^{+} z^n + \sum_{n=\alpha}^{N-1} \Phi_n^{sb}(-D_0)^{-1} z^n D(z) + \sum_{n=\alpha}^{N-1} H_{n(\alpha)}^{+} z^n \\ &= \sum_{n=\alpha}^{\infty} H_{n(\alpha)}^{+} z^n + \sum_{n=\alpha}^{N-1} \Phi_n^{sb}(-D_0)^{-1} z^n D(z). \end{aligned}$$

Now, using $\sum_{n=\alpha}^{\infty} H_{n(\alpha)}^{+} z^n = \sum_{n=\alpha}^{\infty} H_{n(\alpha)}^{-} z^n H(z)$ finishes the proof.

References

1. M. Yadin and P. Naor, Queueing systems with a removable server station, *Operational Research Quarterly,* vol. 14, pp. 393–405, 1963.
2. M. Hersh and I. Brosh, The optimal strategy structure of an intermittently Operated Service Channel, *Europ. Journal Operational Research*, vol. 5, pp. 133–141, 1980.
3. M. Hofri, Queueng systems with a procrastinating server, in *Proc. Performance'86 and ACM-SIGMETRICS (1980), Performance Evaluation Review*, vol. 14, no. 1, pp. 245–253, 1986.
4. O. Kella, The threshold policy in the M/G/1 queue with server vacations, *Naval Research Logist.*, vol. 36, pp. 111–123, 1989.
5. H. S. Lee and M. M. Srinivasan, Control policies for the $M^X/G/1$ queueing system, *Mgmt. Science*, vol. 35, no. 6, pp. 708–721, 1989.
6. H. Takagi, *Queueing Analysis: A Foundation of Performance Evaluation, Vol. I, Vacation and Priority Systems, Part I*. North-Holland, 1991.
7. H. W. Lee, S. S. Lee, and K. C. Chae, Operating characteristics of queue with *N*-policy, *Queueing Systems*, vol. 15, pp. 387–399, 1994.
8. H. W. Lee and B. Y. Ahn, Operational behavior of the MAP/G/1 queue under *N*-policy with single vacation and set-up, *Appl. Math. & Stoch. Analysis*, vol. 15, no. 2, 167–196, 2002.
9. H. W. Lee and J. O. Park, Optimal strategy in *N*-policy system with early setup, *Journal of Operational Research Society*, vol. 48, pp. 306–313, 1997.
10. H. W. Lee, N. I. Park, and J. Jeon, Application of the factorization property to the analysis of production systems with a non-renewal input, bilevel threshold control and maintenance, in *Proc. the Fourth International Conference on Matrix-Analytic Methods in Stochastic Models Matrix-Analytic Methods: Theory and Applications (Eds. Guy Latouche and Peter Taylor)*, pp. 219–236, 2002.
11. D. M. Lucantoni, 'New results on the single server queue with a batch Markovian arrival process, *Stochastic Models*, vol. 7, no. 1, pp. 1–46, 1991.
12. D. M. Lucantoni, The BMAP/G/1 queue: A tutorial, *Models and Technique for Performance Evaluation of Computer and Communications Systems, (L. Donatiello and R. Nelson Ed.)*, Springer Verlag, pp. 330–358, 1993.
13. V. Ramaswami, Stable recursion for the steady state vector for Markov chains of M/G/1 type, *Stochastic Models*, vol. 4, pp. 183–188, 1988.
14. G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM series on Statistics and Applied Probability, 1999.

15. S. H. Chang, T. Takine, K. C. Chae, and H. W. Lee, A unified queue length formula for BMAP/G/1 queue with generalized vacations, *Stochastic Models*, vol. 18, no. 3, pp. 369–386, 2002.
16. H. W. Lee, N. I. Park and J. Jeon, A new approach to queue lengths and waiting times of BMAP/G/1 queues, *Computers & Operations Research*, vol. 30, pp. 2021–2045, 2003.
17. S. Kasahara, T. Takine, Y. Takahashi and T. Hasegawa, MAP/G/1 queues under N-policy with and without vacations, *Journal Operational Research Society of Japan*, vol. 39, no. 2, pp. 188–212, 1996.

Chapter 6
Performance Analysis of an M/E_k/1 Queue with Balking and Two Service Rates Based on a Single Vacation Policy

Chunyan Li, Wuyi Yue, and Dequan Yue

Abstract In this chapter, we present an analysis for an M/E_k/1 queue with balking and two service rates based on a single vacation policy. Customers are served at two different rates depending on the number of customers in the system. If customers on arrival find other customers in the system, they either decide to enter the queue or balk with a constant probability. The server takes a single vacation when the system becomes empty. We first formulate a quasi birth-and-death process for the queueing system. Then, we obtain the equilibrium condition of the system. By using the matrix-geometric solution method, we obtain the matrix-geometric form solution for the steady-state probability vectors. The computation of the boundary steady-state probability vectors is also discussed. Then, we derive explicitly performance measures of the system. Based on this performance analysis, we develop a cost model to determine numerically the system's optimal cost and optimal critical value. Finally, we perform a sensitivity analysis through numerical experiments.

6.1 Introduction

In this chapter, we consider an M/E_k/1 queueing system with balking and two service rates based on a single vacation policy. Customers are served at two different rates depending on the number of customers in the system. If customers on arrival

C. Li
Department of Sciences, College of Zhijiang, Zhejiang University of Technology, Hangzhou 310024, China
e-mail: llccyy1980@126.com

W. Yue
Department of Intelligence and Informatics, Konan University, Kobe 658-8501, Japan
e-mail: yue@konan-u.ac.jp

D. Yue
College of Sciences, Yanshan University, Qinhuangdao 066004, China
e-mail: ydq@ysu.edu.cn

W. Yue et al. (eds.), *Advances in Queueing Theory and Network Applications,*

find other customers in the system, they either decide to enter the queue or balk with a constant probability. Balking is not only a common phenomenon in queues arising in daily activities, but is also found in applications in communication systems, production line systems, and in various machine interference or repair models (see, e.g., [1], [2], and references therein).

The queueing systems with balking, or reneging, or both have been studied by many researchers. Haight [3] was the first person who considered an M/M/1 queue with balking. An M/M/1 queue with customer reneging was also proposed by Haight [4]. The combined effects of balking and reneging in an M/M/1 queue with limited waiting room and unlimited waiting room have been investigated by Ancker and Gafarian [5], [6]. They obtained the steady-state probabilities and some performance measures of the system such as the mean number of customers in the queue, the mean number of customers in the system, and the mean rate of customer loss.

Abou-El-Ata [7] extended the model in [5] to study a state-dependent M/M/1/N queue with reneging and a general balk function, where the server has two service rates depending on the number of customers in the system. They obtained the transient solution of the state probabilities. Al-Seedy [8] extended the model in [7] to a state-dependent M/E_k/1/N queue with balking. By solving the steady-state probability-difference equations, Al-Seedy [8] obtained some iterative expressions of the steady-state probabilities. A state-dependent M/M/1 queue with balking was studied by Al-Seedy and Kotb [9]. Recently, Yue, Li, and Yue [10] extended the model in [8] to a state-dependent M/E_k/1 queueing system with balking. They formulated a quasi birth-and-death (QBD) process, and obtained the steady-state probability vector and some performance measures of the system.

Similarly, queueing models with vacations have been studied by many researchers and have been found to be applicable in analyzing numerous real-world queueing situations such as flexible manufacturing systems, service systems, and telecommunication systems. Excellent surveys of queueing systems with server vacations can be found in the paper by Doshi [11] and the book by Takagi [12]. However, most of the research works on queueing systems with balking have not considered server vacations. There were only a few papers that we know of that considered queueing systems with balking and server vacations (see, e.g., [1], [13], and [14]). In this chapter, we study an M/E_k/1 queueing system with balking and single vacations.

The rest of this chapter is organized as follows. In Sect. 6.2, we formulate a QBD process and obtain the equilibrium condition for the system. In Sect. 6.3, by using a matrix-geometric solution method, we derive the explicit expression for the steady-state probability vector. We also discuss the computation of the boundary steady-state probability vectors. In Sect. 6.4, we derive explicitly some performance measures of the system such as the expected number of the customers in the system, the expected number of customers in the queue, and the mean balking rate of the system. Based on this performance analysis, we develop a cost model to determine numerically the optimal cost and optimal critical value of the system.

In Sect. 6.5, we perform sensitivity analysis through numerical experiments. Conclusions are given in Sect. 6.6.

6.2 System Model and Equilibrium Condition

In this chapter, we consider an M/E$_k$/1 queueing system with balking and two service rates based on a single vacation policy.

6.2.1 System Model

The assumptions of the system model are as follows:

(1) Customers arrive according to a Poisson process with arrival rate λ. There is one server in the system. If customers on arrival find other customers in the system, they either decide to enter the queue with a probability β or balk with a probability $1-\beta$.

(2) Customers are served on a First-Come First-Served (FCFS) basis. Once service commences, it always proceeds to completion. The service times are assumed to be distributed according to an Erlang distribution with mean k/μ_n; that is, the service time is made up of k independent and identical exponential stages, each with mean $1/\mu_n$, given by

$$\mu_n = \begin{cases} \mu_1, & n = 1,2,\ldots,r \\ \mu_2, & n = r+1, r+2, \ldots. \end{cases}$$

This means that the server at each service stage has two rates, say "slow and fast," depending on the number of customers n in the system. When the number of customers n in the system is less than or equal to the critical value r, the server has a slow service rate μ_1; otherwise, the server has a fast service rate μ_2 $(0 < \mu_1 < \mu_2 < \infty)$.

(3) When the system is empty, the server goes on a vacation. If the server returns from a vacation to find customers waiting, it begins to serve those waiting customers; otherwise, the server is idle and begins serving whenever customers arrive. This type of vacation is called "single vacation". The server's vacation time follows an exponential distribution with the vacation rate η $(\eta > 0)$.

(4) The interarrival times, service times, and vacations are mutually independent.

6.2.2 Equilibrium Condition

In the following, we first formulate a QBD process. Then, we provide the equilibrium condition of the system.

Let $N(t)$ denote the number of customers in the system at time t, and let $J(t)$ denote the service stage for the customer being served at time $t(t \geq 0)$. A customer goes into the first stage of the service, then progresses through the remaining stages, and must complete the last stage. Let $J(t) = 0$, if the server goes on vacation at time t; $J(t) = i$, if the server is servicing the customer and the customer goes into the ith service stage at time t; and $J(t) = -1$, if the server is idle at time t. The state space of the two-dimensional process $\{(N(t), J(t)); t \geq 0\}$ is given by

$$S = \{(0, j); j = -1, 0\} \cup \{(i, j); i = 0, 1, \ldots, j = 1, 2, \ldots, k\}.$$

All states of this two-dimensional process are labeled in lexicographic order as follows:

$$(0,0); (0,-1); (1,0), (1,1), \ldots, (1,k); (2,0), (2,1), \ldots, (2,k); \ldots.$$

By probability analysis, the process $\{(N(t), J(t)); t \geq 0\}$ has the following infinitesimal generator.

$$\mathbf{Q} = \begin{pmatrix} \mathbf{B}_0 & \mathbf{C}_0 & & & & \\ \mathbf{A}_0 & \mathbf{B}_1 & \mathbf{C}_1 & & & \\ & \mathbf{A}_1 & \mathbf{B}_1 & \mathbf{C}_1 & & \\ & & \cdots & \cdots & \cdots & \\ & & & \mathbf{A}_1 & \mathbf{B}_1 & \mathbf{C}_1 \\ & & & & \mathbf{A}_2 & \mathbf{B}_2 & \mathbf{C}_1 \\ & & & & & \cdots & \cdots & \cdots \end{pmatrix} \begin{matrix} \cdots 0 \\ \cdots 1 \\ \cdots 2 \\ \vdots \\ \cdots r \\ \cdots r+1 \\ \vdots \end{matrix},$$

where

$$\mathbf{B}_0 = \begin{pmatrix} -(\lambda+\eta) & \eta \\ 0 & -\lambda \end{pmatrix}, \quad \mathbf{C}_0 = \begin{pmatrix} \lambda & 0 & 0 \cdots 0 \\ 0 & \lambda & 0 \cdots 0 \end{pmatrix},$$

$$\mathbf{A}_0 = \begin{pmatrix} 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ \mu_1 & 0 \end{pmatrix}, \quad \mathbf{A}_1 = \begin{pmatrix} 0 & 0 & 0 \cdots 0 \\ \vdots & \vdots & \vdots \quad \vdots \\ 0 & 0 & 0 \cdots 0 \\ 0 & \mu_1 & 0 \cdots 0 \end{pmatrix},$$

$$\mathbf{B}_1 = \begin{pmatrix} -(\beta\lambda+\eta) & \eta & 0 & \cdots & 0 & 0 \\ 0 & -(\beta\lambda+\mu_1) & \mu_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -(\beta\lambda+\mu_1) & \mu_1 \\ 0 & 0 & 0 & \cdots & 0 & -(\beta\lambda+\mu_1) \end{pmatrix},$$

$$\mathbf{C}_1 = \begin{pmatrix} \beta\lambda & 0 & \cdots & 0 \\ 0 & \beta\lambda & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \beta\lambda \end{pmatrix}, \quad \mathbf{A}_2 = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & \mu_2 & 0 & \cdots & 0 \end{pmatrix},$$

$$\mathbf{B}_2 = \begin{pmatrix} -(\beta\lambda+\eta) & \eta & 0 & \cdots & 0 & 0 \\ 0 & -(\beta\lambda+\mu_2) & \mu_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -(\beta\lambda+\mu_2) & \mu_2 \\ 0 & 0 & 0 & \cdots & 0 & -(\beta\lambda+\mu_2) \end{pmatrix},$$

where $\mathbf{B}_0$ is a 2×2 matrix, $\mathbf{C}_0$ is a $2\times(k+1)$ matrix, $\mathbf{A}_0$ is a $(k+1)\times 2$ matrix, and the other matrices are $(k+1)\times(k+1)$ matrices.

From the book written by Neuts [15], we know that $\{(N(t),J(t));t\geq 0\}$ is a QBD process. Let $\mathbf{H}=\mathbf{A}_2+\mathbf{B}_2+\mathbf{C}_1$, then $\mathbf{H}$ can be given by

$$\mathbf{H} = \begin{pmatrix} -\eta & \eta & 0 & \cdots & 0 & 0 \\ 0 & -\mu_2 & \mu_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\mu_2 & \mu_2 \\ 0 & \mu_2 & 0 & \cdots & 0 & -\mu_2 \end{pmatrix}.$$

It is readily known that $\mathbf{H}$ is an irreducible generator. Let $\pi=(\pi_0,\pi_1,\ldots,\pi_k)$ be a $(k+1)$-dimensional row vector of the steady-state probability of $\mathbf{H}$. Then, π satisfies the linear equations: $\pi\mathbf{H}=0$ and $\pi\mathbf{e}=1$, where $\mathbf{e}=(1,1,\ldots,1)$ is a column vector with $(k+1)$ elements. Solving the linear equations, we get

$$\pi_0=0, \quad \pi_i=\frac{1}{k}, \quad i=1,2,\ldots,k. \tag{6.1}$$

By Theorem 3.1.1 in [15], the equilibrium condition of the system is given by $\pi\mathbf{A}_2\mathbf{e}>\pi\mathbf{C}_1\mathbf{e}$. Substituting π with (6.1), we then have the equilibrium condition for the system given by

$$\frac{k\beta\lambda}{\mu_2}<1. \tag{6.2}$$

6.3 Steady-State Probability Vector

From the discussion in Sect. 6.2, we know that the steady-state probability vector of $\mathbf{Q}$ exists under the equilibrium condition given by (6.2). In this section, we derive the explicit expression for the steady-state probability vector.

Let $\mathbf{X} = (\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_r, \mathbf{X}_{r+1}, \ldots)$, where $\mathbf{X}_0 = (x_0, x_{-1})$ is a row vector with two elements, and $\mathbf{X}_i = (x_{i0}, x_{i1}, x_{i2}, \ldots, x_{ik})$ is a row vector with $(k+1)$ elements, $i = 1, 2, \ldots$. By applying the matrix-geometric solution method, the stationary probability vector is given by

$$\mathbf{X}_i = \mathbf{X}_r \mathbf{R}^{i-r}, \quad i = r, r+1, \ldots, \tag{6.3}$$

where $\mathbf{R}$ is the minimal nonnegative solution to the equation $\mathbf{R}^2\mathbf{A}_2 + \mathbf{R}\mathbf{B}_2 + \mathbf{C}_1 = 0$, and the boundary steady-state probability vectors $\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_r$ are given by solving the following equations:

$$\mathbf{X}_0\mathbf{B}_0 + \mathbf{X}_1\mathbf{A}_0 = \mathbf{0}, \tag{6.4}$$

$$\mathbf{X}_0\mathbf{C}_0 + \mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_2\mathbf{A}_1 = \mathbf{0}, \tag{6.5}$$

$$\mathbf{X}_i\mathbf{C}_1 + \mathbf{X}_{i+1}\mathbf{B}_1 + \mathbf{X}_{i+2}\mathbf{A}_1 = \mathbf{0}, \quad i = 1, 2, \ldots, r-2, \tag{6.6}$$

$$\mathbf{X}_{r-1}\mathbf{C}_1 + \mathbf{X}_r(\mathbf{B}_1 + \mathbf{R}\mathbf{A}_2) = \mathbf{0}, \tag{6.7}$$

$$x_0 + x_{-1} + \sum_{i=1}^{r-1} \mathbf{X}_i\mathbf{e} + \mathbf{X}_r(\mathbf{I} - \mathbf{R})^{-1}\mathbf{e} = 1, \tag{6.8}$$

where $\mathbf{e} = (1, 1, \ldots, 1)$ is a column vector with $(k+1)$ elements. In order to solve (6.4)–(6.8), we define the following matrices:

$$\mathbf{M}_r = \mathbf{I}, \tag{6.9}$$

$$\mathbf{M}_{r-1} = -\frac{1}{\beta\lambda}(\mathbf{B}_1 + \mathbf{R}\mathbf{A}_2), \tag{6.10}$$

$$\mathbf{M}_i = -\frac{1}{\beta\lambda}(\mathbf{M}_{i+1}\mathbf{B}_1 + \mathbf{M}_{i+2}\mathbf{A}_1), \quad i = 1, 2, \ldots, r-2, \tag{6.11}$$

$$\mathbf{M}_0 = -(\mathbf{M}_1\mathbf{B}_1 + \mathbf{M}_2\mathbf{A}_1), \tag{6.12}$$

where $\mathbf{I}$ is the $(k+1) \times (k+1)$ identity matrix.

Let $\varepsilon_1 = (1, 0, 0, \ldots, 0)$ and $\varepsilon_2 = (0, 1, 0, \ldots, 0)$ be column vectors with $(k+1)$ elements, respectively. Let $\tilde{\mathbf{B}}_0 = (\varepsilon_1, \varepsilon_2)\mathbf{B}_0$ be the $(k+1) \times (k+1)$ matrix. We have the following theorem.

Theorem 6.1. *The solutions of (6.4)–(6.8) are given by*

$$\mathbf{X}_i = \mathbf{X}_r \mathbf{M}_i, \quad i = 1, 2, \ldots, r, \tag{6.13}$$

$$x_0 = \frac{1}{\lambda} \mathbf{X}_r \mathbf{M}_0 \boldsymbol{\varepsilon}_1, \tag{6.14}$$

$$x_{-1} = \frac{1}{\lambda} \mathbf{X}_r \mathbf{M}_0 \boldsymbol{\varepsilon}_2 \tag{6.15}$$

and $\mathbf{X}_r$ *satisfies the following equations:*

$$\mathbf{X}_r \left(\frac{1}{\lambda} \mathbf{M}_0 \tilde{\mathbf{B}}_0 + \mathbf{M}_1 \mathbf{A}_0 \right) = 0, \tag{6.16}$$

$$\mathbf{X}_r \left[\frac{1}{\lambda} \mathbf{M}_0 (\boldsymbol{\varepsilon}_1 + \boldsymbol{\varepsilon}_2) + \sum_{i=1}^{r-1} \mathbf{M}_i \mathbf{e} + (\mathbf{I} - \mathbf{R})^{-1} \mathbf{e} \right] = 1. \tag{6.17}$$

Proof. Note that $\mathbf{C}_1$ is invertible and $\mathbf{C}_1^{-1} = 1/(\beta\lambda)\mathbf{I}$. We have from (6.7) that

$$\mathbf{X}_{r-1} = \mathbf{X}_r \mathbf{M}_{r-1}. \tag{6.18}$$

This indicates that (6.13) holds for $i = r-1$. It is obvious that (6.13) holds for $i = r$. Suppose that (6.13) holds for $i = k+2, k+1$; then we have from (6.6) that

$$\begin{aligned} \mathbf{X}_k &= -(\mathbf{X}_{k+1} \mathbf{B}_1 + \mathbf{X}_{k+2} \mathbf{A}_1) \mathbf{C}_1^{-1} \\ &= -\frac{1}{\beta\lambda} \mathbf{X}_r (\mathbf{M}_{k+1} \mathbf{B}_1 + \mathbf{M}_{k+2} \mathbf{A}_1) \\ &= \mathbf{X}_r \mathbf{M}_k. \end{aligned} \tag{6.19}$$

Thus, by the inductive method, we conclude that (6.13) holds for $i = 1, 2, \ldots, r$. From (6.5), we have

$$\begin{aligned} \mathbf{X}_0 \mathbf{C}_0 &= -\mathbf{X}_r (\mathbf{M}_1 \mathbf{B}_1 + \mathbf{M}_2 \mathbf{A}_1) \\ &= \mathbf{X}_r \mathbf{M}_0. \end{aligned} \tag{6.20}$$

Note that $\mathbf{C}_0 \boldsymbol{\varepsilon}_1 = (\lambda, 0)$ and $\mathbf{C}_0 \boldsymbol{\varepsilon}_2 = (0, \lambda)$ are column vectors with two elements, respectively, we get (6.14) and (6.15). Substituting (6.14) and (6.15) into (6.4) and (6.8), we get (6.16) and (6.17). □

In general, it is difficult to give an exact expression of $\mathbf{R}$ except for a few special cases. However, the matrix $\mathbf{R}$ can be approximately calculated by the following iterative procedure:

(1) $\mathbf{R}(0) = 0$,

(2) $\mathbf{R}(n+1) = -(\mathbf{C}_1 + \mathbf{R}^2(n) \mathbf{A}_2) \mathbf{B}_2^{-1}, \quad n \geq 0.$

The proof of the convergence for this iterative algorithm is given in [15]. The matrix $\mathbf{B}_2^{-1}$ in the above algorithm exists, and can be explicitly given by

$$\mathbf{B}_2^{-1}=\begin{pmatrix} a & ab(-\eta) & ab^2(-\eta)(-\mu_2) & \cdots & ab^{k-1}(-\eta)(-\mu_2)^{k-2} & ab^k(-\eta)(-\mu_2)^{k-1} \\ 0 & b & b^2(-\mu_2) & \cdots & b^{k-1}(-\mu_2)^{k-2} & b^k(-\mu_2)^{k-1} \\ 0 & 0 & b & \cdots & b^{k-2}(-\mu_2)^{k-3} & b^{k-1}(-\mu_2)^{k-2} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & b & b^2(-\mu_2) \\ 0 & 0 & 0 & \cdots & 0 & b \end{pmatrix},$$

where $a=-1/(\beta\lambda+\eta)$ and $b=-1/(\beta\lambda+\mu_2)$.

6.4 Performance Measures and Cost Model

In this section, we give some useful performance measures of the system. Based on these performance measures, we develop a cost model to determine the optimal critical value to minimize the total expected cost per unit time.

6.4.1 Performance Measures

Using the steady-state probability vector $\mathbf{X}$ presented in Sect. 6.3, we can obtain some performance measures of the system.

Theorem 6.2.

(1) The expected number of customers in the queue is given by

$$\mathrm{E}[N_q]=\mathbf{X}_r\left\{\sum_{n=1}^{r-1} n\mathbf{M}_{n+1}+\mathbf{R}[(r-1)(\mathbf{I}-\mathbf{R})^{-1}+(\mathbf{I}-\mathbf{R})^{-2}]\right\}\mathbf{e}. \quad (6.21)$$

(2) The expected number of customers in the system is given by

$$\mathrm{E}[N]=\mathbf{X}_r\left\{\sum_{n=1}^{r} n\mathbf{M}_{n}+\mathbf{R}[r(\mathbf{I}-\mathbf{R})^{-1}+(\mathbf{I}-\mathbf{R})^{-2}]\right\}\mathbf{e}. \quad (6.22)$$

(3) The mean balking rate of the system is given by

$$BR=(1-\beta)\lambda(1-\mathbf{X}_0\mathbf{e}), \quad (6.23)$$

where $\mathbf{e}=(1,1,\ldots,1)$ *is a column vector with* $(k+1)$ *elements.*

(4) The probability that the server is busy is given by

$$P_B = \mathbf{X}_r \left\{ \sum_{n=1}^{r-1} \mathbf{M}_n + (\mathbf{I} - \mathbf{R})^{-1} \right\} \delta, \tag{6.24}$$

where $\delta = (0, 1, \ldots, 1)$ *is a column vector with* $(k+1)$ *elements.*

(5) The probability that the server goes on vacation is given by

$$P_V = 1 - P_B - x_{-1}. \tag{6.25}$$

(6) The probability that the server is idle is given by

$$P_I = x_{-1}. \tag{6.26}$$

Proof. The expected number of customers in the queue is given by

$$\mathrm{E}[N_q] = \sum_{n=1}^{\infty} \sum_{i=0}^{k} n x_{n+1,i} = \sum_{n=1}^{r-1} n \mathbf{X}_r \mathbf{M}_{n+1} \mathbf{e} + \sum_{n=r}^{\infty} n \mathbf{X}_r \mathbf{R}^{n-r+1} \mathbf{e}. \tag{6.27}$$

Hence, we obtain (6.21) by a summation of series. The expected number of customers in the system is given by

$$\mathrm{E}[N] = \sum_{n=1}^{\infty} \sum_{i=0}^{k} n x_{n,i} = \sum_{n=1}^{r} n \mathbf{X}_r \mathbf{M}_n \mathbf{e} + \sum_{n=r+1}^{\infty} n \mathbf{X}_r \mathbf{R}^{n-r} \mathbf{e}. \tag{6.28}$$

Hence, we get (6.22) by a summation of series. Using the concept of Ancker and Gafarian [5], the average balking rate of the system is given by

$$BR = \sum_{n=1}^{\infty} (1-\beta) \lambda \mathbf{X}_n \mathbf{e} = (1-\beta) \lambda (1 - \mathbf{X}_0 \mathbf{e}). \tag{6.29}$$

The processes of the proofs for (6.24)–(6.26) are obvious, hence the proofs for (6.24)–(6.26) have been omitted. □

6.4.2 Cost Model

In this subsection, we develop a steady-state expected cost function where the critical value r is a decision variable. Our objective is to determine the critical value r to minimize the total expected cost per unit time.

Let C_1 be the cost per unit time when there are customers waiting for service, C_2 be the cost per unit time when the server is busy, C_3 be the cost per unit time when the server goes on vacation, C_4 be the lost cost per unit time when customers balk, and C_5 be the cost per unit time when the server is idle.

According to the definition of each of the cost parameters listed above, the total expected cost function per unit time is given by

$$F(r) = C_1 \mathrm{E}[N_q] + C_2 P_B + C_3 P_V + C_4 BR + C_5 P_I, \tag{6.30}$$

where $\mathrm{E}[N_q]$, P_B, P_V, BR, and P_I are given in (6.21) and (6.23)–(6.26). The first item is the cost incurred by the customer's waiting. The fourth item is the cost incurred by the loss of a customer. The second, the third, and the last items are the costs incurred by the server.

6.5 Sensitivity Analysis

In this section, we perform a sensitivity analysis on the optimal critical value r^* and its expected cost $F(r^*)$ based on changes in values of the system parameters such as the arrival rate λ, the slow service rate μ_1, the fast service rate μ_2, the vacation rate η, and the entering probability β. Let the distribution of the service time be a two-stage Erlang distribution, and the employed cost parameters $C_1 = 150$, $C_2 = 250$, $C_3 = 200$, $C_4 = 300$ and $C_5 = 100$. The numerical results of the optimal critical value r^* and its expected minimum cost $F(r^*)$ are illustrated in Figs. 6.1–6.5.

In Fig. 6.1, we fix $\mu_1 = 0.2$, $\mu_2 = 0.8$, $\eta = 0.5$, and $\beta = 0.5$, and display the optimal critical value r^* as well as its expected minimum cost $F(r^*)$ by varying the arrival rate λ. Figure 6.1 shows that: (i) the optimal critical value r^* decreases as λ increases from 0.05 to 0.1, and it does not change at all when λ varies from 0.1 to 0.3; (ii) the minimum expected cost $F(r^*)$ increases as λ increases. Intuitively, λ affects r^* slightly and affects $F(r^*)$ significantly.

In Fig. 6.2, we fix $\lambda = 0.1$, $\mu_2 = 0.8$, $\eta = 0.5$, and $\beta = 0.5$, and display the optimal critical value r^* as well as its expected minimum cost $F(r^*)$ by varying the slow service rate μ_1. Figure 6.2 shows that: (i) the optimal critical value r^* increases

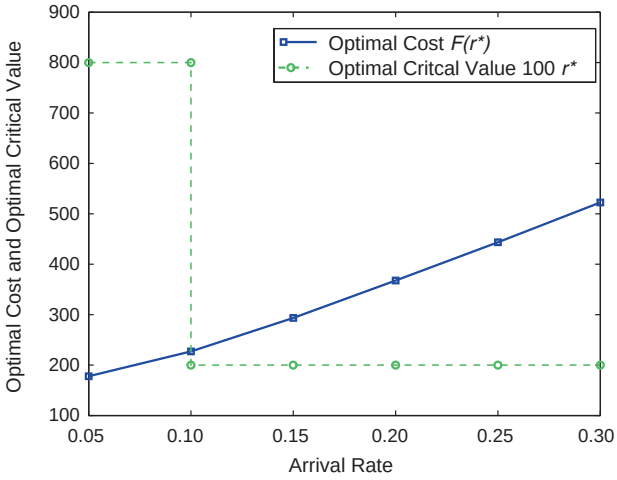


Fig. 6.1 Optimal critical value r^* and optimal cost $F(r^*)$ versus arrival rate λ for $\mu_1 = 0.2$, $\mu_2 = 0.8$, $\eta = 0.5$, and $\beta = 0.5$.

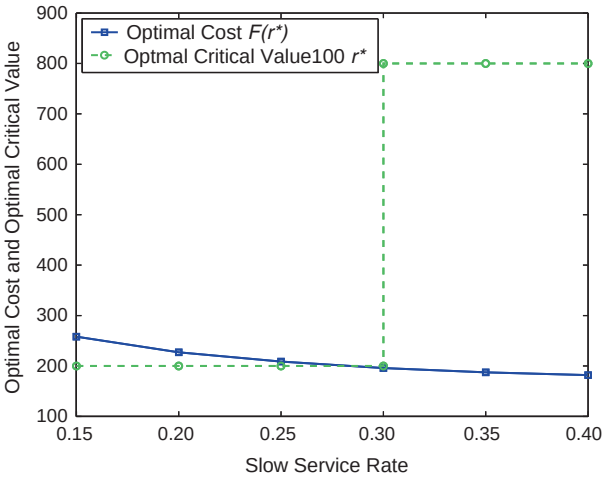


Fig. 6.2 Optimal critical value r^* and optimal cost $F(r^*)$ versus slow service rate μ_1 for $\lambda = 0.1$, $\mu_2 = 0.8$, $\eta = 0.5$, and $\beta = 0.5$.

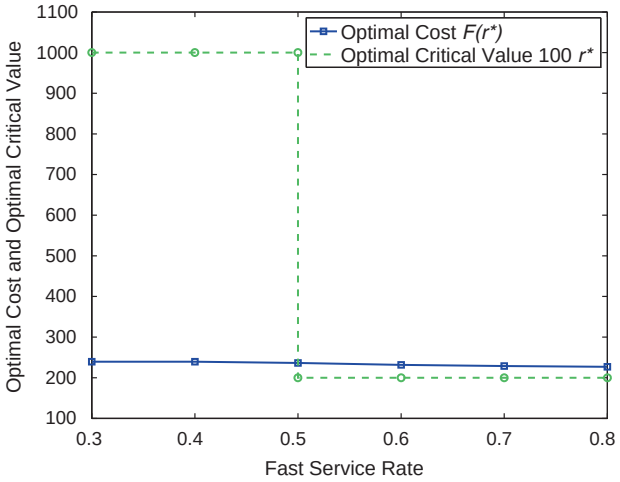


Fig. 6.3 Optimal critical value r^* and optimal cost $F(r^*)$ versus fast service rate μ_2 for $\lambda = 0.1$, $\mu_1 = 0.2$, $\eta = 0.5$, and $\beta = 0.5$.

as μ_1 increases; (ii) its minimum expected cost $F(r^*)$ decreases as μ_1 increases. Intuitively, μ_1 affects r^* and $F(r^*)$ significantly.

In Fig. 6.3, we fix $\lambda = 0.1$, $\mu_1 = 0.2$, $\eta = 0.5$, and $\beta = 0.5$, and display the optimal critical value r^* as well as its expected minimum cost $F(r^*)$ by varying the fast service rate μ_2. Figure 6.3 shows that: (i) the optimal critical value r^* decreases as μ_2 increases from 0.3 to 0.5, whereas it does not change at all when μ_2 varies from 0.5 to 0.8; (ii) the minimum expected cost $F(r^*)$ rarely changes when μ_2 varies from 0.3 to 0.8. Intuitively, μ_2 affects r^* slightly and affects $F(r^*)$ rarely.

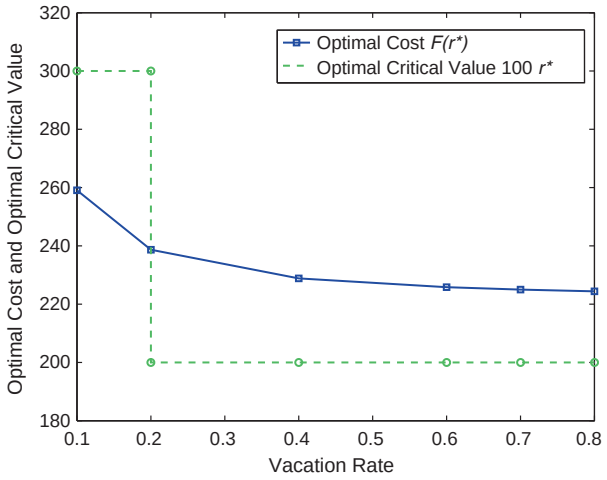


Fig. 6.4 Optimal critical value r^* and optimal cost $F(r^*)$ versus vacation rate η for $\lambda = 0.1$, $\mu_1 = 0.2$, $\mu_2 = 0.8$, and $\beta = 0.5$.

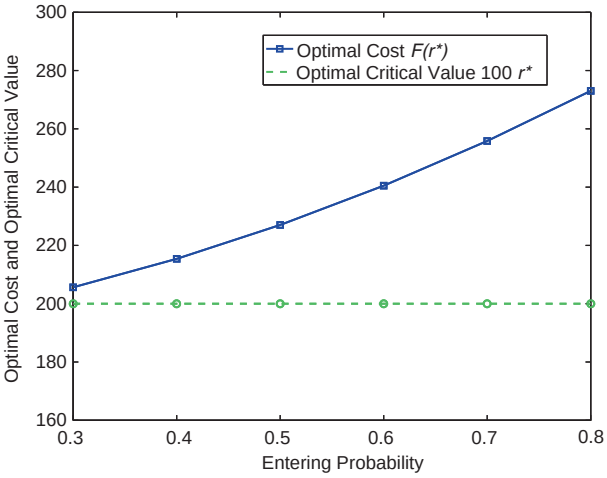


Fig. 6.5 Optimal critical value r^* and the optimal cost $F(r^*)$ versus the probability β for $\lambda = 0.1$, $\mu_1 = 0.2$, $\mu_2 = 0.8$, and $\eta = 0.5$.

In Fig. 6.4, we fix $\lambda = 0.1$, $\mu_1 = 0.2$, $\mu_2 = 0.8$, and $\beta = 0.5$, and display the optimal critical value r^* as well as its expected minimum cost $F(r^*)$ by varying the vacation rate η. Figure 6.4 shows that: (i) the optimal critical value r^* changes slightly when η varies from 0.1 to 0.8; (ii) the minimum expected cost $F(r^*)$ decreases slightly as η increases. Intuitively, η affects r^* and $F(r^*)$ slightly.

In Fig. 6.5, we fix $\lambda = 0.1$, $\mu_1 = 0.2$, $\mu_2 = 0.8$, and $\eta = 0.5$, and display the optimal critical value r^* and its expected minimum cost $F(r^*)$ by varying the entering probability β. Figure 6.5 shows that: (i) the optimal critical value r^* does not

change at all when β varies from 0.3 to 0.8; and (ii) the minimum expected cost $F(r^*)$ increases slightly as β increases. Intuitively, the optimal critical value r^* and its expected minimum cost $F(r^*)$ are insensitive to changes in β.

It appears from Figs. 6.1–6.5 that: (a) λ affects r^* slightly, and affects $F(r^*)$ significantly; (b) μ_1 affects r^* and $F(r^*)$ significantly; and (c) the optimal critical value r^* and its expected minimum cost $F(r^*)$ are insensitive to changes in μ_2, η, and β.

6.6 Conclusions

We considered an M/E$_k$/1 queueing system with balking and two service rates based on a single vacation policy. By using a matrix-geometric solution, we obtained the matrix solution of the steady-state probability distribution and the explicit expressions for some performance measures of the system. Based on these performance measures, we developed a cost model to determine the optimal critical value to minimize the total expected cost per unit time. Furthermore, we performed sensitivity analysis for the optimal critical value and its expected minimum cost with various parameters.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China (No. 70671088) and the Natural Science Foundation of Hebei Province (No. A2004000185), China, and was supported in part by GRANT-IN-AID FOR SCIENTIFIC RESEARCH (No. 19500070) and MEXT.ORC (2004-2008), Japan.

References

1. J. Ke, Operating characteristic analysis on the M^X/G/1 system with a variant vacation policy and balking, *Applied Mathematical Modelling*, vol. 31, pp. 1321–1337, 2007.
2. A. I. Shawky, The single-server machine interference model with balking, reneging and an additional server for longer queues, *Microelectronic and Reliability*, vol. 37, pp. 355–357, 1997.
3. F. A. Haight, Queueing with balking, *Biometrika*, vol. 44, pp. 360–369, 1957.
4. F. A. Haight, Queueing with reneging, *Metrika*, vol. 2, pp. 186–197, 1959.
5. C. J. Ancker Jr. and A. V. Gafarian, Some queueing problems with balking and reneging: I, *Operation Research*, vol. 11, pp. 88–100, 1963.
6. C. J. Ancker Jr. and A. V. Gafarian, Some queueing problems with balking and reneging: II, *Operation Research*, vol. 11, pp. 928–937, 1963.
7. M. O. Abou-EI-Ata, The state-dependent queue: M/M/1/N with reneging and general balk functions, *Microelectronic and Reliability*, vol. 31, pp. 1001–1007, 1991.
8. R. O. Al-Seedy, Analytical solution of the state-dependent Erlangian queue: M/E$_j$/1/N with balking, *Microelectronic and Reliability*, vol. 36, pp. 203–206, 1996.
9. R. O. Al-Seedy and K. A. M. Kotb, Transient solution of the state-dependent queue: M/M/1 with balking, *Advances in Modelling & Simulation*, vol. 35, pp. 55–64, 1991.

10. D. Yue, C. Li, and W. Yue, The matrix-geometric solution of the M/E_k/1 queue with balking and state-dependent service, *Nonlinear Dynamics and Systems Theory*, vol. 3, pp. 295–308, 2006.
11. B. Doshi, Single server queues with vacations, in: H. Takagi (Ed.), *Stochastic Analysis of Computer and Communication Systems*, pp. 217–265. The Netherlands: North-Holland, 1990.
12. H. Takagi, *Queueing Analysis: A Foundation of Performance Evaluation, Volume 1: Vacation and Priority Systems*. Amsterdam: Elsevier, 1991.
13. D. Yue, Y. Zhang, and W. Yue, Optimal performance analysis of an M/M/1/N queue system with balking, reneging and server vacation, *International Journal of Pure and Applied Mathematics*, vol. 28, pp. 101–115, 2006.
14. D. Yue, W. Yue, and Y. Sun, Performance analysis of an M/M/*c*/N queueing system with balking, reneging and synchronous vacations of partial servers, in *Proc. 6th International Symposium on Operations Research and Its Applications*, pp. 128–143, 2006.
15. M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models*. Baltimore: Johns Hopkins University Press, 1981.

Chapter 7
Markovian Polling Systems: Functional Computation for Mean Waiting Times and its Computational Complexity

Tetsuji Hirayama

Abstract We consider Markovian polling systems in which a single server serves J stations with Poisson arrivals and general service times. After completing a service period at station i, the server selects station j with probability p_{ij} and visits the station after spending a switchover time. We use the functional computation for mean waiting times that has been investigated in our previous research on multiclass M/G/1 type systems (e.g., [1] and [2]), which is different from the buffer occupancy method used in [3]. The advantages of the functional computation method are (1) its wide applicability to the analysis of M/G/1 type multiclass queues, and (2) its rather small computational complexity compared with the buffer occupancy method.

7.1 Introduction

A polling system is a multiclass queueing system in which a single server serves customers arriving at J stations according to some scheduling algorithm. It has been receiving much attention because of its ability to model a large variety of systems including computer communication networks, intelligent production systems, and transportation systems (e.g., [4] and [5]).

Several methods of analyzing various polling systems have been investigated. The leading method is the buffer occupancy method (e.g., [6]–[8]). This method has been used to analyze not only the standard system models but also various variants of the models that include a system with a mixture of exhaustive and gated disciplines [9], a system with simultaneous arrivals [10], a system with customers' feedback [11], and a nondeterministic polling system [3], and so on.

T. Hirayama
Graduate School of Systems and Information Engineering, University of Tsukuba,
Ibaraki 305-8573, Japan
e-mail: hirayama@cs.tsukuba.ac.jp

W. Yue et al. (eds.), *Advances in Queueing Theory and Network Applications,*

The other methods have also been investigated (e.g., [2], [12]–[14]). A fundamental survey of the analysis of polling systems was given in [15], and a detailed explanatory survey of these methods was given in [4]. The descendant set technique [16] has taken another approach to obtain the moments of the buffer occupancy variables, and was used to analyze a state-dependent polling system [17]. A stochastic decomposition was used to obtain a pseudo-conservation law for a weighted sum of the mean waiting times under various scheduling algorithms [18]. Another type of a decomposition theorem that relates a system with nonzero switchover times to a system with zero switchover times was investigated in [19].

Many of the research efforts listed above were concerned with the cyclic systems. On the other hand, various polling schemes other than cyclic have been investigated. Random polling systems in which the server next visits station j stochastically with probability p_j were considered by Kleinrock and Levy [20]. They were used to analyze the distributed access scheme to communication channels [4], [20].

Srinivasan [3] extended their analysis to nondeterministic polling systems (including Markovian polling systems) in which the server moves among stations according to general stochastic rules. Markovian polling systems with single buffers were investigated by Chung, Un, and Jung [21]. A system in which the server visits stations according to an arbitrary polling sequence (or table) of stations was considered by Baker and Rubin [22]. In this system, stations can be given higher priority by being listed more frequently in the polling table. Boxma, Levy, and Weststrate [23] found (approximate) formulas and procedures for determining the visit frequencies that optimize the system performances.

In this chapter, we consider Markovian polling systems in which a single server serves customers at J stations, and obtain their mean waiting times. Customers arrive according to Poisson processes and their service time distributions are general. After completing a service period at station i, the server selects station j with probability p_{ij} and visits it after spending a switchover time. The customer selection rule at each station is either gated or exhaustive. Although the system was already solved by the buffer occupancy method in [3], we take the other method (functional computation) that has been investigated in our previous research on multiclass M/G/1 type systems (e.g., [1], [2], [24], and [25]).

The key skill of our functional computation method is to consider the expected waiting time of a customer conditioned on the system state at its arrival epoch and represent it as a function of the system state. The advantages of the method are

(1) Its wide applicability to the analysis of mean waiting times in M/G/1 type multiclass queues
(2) Its rather small computational complexity necessary to calculate the mean waiting times for all stations as compared with the buffer occupancy method

Our method was initially applied to multiclass M/G/1 queues with priority [24], and then extended to the systems with customers' feedback [1]. Polling systems were initially investigated by our method in [2], and their multiclass extensions with customers' feedback were investigated in [25]. In all of the models, we have obtained the linear functional expressions for the conditional expected waiting (or

sojourn) times, which are the key property of our method, although the derivation procedures themselves are distinct among the models.

As for the computational complexity for computing the mean waiting times for all stations, our functional computation (for the Markovian polling system) requires us to solve $2J$ sets of $O(J)$ linear equations and a set of $O(J^2)$ (steady-state) linear equations. This means that our method at most requires $O(J^6)$ numerical operations. Furthermore a successive approximation method can be applied to solving the set of the latter steady-state linear equations, and then it can be shown that our method requires $O(J^4) + O(J^3 N)$ numerical operations where N is the number of its iterations. On the other hand, the buffer occupancy method requires us to solve the $O(J^3)$ linear equations for deriving the mean waiting times for all stations. If a successive approximation is applied to solving them, the method requires $O(J^4 N')$ numerical operations where N' is the number of its iterations.[1] Numerical examples are given in Sect. 7.6 of this chapter in order to compare the actual computational times in our method with those in the buffer occupancy method.

The rest of this chapter is organized as follows. In Sect. 7.2 we first define the system state that represents an evolution of the system. Its components include the numbers of customers and the remaining service time of a customer being served, and so on. Then we define some types of the expected waiting times for each customer conditioned on the system state at its arrival or relative polling instants. It is shown that these conditional expectations satisfy the "polling equation." In Sect. 7.3 we obtain the explicit expressions for some of the conditional expected waiting times. We further obtain the conditional expected numbers of customers at the next polling instants. In Sect. 7.4 the explicit expression for the overall expected waiting time is obtained by solving the polling equation. It can be shown that the expression has the linear functional form. In Sect. 7.5 the mean waiting times and the mean numbers of customers in a steady-state are obtained from the expression by using the generalized Little's formula and the PASTA property. Then we discuss the computational complexity of our functional computation method in detail in Sect. 7.6.

7.2 Model Description

In this section, we describe our model of the Markovian polling systems. A single server serves J groups of customers at J stations with infinite buffer capacities. Customers arrive at station i from outside the system according to a Poisson process with rate λ_i, and are called i-customers $(i = 1, \ldots, J)$. The overall arrival rate is denoted by $\lambda = \sum_{i=1}^{J} \lambda_i$. These customers are numbered in order of arrival, and let

[1] For a cyclic or random polling system, only $O(J^3)$ numerical operations are required for each iteration of the approximation for the buffer occupancy equations. But for a Markovian polling system, $O(J^4)$ operations are required for each iteration and the overall complexity becomes $O(J^4 N')$. For more detail, see [26] and (4.14) in [3].

c^e and τ_0^e denote the eth arriving customer itself and its arrival epoch, respectively $(e = 1, 2, \ldots)$.[2]

Service times S_i of i-customers are independently, identically, and arbitrarily distributed with mean $\mathrm{E}[S_i] > 0$ and second moment $\overline{s_i^2}$. Customers are served according to a predetermined scheduling algorithm defined below. The service is nonpreemptive. After receiving a service, each customer departs from the system. We define resource utilizations $\rho_i = \lambda_i \mathrm{E}[S_i]$, and put the usual assumption that $\rho = \sum_{i=1}^{J} \rho_i < 1$.

After completing a service period (defined below) at station i, the server selects a station in a Markovian manner where station j is selected with probability p_{ij}, and then visits station j after spending an arbitrarily distributed switchover time with mean $\overline{s_{ij}^o}$ and second moment $\overline{s_{ij}^{o2}}$, $(i, j = 1, \ldots, J)$. Let $\mathbf{P} = (p_{ij} : i, j = 1, \ldots, J)$ be the switching probability matrix, and assume that the Markov chain generated by the transition probability matrix $\mathbf{P}$ is irreducible. Furthermore, the arrival processes, the service times, and the server switching processes are assumed to be independent of each other.

The system is separated into two parts which are called the "service facility" and the "waiting room." There is a gate at each station between its queue in the waiting room and its queue in the service facility. And each arriving customer enters the queue in the service facility when the gate is opened; otherwise, it enters the queue in the waiting room. When the server visits a station, its gate is opened in order to admit some customers at the station to the service facility. The server serves the customers in the service facility until the server empties it, and then visits another station. Because the gates of the stations that are not visited by the server are closed, all customers at such stations must wait for service in the waiting room.

Each time interval from when the server visits a station until the first time when the server empties the service facility is called a *service period*.[3] Each time interval when the server switches from a station to another station is called a *switchover period*. Let $\Pi = \{1, \ldots, J\}$ be the set of (indices of) the service periods where $i \in \Pi$ denotes the service period of station i. And let $\Pi^s = \{(i, j) : i, j = 1, \ldots, J\}$ be the set of (indices of) the switchover periods where (i, j) denotes a switchover period from station i to station j.

A scheduling algorithm is specified as follows: (1) Selection order of the stations by the server, which is the Markovian as described before, (2) customer selection rule at each station used when the server admits customers into the service facility, which is either gated or exhaustive, and (3) service order of customers in the service facility, which is First-Come First-Served (FCFS).

When the server selects one of the stations with the gated rule, all customers staying at the station just when the server visits it enter its queue in the service facility, and then the gate is immediately closed. $\mathscr{H}_g$ denotes the set of stations with the gated rule. When the server selects one of the stations with the exhaustive rule,

[2] These customers arrive from outside the system according to a Poisson process with rate λ, and each of them becomes an i-customer with probability λ_i/λ when it arrives $(i = 1, \ldots, J)$.

[3] A time epoch when the server visits a station is called a service period beginning epoch or a polling instant.

the gate of the station remains open (i.e., customers arriving at the station later may still enter the service facility) and the server continues to serve all customers until the station is cleared of customers for the first time. The service period of the station finishes at this time, and its gate is closed. $\mathscr{H}_e$ denotes the set of the stations with the exhaustive rule.

Let us consider the system operating under a specified scheduling algorithm. For any e $(e = 1,2,\ldots)$, let $\{\tau_k^e : k = 1,2,\ldots\}$ be a sequence of all polling instants (i.e., service period beginning epochs) of all stations that occur after the c^es' arrival epoch.[4] Furthermore let $X_S^e(t)$ denote the station at which c^e stays at time t, or $X_S^e(t) = 0$ if it does not stay in the system at time t. Let $\mathscr{R}, \mathscr{R}_+, \mathscr{I}_+$ be, respectively, the set of real numbers, the set of nonnegative real numbers, and the set of nonnegative integers. For any event $\mathscr{K}$, let

$$\mathbf{1}\{\mathscr{K}\} = \begin{cases} 1, & \text{if event } \mathscr{K} \text{ is true} \\ 0, & \text{if event } \mathscr{K} \text{ is false.} \end{cases}$$

Then let $\kappa(t)$ denote a period that the system experiences at time t; that is the server is in a service period of station $\kappa(t)$ if $\kappa(t) \in \Pi$, or the server is in a switchover period from station i to station j if $\kappa(t) = (i,j) \in \Pi^s$. Let $r(t)$ denote the remaining service time of a customer being served at time t if $\kappa(t) \in \Pi$, or the remaining length of a switchover period if $\kappa(t) \in \Pi^s$.

The number of i-customers in the service facility at time t (who are not being served) is denoted by $g_i(t)$, and the number of i-customers in the waiting room at time t is denoted by $n_i(t)$. Let $\mathbf{g}(t) = (g_1(t),\ldots,g_J(t)) \in \mathscr{I}_+^J$, and let $\mathbf{n}(t) = (n_1(t),\ldots,n_J(t)) \in \mathscr{I}_+^J$. We also specify the other information $L(t)$ of the system at time t. The sample paths of these processes are assumed to be left-continuous with right-hand limits, except for $X_S^e(t), \kappa(t)$, and $L(t)$ which are right-continuous with left-hand limits.

Let us consider transition epochs of these processes consisting of customer arrival epochs, service completion epochs, and switchover period completion epochs. Then we define the stochastic process as

$$\mathscr{Q} = \{\mathbf{Y}(t) = (\kappa(t), r(t), \mathbf{g}(t), \mathbf{n}(t), L(t)) : t \geq 0\} \tag{7.1}$$

which represents an evolution of the system. For any scheduling algorithm defined above, $\mathscr{Q}$ may embed a Markov process. Possible values of $\mathbf{Y}(t)$ $(t \geq 0)$ are called *states*, and the state space of $\mathscr{Q}$ is denoted by $\mathscr{E}$.

We define three types of the performance measures of customer c^e $(e = 1,2,\ldots)$. The first type is related to the c^es' waiting times in the waiting room. We define for any $t \geq 0$ and $i = 1,\ldots,J$,

$$C_{Wi}^e(t) = \begin{cases} 1, & \text{if } c^e \text{ stays in the waiting room as an } i\text{-customer at time } t \\ 0, & \text{otherwise.} \end{cases} \tag{7.2}$$

[4] Note that τ_0^e is the customer's arrival epoch, and we assume that $\tau_0^e < \tau_1^e < \tau_2^e < \cdots$.

The c^es' waiting time spent in the waiting rooms is defined by

$$W_i^e = \int_0^\infty C_{Wi}^e(t)dt, \qquad (i = 1,\ldots,J). \tag{7.3}$$

Then, for $l = 0,1,2\ldots$, the expected waiting times in the waiting room during the time interval $[\tau_l^e, \tau_{l+1}^e)$ conditioned on the state of the system are defined by

$$W_i^0(\mathbf{Y},e,l) = \mathrm{E}\left[\int_{\tau_l^e}^{\tau_{l+1}^e} C_{Wi}^e(t)dt \,\middle|\, \mathbf{Y}(\tau_l^e) = \mathbf{Y}, X_S^e(\tau_l^e) = i\right] \tag{7.4}$$

for $\mathbf{Y} \in \mathscr{E},\ i = 1,\ldots,J$.

The second type of the performance measures is related to the pieces of the c^es' waiting times in the waiting room. Let

$$H_i^e(k) = \int_0^\infty C_{Wi}^e(t)\mathbf{1}\{\kappa(t) = k\}dt, \qquad (i = 1,\ldots,J,\ k \in \Pi \cup \Pi^s). \tag{7.5}$$

$H_i^e(k)$ denotes the c^es' waiting times in the waiting room spent while the system is in period k. For $l = 0,1,2,\ldots$, the expected waiting times after time τ_l^e conditioned on the state of the system are defined by

$$H_i(\mathbf{Y},e,l,k) = \mathrm{E}\left[\int_{\tau_l^e}^\infty C_{Wi}^e(t)\mathbf{1}\{\kappa(t) = k\}dt \,\middle|\, \mathbf{Y}(\tau_l^e) = \mathbf{Y}, X_S^e(\tau_l^e) = i\right], \tag{7.6}$$

$$H_i^0(\mathbf{Y},e,l,k) = \mathrm{E}\left[\int_{\tau_l^e}^{\tau_{l+1}^e} C_{Wi}^e(t)\mathbf{1}\{\kappa(t) = k\}dt \,\middle|\, \mathbf{Y}(\tau_l^e) = \mathbf{Y}, X_S^e(\tau_l^e) = i\right] \tag{7.7}$$

for $i = 1,\ldots,J,\ k \in \Pi \cup \Pi^s,\ \mathbf{Y} \in \mathscr{E}$. Then the following "polling equation" holds.

$$H_i(\mathbf{Y},e,l,k) = \begin{cases} H_i^0(\mathbf{Y},e,l,k) + \mathrm{E}[H_i(\mathbf{Y}(\tau_{l+1}^e),e,l+1,k)|\mathbf{Y}(\tau_l^e) = \mathbf{Y}, X_S^e(\tau_l^e) = i], \\ \qquad \text{if } (\kappa_0 \neq i) \text{ or } (\kappa_0 = i \in \mathscr{H}_g, l = 0) \\ 0, \quad \text{if } (\kappa_0 = i \in \mathscr{H}_e) \text{ or } (\kappa_0 = i \in \mathscr{H}_g, l > 0) \end{cases} \tag{7.8}$$

for $\mathbf{Y} = (\kappa_0, r, \mathbf{g}, \mathbf{n}, L) \in \mathscr{E},\ i = 1,\ldots,J,\ l = 0,1,\ldots,\ k \in \Pi \cup \Pi^s$.

The third type of the performance measures is related to the c^es' waiting times in the service facility. We define for any $t \geq 0$ and $i = 1,\ldots,J$,

$$C_{Fi}^e(t) = \begin{cases} 1, \text{ if } c^e \text{ is in the service facility as an } i\text{-customer and} \\ \quad \text{is not served at time } t \\ 0, \text{ otherwise.} \end{cases} \tag{7.9}$$

The c^es' waiting time in the service facility is defined by

$$F_i^e = \int_0^\infty C_{Fi}^e(t)dt, \qquad (i = 1,\ldots,J). \tag{7.10}$$

The expected waiting times in the service facility after time τ_0^e conditioned on the state of the system are defined by

$$F_i(\mathbf{Y},e)=\mathrm{E}\left[\int_{\tau_0^e}^{\infty}C_{Fi}^e(t)dt\,\middle|\,\mathbf{Y}(\tau_0^e)=\mathbf{Y},X_S^e(\tau_0^e)=i\right] \tag{7.11}$$

for $\mathbf{Y}\in\mathscr{E}$, $i=1,\ldots,J$.

7.3 Expressions for $W_j^0(\cdot),H_j^0(\cdot),F_j(\cdot)$, and Related Quantities

In this section we obtain the conditional expected waiting times $W_j^0(\cdot),H_j^0(\cdot)$, and $F_j(\cdot)$ of a j-customer ($j=1,\ldots,J$). We also consider the expected number of customers at the next polling instant. We observe a specific customer c^e assuming that it is a j-customer ($e=1,2,\ldots$).

7.3.1 Expressions for $W_j^0(\cdot),H_j^0(\cdot)$, and $F_j(\cdot)$

Let $l=0,1,2,\ldots$ and let $\mathbf{Y}=(\kappa_0,r,\mathbf{g},\mathbf{n},L)\in\mathscr{E}$ be the system state at time τ_l^e where $\mathbf{g}=(g_1,\ldots,g_J)$ and $\mathbf{n}=(n_1,\ldots,n_J)$. Because we assume that c^e is a j-customer, $X_S^e(\tau_l^e)=j$. When we consider the c^es' expected waiting time in the waiting room $W_j^0(\mathbf{Y},e,l)$ during the time interval $[\tau_l^e,\tau_{l+1}^e)$, we consider the following cases according to $\kappa_0=\kappa(\tau_l^e)$, which is the period at time τ_l^e. For $\kappa_0\in\mathscr{H}_g$, we have

$$W_j^0(\mathbf{Y},e,l)=\begin{cases}n_{\kappa_0}\mathrm{E}[S_{\kappa_0}]+\sum_{\kappa_1=1}^{J}p_{\kappa_0\kappa_1}\overline{s^o_{\kappa_0\kappa_1}}, & \kappa_0\neq j,\ (l>0)\\ 0, & \kappa_0=j,\ (l>0)\\ r+g_{\kappa_0}\mathrm{E}[S_{\kappa_0}]+\sum_{\kappa_1=1}^{J}p_{\kappa_0\kappa_1}\overline{s^o_{\kappa_0\kappa_1}}, & (l=0).\end{cases} \tag{7.12}$$

For $\kappa_0\in\mathscr{H}_e$, we have

$$W_j^0(\mathbf{Y},e,l)$$
$$=\begin{cases}(n_{\kappa_0}\mathrm{E}[S_{\kappa_0}])/(1-\rho_{\kappa_0})+\sum_{\kappa_1=1}^{J}p_{\kappa_0\kappa_1}\overline{s^o_{\kappa_0\kappa_1}}, & \kappa_0\neq j,\ (l>0)\\ (r+g_{\kappa_0}\mathrm{E}[S_{\kappa_0}])/(1-\rho_{\kappa_0})+\sum_{\kappa_1=1}^{J}p_{\kappa_0\kappa_1}\overline{s^o_{\kappa_0\kappa_1}}, & \kappa_0\neq j,\ (l=0)\\ 0, & \kappa_0=j,\ (l\geq 0).\end{cases} \tag{7.13}$$

For $\kappa_0\in\Pi^s$, we have

$$W_j^0(\mathbf{Y},e,l)=\begin{cases}0, & (l>0)\\ r, & (l=0).\end{cases} \tag{7.14}$$

Because $H_j^0(\mathbf{Y},e,l,k)$ is a piece of the expected waiting time $W_j^0(\mathbf{Y},e,l)$, it is given by appropriately choosing the parts of $W_j^0(\cdot)$. For $\kappa_0 \in \mathscr{H}_g$, we have

$$H_j^0(\mathbf{Y},e,l,k) = \begin{cases} n_{\kappa_0}\mathrm{E}[S_{\kappa_0}], & k=\kappa_0,\ (\kappa_0 \neq j,\ l>0) \\ p_{\kappa_0\kappa_1}\overline{s^o_{\kappa_0\kappa_1}}, & k=(\kappa_0,\kappa_1)\in\Pi^s,\ (\kappa_0\neq j,\ l>0) \\ 0, & (\kappa_0 = j,\ l>0) \\ r+g_{\kappa_0}\mathrm{E}[S_{\kappa_0}], & k=\kappa_0,\ (l=0) \\ p_{\kappa_0\kappa_1}\overline{s^o_{\kappa_0\kappa_1}}, & k=(\kappa_0,\kappa_1)\in\Pi^s,\ (l=0) \\ 0, & \text{otherwise}, \end{cases} \tag{7.15}$$

where $\kappa_1 = 1,\ldots,J$. And for $\kappa_0 \in \mathscr{H}_e$, we have

$$\begin{aligned} &H_j^0(\mathbf{Y},e,l,k) \\ &= \begin{cases} (n_{\kappa_0}\mathrm{E}[S_{\kappa_0}])/(1-\rho_{\kappa_0}), & k=\kappa_0,\ (\kappa_0\neq j,\ l>0) \\ p_{\kappa_0\kappa_1}\overline{s^o_{\kappa_0\kappa_1}}, & k=(\kappa_0,\kappa_1)\in\Pi^s,\ (\kappa_0\neq j,\ l>0) \\ (r+g_{\kappa_0}\mathrm{E}[S_{\kappa_0}])/(1-\rho_{\kappa_0}), & k=\kappa_0,\ (\kappa_0\neq j,\ l=0) \\ p_{\kappa_0\kappa_1}\overline{s^o_{\kappa_0\kappa_1}}, & k=(\kappa_0,\kappa_1)\in\Pi^s,\ (\kappa_0\neq j,\ l=0) \\ 0, & (\kappa_0 = j,\ l\geq 0) \\ 0, & \text{otherwise}, \end{cases} \end{aligned} \tag{7.16}$$

where $\kappa_1 = 1,\ldots,J$. For $\kappa_0 \in \Pi^s$, we have

$$H_j^0(\mathbf{Y},e,l,k) = \begin{cases} 0, & (l>0) \\ r, & k=\kappa_0,\ (l=0) \\ 0, & \text{otherwise},\ (l=0). \end{cases} \tag{7.17}$$

Because $F_j(\mathbf{Y},e)$ is the expected waiting time in the service facility, it is equal to the expected (remaining) service times of customers at station j at the c^es' arrival epoch τ_0^e. Then we have

$$F_j(\mathbf{Y},e) = \begin{cases} n_j\mathrm{E}[S_j], & j\in\mathscr{H}_g \\ n_j\mathrm{E}[S_j], & j\in\mathscr{H}_e \text{ and } j\neq\kappa_0 \\ r+g_j\mathrm{E}[S_j], & j\in\mathscr{H}_e \text{ and } j=\kappa_0. \end{cases} \tag{7.18}$$

7.3.2 System State at the Next Polling Instant

Let $l = 0, 1, 2, \ldots$ and let $\mathbf{Y} = (\kappa_0, r, \mathbf{g}, \mathbf{n}, L) \in \mathscr{E}$ be the system state at time τ_l^e where $\mathbf{g} = (g_1, \ldots, g_J)$ and $\mathbf{n} = (n_1, \ldots, n_J)$. We consider the system state at the next polling instant τ_{l+1}^e.

When we consider the system state (especially, the numbers of customers) at the next polling instant, we consider the following cases according to $\kappa_0 = \kappa(\tau_l^e)$, which is the period at time τ_l^e. For $\kappa_0 \in \mathscr{H}_g$, we can show that

$$\begin{aligned}
&\mathrm{E}[n_m(\tau_{l+1}^e) | \kappa(\tau_{l+1}^e) = \kappa_1, \mathbf{Y}(\tau_l^e) = \mathbf{Y}, X_S^e(\tau_l^e) = j] \\
&\quad = \begin{cases} n_m + \lambda_m \{ n_{\kappa_0} \mathrm{E}[S_{\kappa_0}] + \overline{s^o_{\kappa_0 \kappa_1}} \}, & m \neq \kappa_0, \ (l > 0) \\ \lambda_{\kappa_0} \{ n_{\kappa_0} \mathrm{E}[S_{\kappa_0}] + \overline{s^o_{\kappa_0 \kappa_1}} \}, & m = \kappa_0, \ (l > 0) \\ n_m + \mathbf{1}_{mj} + \lambda_m \{ r + g_{\kappa_0} \mathrm{E}[S_{\kappa_0}] + \overline{s^o_{\kappa_0 \kappa_1}} \}, & (l = 0) \end{cases}
\end{aligned} \tag{7.19}$$

for any $m, j, \kappa_1 \in \Pi$, where $\mathbf{1}_{mj} = \mathbf{1}\{m = j\}$. For $\kappa_0 \in \mathscr{H}_e$, we have

$$\begin{aligned}
&\mathrm{E}[n_m(\tau_{l+1}^e) | \kappa(\tau_{l+1}^e) = \kappa_1, \mathbf{Y}(\tau_l^e) = \mathbf{Y}, X_S^e(\tau_l^e) = j] \\
&= \begin{cases} n_m + \lambda_m \{ (n_{\kappa_0} \mathrm{E}[S_{\kappa_0}]) / (1 - \rho_{\kappa_0}) + \overline{s^o_{\kappa_0 \kappa_1}} \}, & m \neq \kappa_0, \ (l > 0) \\ n_m + \mathbf{1}_{mj} + \lambda_m \{ (r + (g_{\kappa_0} + \mathbf{1}_{\kappa_0 j}) \mathrm{E}[S_{\kappa_0}]) / (1 - \rho_{\kappa_0}) + \overline{s^o_{\kappa_0 \kappa_1}} \}, & \\ \qquad\qquad m \neq \kappa_0, \ (l = 0) & \\ \lambda_{\kappa_0} \overline{s^o_{\kappa_0 \kappa_1}}, & m = \kappa_0, \ (l \geq 0) \end{cases}
\end{aligned} \tag{7.20}$$

for any $m, j, \kappa_1 \in \Pi$. For $\kappa_0 \in \Pi^s$, we have

$$\begin{aligned}
&\mathrm{E}[n_m(\tau_{l+1}^e) | \kappa(\tau_{l+1}^e) = \kappa_1, \mathbf{Y}(\tau_l^e) = \mathbf{Y}, X_S^e(\tau_l^e) = j] \\
&= \begin{cases} 0, & (l > 0) \\ n_m + \mathbf{1}_{mj} + \lambda_m r, & (l = 0). \end{cases}
\end{aligned} \tag{7.21}$$

Furthermore for any $m, j, \kappa_1 \in \Pi$, we obviously have

$$\mathrm{E}[g_m(\tau_{l+1}^e) | \kappa(\tau_{l+1}^e) = \kappa_1, \mathbf{Y}(\tau_l^e) = \mathbf{Y}, X_S^e(\tau_l^e) = j] = 0. \tag{7.22}$$

7.3.3 Unified Forms: Linear Functional Expressions

From the analysis in this section, we can easily see the following important properties.

- The component $(\kappa_0, r, \mathbf{g}, \mathbf{n})$ of state $\mathbf{Y} = (\kappa_0, r, \mathbf{g}, \mathbf{n}, L) \in \mathscr{E}$ at epoch τ_l^e is sufficient to derive $W_i^0(\mathbf{Y}, e, l), H_j^0(\mathbf{Y}, e, l, k), F_j(\mathbf{Y}, e)$, and the conditional expected numbers of customers at time τ_{l+1}^e.
- These quantities are linear with respect to r and $(\mathbf{g}, \mathbf{n}) = (g_1, \ldots, g_J, n_1, \ldots, n_J)$.

For convenience, let $\mathbf{e}_j = (0,\ldots,0,\underbrace{1}_{j^{th}\text{ place}},0,\ldots,0) \in \mathscr{R}^{1\times J}$, and let $p_k = p_{\kappa_0,\kappa_1}$ for $k = (\kappa_0,\kappa_1) \in \Pi^s$. Then we have the following.

Proposition 7.1. *Let* $\mathbf{Y} = (\kappa_0, r, \mathbf{g}, \mathbf{n}, L) \in \mathscr{E}$, $j = 1,\ldots,J$, $e = 1,2,\ldots$, $l = 0,1,2,\ldots$ *and* $k \in \Pi \cup \Pi^s$. *Then we have*

$$H_j^0(\mathbf{Y},e,l,k)$$

$$= \begin{cases} (\mathbf{g},\mathbf{n})\mathbf{h}_{10}^0(\kappa_0,j,k), & \kappa_0 \in \Pi,\ l>0,\ k \in \Pi \\ p_k h_{11}^0(\kappa_0,j,k), & \kappa_0 \in \Pi,\ l>0,\ k \in \Pi^s \\ r\varphi^0(\kappa_0,j,k) + (\mathbf{g},\mathbf{n})\mathbf{h}_{00}^0(\kappa_0,j,k), & \kappa_0 \in \Pi,\ l=0,\ k \in \Pi \\ p_k h_{01}^0(\kappa_0,j,k), & \kappa_0 \in \Pi,\ l=0,\ k \in \Pi^s \\ 0, & \kappa_0 \in \Pi^s,\ l>0,\ k \in \Pi \cup \Pi^s \\ 0, & \kappa_0 \in \Pi^s,\ l=0,\ k \in \Pi \\ r\varphi^0(\kappa_0,j,k), & \kappa_0 \in \Pi^s,\ l=0,\ k \in \Pi^s, \end{cases} \tag{7.23}$$

$$F_j(\mathbf{Y},e) = r\psi(\kappa_0,j) + (\mathbf{g},\mathbf{n})\mathbf{f}(\kappa_0,j), \tag{7.24}$$

where the above coefficients

$$\mathbf{h}_{a0}^0(\kappa_0,j,k) \in \mathscr{R}^{2J\times 1}, \qquad h_{a1}^0(\kappa_0,j,k) \in \mathscr{R}, \qquad (a=0,1),$$
$$\varphi^0(\kappa_0,j,k) \in \mathscr{R}, \qquad \psi(\kappa_0,j) \in \mathscr{R}, \qquad \mathbf{f}(\kappa_0,j) \in \mathscr{R}^{2J\times 1}$$

can be determined from the given system parameters through the expressions obtained in this section. Furthermore we have

$$\mathrm{E}[(\mathbf{g}(\tau_{l+1}^e),\mathbf{n}(\tau_{l+1}^e))|\kappa(\tau_{l+1}^e)=\kappa_1,\mathbf{Y}(\tau_l^e)=\mathbf{Y},X_S^e(\tau_l^e)=j]$$

$$= \begin{cases} (\mathbf{g},\mathbf{n})\mathbf{U}_1(\kappa_0) + \mathbf{u}_1(\kappa_0,\kappa_1), & \kappa_0 \in \Pi,\ l>0 \\ r\upsilon(\kappa_0) + (\mathbf{g},\mathbf{n})\mathbf{U}_0(\kappa_0) + \mathbf{u}_0(j,\kappa_0,\kappa_1), & \kappa_0 \in \Pi,\ l=0 \\ \mathbf{0}, & \kappa_0 \in \Pi^s,\ l>0 \\ r\upsilon + (\mathbf{g},\mathbf{n})\mathbf{U}_0 + (\mathbf{0},\mathbf{e}_j), & \kappa_0 \in \Pi^s,\ l=0 \end{cases} \tag{7.25}$$

for $\kappa_1 \in \Pi$. *The above coefficients*

$$\mathbf{U}_1(\kappa_0) \in \mathscr{R}^{2J\times 2J}, \qquad \mathbf{u}_1(\kappa_0,\kappa_1) \in \mathscr{R}^{1\times 2J}, \qquad \upsilon(\kappa_0) \in \mathscr{R}^{1\times 2J},$$
$$\mathbf{U}_0(\kappa_0) \in \mathscr{R}^{2J\times 2J}, \qquad \mathbf{u}_0(j,\kappa_0,\kappa_1) \in \mathscr{R}^{1\times 2J}, \qquad \upsilon \in \mathscr{R}^{1\times 2J}, \qquad \mathbf{U}_0 \in \mathscr{R}^{2J\times 2J}$$

can be determined from the given system parameters through the expressions obtained in this section.

Note 1. We can simplify the expression for $H_j^0(\cdot)$ as follows:

$$H_j^0(\mathbf{Y},e,l,k) = \begin{cases} (\mathbf{g},\mathbf{n})\mathbf{h}_{10}^0(\kappa_0,j,k) + p_k h_{11}^0(\kappa_0,j,k), & l>0 \\ r\varphi^0(\kappa_0,j,k) + (\mathbf{g},\mathbf{n})\mathbf{h}_{00}^0(\kappa_0,j,k) + p_k h_{01}^0(\kappa_0,j,k), & l=0. \end{cases}$$

Because this expression introduces much labor into the numerical calculation, we adopt the above somewhat complicated expression. A similar result holds for the expression in Equation (7.25). □

Note 2. It should be noted from Equations (7.15) and (7.16) that

$$H_j^0(\mathbf{Y},e,l,k) = 0,$$
$$(j \in \Pi,\ \mathbf{Y} = (\kappa_0,r,\mathbf{g},\mathbf{n},L) \in \mathscr{E},\ e = 1,2,\ldots,\ l \geq 0,\ k \in \Pi \cup \Pi^s)$$

if $(\kappa_0 = j \in \mathscr{H}_e)$ or $(\kappa_0 = j \in \mathscr{H}_g$ and $l > 0)$. □

7.4 The Linear Functional Expression

In this section we obtain the expression for the performance measure $H_j(\cdot)$ by solving the polling equation. It can be shown that it has the linear functional form.

We define constants $\mathbf{h}_{10}(\kappa_0,j,k) \in \mathscr{R}^{2J\times 1}$ and $h_{11}(\kappa_0,j,k) \in \mathscr{R}$ that satisfy the following equations:

$$\mathbf{h}_{10}(\kappa_0,j,k) = \begin{cases} \mathbf{h}_{10}^0(\kappa_0,j,k) + \mathbf{U}_1(\kappa_0) \displaystyle\sum_{\kappa_1 \in \Pi\setminus\{j\}} p_{\kappa_0\kappa_1}\mathbf{h}_{10}(\kappa_1,j,k), \\ \qquad\qquad \kappa_0 \neq j,\ \kappa_0 \in \Pi,\ k \in \Pi \\ \mathbf{0}, \qquad \text{otherwise}, \end{cases} \tag{7.26}$$

$$h_{11}(\kappa_0,j,k) = \begin{cases} \displaystyle\sum_{\kappa_1 \in \Pi\setminus\{j\}} p_{\kappa_0\kappa_1}\mathbf{u}_1(\kappa_0,\kappa_1)\mathbf{h}_{10}(\kappa_1,j,k) \\ \qquad + \displaystyle\sum_{\kappa_1 \in \Pi\setminus\{j\}} p_{\kappa_0\kappa_1}h_{11}(\kappa_1,j,k), \\ \qquad\qquad \kappa_0 \neq j,\ \kappa_0 \in \Pi,\ k \in \Pi \\ p_k h_{11}^0(\kappa_0,j,k) + \displaystyle\sum_{\kappa_1 \in \Pi\setminus\{j\}} p_{\kappa_0\kappa_1}h_{11}(\kappa_1,j,k), \\ \qquad\qquad \kappa_0 \neq j,\ \kappa_0 \in \Pi,\ k \in \Pi^s \\ 0, \qquad \text{otherwise} \end{cases} \tag{7.27}$$

for $j \in \Pi$. Furthermore for $k \in \Pi$, $\kappa_0 \in \Pi \cup \Pi^s$, and $j \in \Pi$, let[5]

[5] Case 1: $(\kappa_0 \neq j$ or $j \in \mathscr{H}_g)$ and $(\kappa_0 \in \Pi)$; Case 2: $\kappa_0 = j \in \mathscr{H}_e$; Case 3: $\kappa_0 = (k_0,k_1) \in \Pi^s$.

$$\varphi(\kappa_0,j,k)=\begin{cases}\varphi^0(\kappa_0,j,k)+\upsilon(\kappa_0)\displaystyle\sum_{\kappa_1\in\Pi\setminus\{j\}}p_{\kappa_0\kappa_1}\mathbf{h}_{10}(\kappa_1,j,k), & \text{case 1}\\ 0, & \text{case 2}\\ \upsilon\mathbf{h}_{10}(k_1,j,k), & \text{case 3},\end{cases}$$

$$\mathbf{h}_{00}(\kappa_0,j,k)=\begin{cases}\mathbf{h}^0_{00}(\kappa_0,j,k)+\mathbf{U}_0(\kappa_0)\displaystyle\sum_{\kappa_1\in\Pi\setminus\{j\}}p_{\kappa_0\kappa_1}\mathbf{h}_{10}(\kappa_1,j,k), & \text{case 1}\\ \mathbf{0}, & \text{case 2}\\ \mathbf{U}_0\mathbf{h}_{10}(k_1,j,k), & \text{case 3},\end{cases}$$

$$h_{01}(\kappa_0,j,k)=\begin{cases}\displaystyle\sum_{\kappa_1\in\Pi\setminus\{j\}}p_{\kappa_0\kappa_1}\{\mathbf{u}_0(j,\kappa_0,\kappa_1)\mathbf{h}_{10}(\kappa_1,j,k)+h_{11}(\kappa_1,j,k)\}, & \text{case 1}\\ 0, & \text{case 2}\\ (\mathbf{0},\mathbf{e}_j)\mathbf{h}_{10}(k_1,j,k)+h_{11}(k_1,j,k), & \text{case 3}.\end{cases}$$

And for $k\in\Pi^s$, $\kappa_0\in\Pi\cup\Pi^s$ and $j\in\Pi$, let

$$\varphi(\kappa_0,j,k)=\begin{cases}0, & \kappa_0\in\Pi\\ \varphi^0(\kappa_0,j,k), & \kappa_0\in\Pi^s,\end{cases}$$

$$\mathbf{h}_{00}(\kappa_0,j,k)=\mathbf{0},$$

$$h_{01}(\kappa_0,j,k)=\begin{cases}p_k h^0_{01}(\kappa_0,j,k)+\displaystyle\sum_{\kappa_1\in\Pi\setminus\{j\}}p_{\kappa_0\kappa_1}h_{11}(\kappa_1,j,k), & \text{case 1}\\ 0, & \text{case 2}\\ h_{11}(k_1,j,k), & \text{case 3}.\end{cases}$$

Now we define the following function, and show that it gives the linear functional expression for the performance measure $H_j(\cdot)$ defined by (7.6).

Definition 7.1. The linear function is defined by

$$\hat{H}_j(\mathbf{Y},e,l,k)=\begin{cases}r\varphi(\kappa_0,j,k)+(\mathbf{g},\mathbf{n})\mathbf{h}_{00}(\kappa_0,j,k)+h_{01}(\kappa_0,j,k), & l=0,\ k\in\Pi\\ r\varphi(\kappa_0,j,k)+h_{01}(\kappa_0,j,k), & l=0,\ k\in\Pi^s\\ (\mathbf{g},\mathbf{n})\mathbf{h}_{10}(\kappa_0,j,k)+h_{11}(\kappa_0,j,k), & l>0,\ k\in\Pi\\ h_{11}(\kappa_0,j,k), & l>0,\ k\in\Pi^s\end{cases}\tag{7.28}$$

for any $j\in\Pi$; $\mathbf{Y}=(\kappa_0,r,\mathbf{g},\mathbf{n},L)\in\mathscr{E}$; $e=1,2,\ldots$; $l=0,1,2,\ldots$ and $k\in\Pi\cup\Pi^s$.

Proposition 7.2. *The function $\hat{H}_\cdot(\cdot,\cdot,\cdot,k)$ $(k\in\Pi\cup\Pi^s)$ defined by (7.28) satisfies the "polling equation" (7.8).*

Proof. See the appendix. □

Proposition 7.3. *The solution of the "polling equation" (7.8) is unique and hence*

$$H_j(\mathbf{Y},e,l,k) = \hat{H}_j(\mathbf{Y},e,l,k),$$
$$(j \in \Pi;\ \mathbf{Y} \in \mathscr{E};\ e = 1,2,\ldots;\ l = 0,1,2,\ldots;\ k \in \Pi \cup \Pi^s).$$

Proof. Because the proof of this proposition is similar to the proof of uniqueness of the solution for the feedback equation given in [1], it is omitted. □

7.5 Steady-State Values

We would like to obtain the steady-state values of the performance measures. We define the mean waiting time of j-customers[6] as follows:

$$\bar{w}_j = \lim_{N\to\infty} \frac{1}{N} \sum_{e=1}^{N} \mathrm{E}[W_j^e + F_j^e | X_S^e(\tau_0^e) = j], \qquad j = 1,\ldots,J. \tag{7.29}$$

In order to obtain the quantity, we define the following interim quantities:

$$\bar{H}_j(\kappa_0,k) = \lim_{N\to\infty} \frac{1}{N} \sum_{e=1}^{N} \mathrm{E}[H_j^e(k)\mathbf{1}\{\kappa(\tau_0^e) = \kappa_0\} | X_S^e(\tau_0^e) = j], \tag{7.30}$$

$$\bar{F}_j(\kappa_0) = \lim_{N\to\infty} \frac{1}{N} \sum_{e=1}^{N} \mathrm{E}[F_j^e \mathbf{1}\{\kappa(\tau_0^e) = \kappa_0\} | X_S^e(\tau_0^e) = j] \tag{7.31}$$

for $j \in \Pi$ and $\kappa_0, k \in \Pi \cup \Pi^s$. The time average values of the system state are defined by

$$\tilde{\mathbf{Y}}^k = (k\tilde{q}^k, \tilde{r}^k, \tilde{\mathbf{g}}^k, \tilde{\mathbf{n}}^k, \tilde{L}^k) = \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathrm{E}[\mathbf{Y}(s)\mathbf{1}\{\kappa(s) = k\}]ds \tag{7.32}$$

for $k \in \Pi \cup \Pi^s$ where $\tilde{\mathbf{g}}^k = (\tilde{g}_1^k,\ldots,\tilde{g}_J^k), \tilde{\mathbf{n}}^k = (\tilde{n}_1^k,\ldots,\tilde{n}_J^k)$.

For $k \in \Pi$, the steady-state value $\tilde{q}^k$, which is the long-run fraction of time that the system is in period k, is calculated as

$$\tilde{q}^k = \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathrm{E}[\mathbf{1}\{\kappa(s) = k\}]ds = \lambda_k \mathrm{E}[S_k]. \tag{7.33}$$

For $k \in \Pi^s$, the steady-state value $\tilde{q}^k$ can be obtained in the following manner. Let π_i be the steady-state probability that the server selects station i at a polling instant. It can be easily shown that $\pi = (\pi_1,\ldots,\pi_J)$ is the steady-state probability of the Markov chain with the transition probability matrix $\mathbf{P}$. We can obtain it by solving $\pi\mathbf{P} = \pi$ and $\pi\mathbf{1} = 1$. Then the long-run fraction of time that the server is moving from station i to station j given that the system is in a switchover period is given by

[6] The time average values and the customer average values defined in this section are assumed to exist.

$$\frac{\pi_i p_{ij} \overline{s_{ij}^o}}{\sum_{i=1}^{J} \sum_{j=1}^{J} \pi_i p_{ij} \overline{s_{ij}^o}}, \qquad (i,j) \in \Pi^s. \tag{7.34}$$

Furthermore the long-run fraction of time that the system is in a switchover period is $1-\rho$. Hence we obtain

$$\tilde{q}^{(i,j)} = (1-\rho) \frac{\pi_i p_{ij} \overline{s_{ij}^o}}{\sum_{i=1}^{J} \sum_{j=1}^{J} \pi_i p_{ij} \overline{s_{ij}^o}}, \qquad (i,j) \in \Pi^s. \tag{7.35}$$

The expected remaining service time of a customer being served given that the current period is $k \in \Pi$ is equal to $\overline{s_k^2}/(2\mathrm{E}[S_k])$, and the expected value of the remaining switchover period given that the current period is $(i,j) \in \Pi^s$ is equal to $\overline{s_{ij}^{o2}}/(2\overline{s_{ij}^o})$. Then we have

$$\tilde{r}^k = \left(\frac{\overline{s_k^2}}{2\mathrm{E}[S_k]} \right) \tilde{q}^k = \frac{\lambda_k \overline{s_k^2}}{2}, \quad k \in \Pi, \tag{7.36}$$

$$\tilde{r}^{(i,j)} = \left(\frac{\overline{s_{ij}^{o2}}}{2\overline{s_{ij}^o}} \right) \tilde{q}^{(i,j)} = (1-\rho) \frac{\pi_i p_{ij} \overline{s_{ij}^{o2}}}{2 \sum_{i=1}^{J} \sum_{j=1}^{J} \pi_i p_{ij} \overline{s_{ij}^o}}, \quad (i,j) \in \Pi^s. \tag{7.37}$$

From the results in the previous sections and the PASTA property, we have

$$\bar{H}_j(\kappa_0, k) = \begin{cases} \tilde{r}^{\kappa_0} \varphi(\kappa_0, j, k) + (\tilde{\mathbf{g}}^{\kappa_0}, \tilde{\mathbf{n}}^{\kappa_0}) \mathbf{h}_{00}(\kappa_0, j, k) + \tilde{q}^{\kappa_0} h_{01}(\kappa_0, j, k), & k \in \Pi \\ \tilde{r}^{\kappa_0} \varphi(\kappa_0, j, k) + \tilde{q}^{\kappa_0} h_{01}(\kappa_0, j, k), & k \in \Pi^s, \end{cases} \tag{7.38}$$

$$\bar{F}_j(\kappa_0) = \tilde{r}^{\kappa_0} \psi(\kappa_0, j) + (\tilde{\mathbf{g}}^{\kappa_0}, \tilde{\mathbf{n}}^{\kappa_0}) \mathbf{f}(\kappa_0, j) \tag{7.39}$$

for $j \in \Pi$ and $\kappa_0 \in \Pi \cup \Pi^s$. Then from the generalized version of Little's formula ($H = \lambda G$) [27], we have

$$\begin{aligned} \tilde{n}_j^k &= \lambda_j \sum_{\kappa_0 \in \Pi \cup \Pi^s} \bar{H}_j(\kappa_0, k), \\ \tilde{g}_j &= \lambda_j \sum_{\kappa_0 \in \Pi \cup \Pi^s} \bar{F}_j(\kappa_0), \end{aligned} \qquad j \in \Pi \quad \text{and} \quad k \in \Pi \cup \Pi^s, \tag{7.40}$$

where $\tilde{g}_j = \sum_{k \in \Pi \cup \Pi^s} \tilde{g}_j^k$. Furthermore it can be shown that

$$\tilde{g}_j^k = \begin{cases} \tilde{g}_j, & k = j, \\ 0, & k \neq j, \end{cases} \qquad j \in \Pi \quad \text{and} \quad k \in \Pi \cup \Pi^s. \tag{7.41}$$

Then we obtain the following set of linear equations for the average numbers of customers in the system.

$$\tilde{n}_j^k = \begin{cases} \lambda_j \displaystyle\sum_{\kappa_0 \in \Pi \cup \Pi^s} \{\tilde{r}^{\kappa_0} \varphi(\kappa_0, j, k) + \tilde{q}^{\kappa_0} h_{01}(\kappa_0, j, k) \\ \qquad\qquad + (\tilde{\mathbf{g}}^{\kappa_0}, \tilde{\mathbf{n}}^{\kappa_0}) \mathbf{h}_{00}(\kappa_0, j, k)\}, & k \in \Pi \\ \lambda_j \displaystyle\sum_{\kappa_0 \in \Pi \cup \Pi^s} \{\tilde{r}^{\kappa_0} \varphi(\kappa_0, j, k) + \tilde{q}^{\kappa_0} h_{01}(\kappa_0, j, k)\}, & k \in \Pi^s, \end{cases} \tag{7.42}$$

$$\tilde{g}_j^k = \begin{cases} \lambda_j \displaystyle\sum_{\kappa_0 \in \Pi \cup \Pi^s} \{\tilde{r}^{\kappa_0} \psi(\kappa_0, j) + (\tilde{\mathbf{g}}^{\kappa_0}, \tilde{\mathbf{n}}^{\kappa_0}) \mathbf{f}(\kappa_0, j)\}, & k = j \\ 0, & k \neq j \text{ or } k \in \Pi^s \end{cases} \tag{7.43}$$

for $j \in \Pi$ and $k \in \Pi \cup \Pi^s$. Then we finally obtain the following proposition.

Proposition 7.4. *The mean waiting time of j-customers ($j = 1, \ldots, J$) is given by*

$$\bar{w}_j = \sum_{\kappa_0 \in \Pi \cup \Pi^s} \left\{ \sum_{k \in \Pi \cup \Pi^s} \bar{H}_j(\kappa_0, k) + \bar{F}_j(\kappa_0) \right\} = \frac{1}{\lambda_j} \left(\tilde{g}_j^j + \sum_{k \in \Pi \cup \Pi^s} \tilde{n}_j^k \right), \tag{7.44}$$

where $\tilde{g}_j^j$ and $\tilde{n}_j^k$ ($j \in \Pi$; $k \in \Pi \cup \Pi^s$) can be obtained by solving the set of (7.42) and (7.43).

7.6 Computational Complexity

We now evaluate the computational complexity to calculate the mean waiting times. In Sect. 7.4 calculation of the coefficients $\mathbf{h}_{10}(\kappa_0, j, k)$ $(\kappa_0, j, k \in \Pi)$ takes much time. Then from (7.26) we have

$$\begin{pmatrix} \mathbf{h}_{10}(1, j, k) \\ \mathbf{h}_{10}(2, j, k) \\ \vdots \\ \mathbf{h}_{10}(J, j, k) \end{pmatrix} = (\mathbf{I} - \mathbf{I}(j)\mathbf{U}\mathbf{Q})^{-1} \mathbf{I}(j) \begin{pmatrix} \mathbf{h}_{10}^0(1, j, k) \\ \mathbf{h}_{10}^0(2, j, k) \\ \vdots \\ \mathbf{h}_{10}^0(J, j, k) \end{pmatrix},$$

where $\mathbf{I} \in \mathscr{R}^{2J^2 \times 2J^2}$ and $\mathbf{I}_0 \in \mathscr{R}^{2J \times 2J}$ are identity matrices, and where

$$\mathbf{I}(j) = \mathrm{diag}(\mathbf{I}_0, \ldots, \mathbf{I}_0, \underbrace{\mathbf{O}}_{j^{th} \text{ place}}, \mathbf{I}_0, \ldots, \mathbf{I}_0) \in \mathscr{R}^{2J^2 \times 2J^2},$$

$$\mathbf{Q} = (p_{i,j}\mathbf{I}_0 : i, j = 1, \ldots, J) \in \mathscr{R}^{2J^2 \times 2J^2},$$

$$\mathbf{U} = \mathrm{diag}(\mathbf{U}_1(j) : j = 1, \ldots, J) \in \mathscr{R}^{2J^2 \times 2J^2}.$$

The calculation of the J inverse matrices $(\mathbf{I}-\mathbf{I}(j)\mathbf{U}\mathbf{Q})^{-1}\mathbf{I}(j) \in \mathscr{R}^{2J^2\times 2J^2}$, $(j \in \Pi)$ takes $J \times O(J^6)$ numerical operations,[7] and the whole calculation of the products of the inverse matrices and the right end vectors of vectors $\{\mathbf{h}^0_{10}(\kappa_0, j, k) : \kappa_0 \in \Pi\}$, $(j,k \in \Pi)$ take $O(J^6)$ numerical operations. The calculations of the other constants in Sect. 7.4 take at most $O(J^5)$ numerical operations. In Sect. 7.5 it takes much time to solve (7.42) and (7.43). Because the set of these equations essentially has $J(J+1)$ unknowns, $O(J^6)$ numerical operations are required in order to solve them. The other calculations in this section take at most $O(J^5)$ numerical operations. Hence the overall complexity of our method is primarily $O(J^7)$ numerical operations.

The primal algorithm has somewhat excessive computational complexity, therefore we would like to reduce it. As noted above, much of the computational complexity comes from the calculations of the constants $\mathbf{h}_{10}(\kappa_0, j, k)$ and the calculations of the steady-state values from (7.42) and (7.43).

7.6.1 *Reduction of Calculations of* $\mathbf{h}_{10}(\cdot)$

This reduction has three steps.

First Reduction Step:

We can reduce the computational complexity by checking the following facts. Because it can be shown from (7.15) and (7.16) that $H^0_j(\mathbf{Y}, e, l, k)$ for $\kappa_0, k \in \Pi$ and $l > 0$ is not affected by the vector $\mathbf{g}$ of the numbers of customers in the service facility, the elements in the upper half of $\mathbf{h}^0_{10}(\kappa_0, j, k)$ in (7.23) are 0. Similarly, because it can be shown from (7.19) and (7.20) that the conditional expectation of $\mathbf{n}(\tau^e_{l+1})$ for $\kappa_0 \in \Pi$ and $l > 0$ is not affected by $\mathbf{g}$, and because $\mathbf{g}(\tau^e_{l+1}) = \mathbf{0}$, the elements in the upper half and the left half of $\mathbf{U}_1(\kappa_0)$ in (7.25) are 0. That is, we have

$$\mathbf{h}^0_{10}(\kappa_0, j, k) = \begin{pmatrix} \mathbf{0} \\ \mathbf{h}^{0*}_{10}(\kappa_0, j, k) \end{pmatrix}, \qquad \mathbf{U}_1(\kappa_0) = \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{U}^*_1(\kappa_0) \end{pmatrix},$$

where $\mathbf{h}^{0*}_{10}(\kappa_0, j, k) \in \mathscr{R}^{J\times 1}$, $\mathbf{U}^*_1(\kappa_0) \in \mathscr{R}^{J\times J}$, and then the size of (7.26) can be reduced by half. Let $\mathbf{h}^*_{10}(\kappa_0, j, k) \in \mathscr{R}^{J\times 1}$ be the vector composed of the lower half elements of $\mathbf{h}_{10}(\kappa_0, j, k)$, and we have the following reduced version of (7.26).

$$\mathbf{h}^*_{10}(\kappa_0, j, k) = \begin{cases} \mathbf{h}^{0*}_{10}(\kappa_0, j, k) + \mathbf{U}^*_1(\kappa_0) \displaystyle\sum_{\kappa_1 \in \Pi\setminus\{j\}} p_{\kappa_0\kappa_1} \mathbf{h}^*_{10}(\kappa_1, j, k), \\ \qquad\qquad \kappa_0 \neq j,\ \kappa_0 \in \Pi,\ k \in \Pi \\ \mathbf{0}, \qquad\qquad \text{otherwise} \end{cases} \tag{7.45}$$

for $\kappa_0, k \in \Pi \cup \Pi^s$ and $j \in \Pi$.

[7] For simplicity, we estimate that an $n \times n$ matrix can be inverted in $O(n^3)$ numerical operations.

Second Reduction Step:

We reduce the calculations by using sparsity of the constants $\mathbf{h}_{10}^{0*}(\kappa_0, j, k)$ and $\mathbf{U}_1^*(\kappa_0)$ in (7.45). From (7.15), (7.16), and (7.23), for $\kappa_0 \in \Pi,\ l > 0,\ k \in \Pi$, we have

$$\mathbf{h}_{10}^{0*}(\kappa_0, j, k) = \begin{cases} \mathbf{e}'_{\kappa_0}\delta(\kappa_0), & \kappa_0 = k,\ \kappa_0 \neq j \\ \mathbf{0}, & \text{otherwise} \end{cases} \quad (j \in \Pi), \tag{7.46}$$

where $\mathbf{e}'_{\kappa_0}$ is the transpose of $\mathbf{e}_{\kappa_0} = (0, \ldots, 0, \underbrace{1}_{\kappa_0^{th}\text{ place}}, 0, \ldots, 0)$ defined in Sect. 7.3, and

where

$$\delta(\kappa_0) = \begin{cases} \mathrm{E}[S_{\kappa_0}], & \kappa_0 \in \mathscr{H}_g \\ \mathrm{E}[S_{\kappa_0}]/(1-\rho_{\kappa_0}), & \kappa_0 \in \mathscr{H}_e. \end{cases}$$

From (7.19), (7.20), and (7.25), it can be shown that

$$\mathbf{U}_1^*(\kappa_0) = \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ u_1^*(\kappa_0) & \cdots & u_{\kappa_0-1}^*(\kappa_0) & u_{\kappa_0}^*(\kappa_0) & u_{\kappa_0+1}^*(\kappa_0) & \cdots & u_J^*(\kappa_0) \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{pmatrix},$$

where $u_m^*(\kappa_0) = \lambda_m \delta(\kappa_0)\mathbf{1}\{m \neq \kappa_0 \text{ or } \kappa_0 \in \mathscr{H}_g\}$, $(m, \kappa_0 \in \Pi)$.

Let

$$\xi_{10}^*(\kappa_0, j, k) = \begin{cases} \sum_{\kappa_1 \in \Pi \setminus \{j\}} p_{\kappa_0\kappa_1}\mathbf{h}_{10}^*(\kappa_1, j, k), & \kappa_0 \neq j \\ \mathbf{0}, & \kappa_0 = j \end{cases} \quad (\kappa_0, j, k \in \Pi).$$

Then from (7.45) and (7.46), we have

$$\mathbf{h}_{10}^*(\kappa_0, j, k) = \mathbf{e}'_{\kappa_0}\delta(\kappa_0)\mathbf{1}\{\kappa_0 = k, \kappa_0 \neq j\} + \mathbf{U}_1^*(\kappa_0)\xi_{10}^*(\kappa_0, j, k)$$

for $\kappa_0, j, k \in \Pi$. Now we define the following notation.

- For any vector $\mathbf{a}$, let $\mathbf{a}|_m$ be its mth element.

Then we have

$$\mathbf{h}_{10}^*(\kappa_0, j, k)|_m = \begin{cases} \xi_{10}^*(\kappa_0, j, k)|_m, & m \neq \kappa_0 \\ \delta(\kappa_0)\mathbf{1}\{\kappa_0 = k,\ \kappa_0 \neq j\} + \sum_{l=1}^{J} u_l^*(\kappa_0)\xi_{10}^*(\kappa_0, j, k)|_l, & m = \kappa_0, \end{cases} \tag{7.47}$$

$$\xi^*_{10}(\kappa_0,j,k)|_m \tag{7.48}$$
$$= \begin{cases} \sum_{\kappa_1 \in \Pi\setminus\{j,m\}} p_{\kappa_0\kappa_1}\xi^*_{10}(\kappa_1,j,k)|_m + p_{\kappa_0 m}\mathbf{h}^*_{10}(m,j,k)|_m, & m\neq j,\ \kappa_0\neq j \\ \sum_{\kappa_1 \in \Pi\setminus\{j\}} p_{\kappa_0\kappa_1}\xi^*_{10}(\kappa_1,j,k)|_j, & m=j,\ \kappa_0\neq j \\ 0, & \kappa_0=j \end{cases}$$

for $m=1,\ldots,J,\ \kappa_0,j,k\in\Pi$. Let

$$\xi^*_{10}(j,k)_m = \begin{pmatrix} \xi^*_{10}(1,j,k)|_m \\ \vdots \\ \xi^*_{10}(J,j,k)|_m \end{pmatrix} \in \mathscr{R}^{J\times 1}, \qquad \mathbf{p}(j)_m = \begin{pmatrix} p_{1,m} \\ \vdots \\ p_{j-1,m} \\ 0 \\ p_{j+1,m} \\ \vdots \\ p_{J,m} \end{pmatrix} \in \mathscr{R}^{J\times 1},$$

$$\mathbf{I}_0(m) = \mathrm{diag}(1,\cdots,1,\underbrace{0}_{m^{th}\text{ place}},1,\cdots,1) \in \mathscr{R}^{J\times J},$$

$$\mathbf{P}(j) = \left(\mathbf{p}(j)_1,\cdots,\mathbf{p}(j)_{j-1},\mathbf{0},\mathbf{p}(j)_{j+1},\cdots,\mathbf{p}(j)_J\right) \in \mathscr{R}^{J\times J}$$

for $m=1,\ldots,J,\ j,k\in\Pi$. Then from (7.48), we have

$$\xi^*_{10}(j,k)_m = \begin{cases} \mathbf{P}(j)\mathbf{I}_0(m)\xi^*_{10}(j,k)_m + \mathbf{p}(j)_m\mathbf{h}^*_{10}(m,j,k)|_m, & m\neq j \\ \mathbf{P}(j)\xi^*_{10}(j,k)_j, & m=j \end{cases}$$

or

$$\xi^*_{10}(j,k)_m = \begin{cases} (\mathbf{I}-\mathbf{P}(j)\mathbf{I}_0(m))^{-1}\mathbf{p}(j)_m\mathbf{h}^*_{10}(m,j,k)|_m, & m\neq j \\ \mathbf{0}, & m=j. \end{cases} \tag{7.49}$$

Now let

$$\eta^*_{10}(j,k) = \begin{pmatrix} \mathbf{h}^*_{10}(1,j,k)|_1 \\ \vdots \\ \mathbf{h}^*_{10}(J,j,k)|_J \end{pmatrix} \in \mathscr{R}^{J\times 1},$$

$$\mathbf{U}^*(j)_m = \mathrm{diag}(u^*_m(1),\cdots,u^*_m(j-1),0,u^*_m(j+1),\cdots,u^*_m(J)) \in \mathscr{R}^{J\times J},$$

$$\delta(k) = (0,\cdots,0,\underbrace{\delta(k)}_{k^{th}\text{ place}},0,\cdots,0)' \in \mathscr{R}^{J\times 1}.$$

Then from (7.47) and (7.49), we have

$$\begin{aligned}\eta^*_{10}(j,k) &= \delta(k)\mathbf{1}\{k \neq j\} + \sum_{m=1}^{J} \mathbf{U}^*(j)_m \xi^*_{10}(j,k)_m \\ &= \delta(k)\mathbf{1}\{k \neq j\} + \sum_{m \neq j} \mathbf{U}^*(j)_m \mathbf{q}(j)_m \mathbf{h}^*_{10}(m,j,k)|_m,\end{aligned}$$

where $\mathbf{q}(j)_m = (\mathbf{I} - \mathbf{P}(j)\mathbf{I}_0(m))^{-1}\mathbf{p}(j)_m$. Let

$$\begin{aligned}\mathscr{U}^*(j) = \big(&\mathbf{U}^*(j)_1\mathbf{q}(j)_1, \ldots, \mathbf{U}^*(j)_{j-1}\mathbf{q}(j)_{j-1}, \mathbf{0}, \\ &\mathbf{U}^*(j)_{j+1}\mathbf{q}(j)_{j+1}, \ldots, \mathbf{U}^*(j)_J\mathbf{q}(j)_J\big) \in \mathscr{R}^{J\times J}.\end{aligned} \tag{7.50}$$

Then we have

$$\eta^*_{10}(j,k) = \delta(k)\mathbf{1}\{k \neq j\} + \mathscr{U}^*(j)\eta^*_{10}(j,k), \qquad (j,k \in \Pi). \tag{7.51}$$

Algorithm for the second reduction: Repeat the following steps for $j = 1, \ldots, J$.

1. Solve $(\mathbf{I} - \mathbf{P}(j)\mathbf{I}_0(m))\mathbf{q}(j)_m = \mathbf{p}(j)_m$ to obtain $\mathbf{q}(j)_m$ for $m \neq j$.
2. Set matrix $\mathscr{U}^*(j)$ defined in (7.50).
3. Solve the set of the equations given by (7.51) to obtain $\eta^*_{10}(j,k)$ for $k \neq j$.
4. From (7.49), $\xi^*_{10}(j,k)_m = \mathbf{q}(j)_m \mathbf{h}^*_{10}(m,j,k)|_m = \mathbf{q}(j)_m \eta^*_{10}(j,k)|_m$ for $m \neq j$[8].
5. From the definition of $\eta^*_{10}(\cdot)$ and (7.47), we have[9]

$$\mathbf{h}^*_{10}(\kappa_0,j,k)|_m = \begin{cases} \eta^*_{10}(j,k)|_{\kappa_0}, & m = \kappa_0 \\ \xi^*_{10}(\kappa_0,j,k)|_m = \mathbf{q}(j)_m|_{\kappa_0}\mathbf{h}^*_{10}(m,j,k)|_m, & m \neq \kappa_0. \end{cases} \tag{7.52}$$

Third Reduction Step:

The computational effort in the algorithm for the second reduction can be further reduced in the following manner. It can be easily shown that $(\mathbf{I} - \mathbf{P}(j)\mathbf{I}_0(m))\mathbf{q}(j)_m = \mathbf{p}(j)_m$ for $m \neq j$ can be written as follows:

$$(\mathbf{I} - \mathbf{P}(j))\mathbf{q}(j)_m = \mathbf{p}(j)_m(1 - \mathbf{q}(j)_m|_m),$$

where $\mathbf{q}(j)_m|_m$ is the mth element of the vector $\mathbf{q}(j)_m$. Hence we have

$$\mathbf{q}(j)_m = (\mathbf{I} - \mathbf{P}(j))^{-1}\mathbf{p}(j)_m(1 - \mathbf{q}(j)_m|_m).$$

Then it can be easily shown that

$$\mathbf{q}(j)_m|_m = \mathbf{q}'(j)_m|_m(1 + \mathbf{q}'(j)_m|_m)^{-1},$$

[8] Because $\mathbf{h}^*_{10}(j,j,k)|_j = \eta^*_{10}(j,k)|_j = 0$, $\xi^*_{10}(j,k)_m = \mathbf{q}(j)_m\mathbf{h}^*_{10}(m,j,k)|_m$ is also true for $m = j$.

[9] $\eta^*_{10}(j,k)|_{\kappa_0}$ and $\mathbf{q}(j)_m|_{\kappa_0}$ are the κ_0th elements of $\eta^*_{10}(j,k)$ and $\mathbf{q}(j)_m$, respectively.

where $\mathbf{q}'(j)_m|_m$ is the mth element of the vector $\mathbf{q}'(j)_m = (\mathbf{I}-\mathbf{P}(j))^{-1}\mathbf{p}(j)_m$. The first step 1 of the algorithm for the second reduction then can be arranged as

1′. Solve $(\mathbf{I}-\mathbf{P}(j))\mathbf{q}'(j)_m = \mathbf{p}(j)_m$ to obtain $\mathbf{q}'(j)_m$ for $m \neq j$. Then set

$$\mathbf{q}(j)_m|_m = \mathbf{q}'(j)_m|_m(1+\mathbf{q}'(j)_m|_m)^{-1}, \qquad \mathbf{q}(j)_m = \mathbf{q}'(j)_m(1-\mathbf{q}(j)_m|_m).$$

The computational complexity of the algorithm is evaluated later.

7.6.2 Reduction of Calculations of Steady-State Values

Because we cannot further reduce the number of the steady-state equations, we would like to solve them by a successive approximation instead of directly solving them. Because it takes much computational effort to apply the original equations (7.42) and (7.43) to the approximation, we would like to reduce it by arranging coefficients $\mathbf{h}_{00}(\cdot)$ as follows.

From (7.15) (7.16), and (7.23), we have

$$\mathbf{h}^0_{00}(\kappa_0,j,k) = \begin{pmatrix} * \\ \mathbf{0} \end{pmatrix}, \qquad (\kappa_0,j,k \in \Pi).$$

That is, $H^0_j(\mathbf{Y},e,l,k)$ for $\kappa_0,k \in \Pi$ and $l=0$ is not affected by the number of customers in the waiting room $\mathbf{n}$. From (7.19), (7.20), (7.22), and (7.25), we have

$$\mathbf{U}_0(\kappa_0) = \begin{pmatrix} \mathbf{O} & * \\ \mathbf{O} & \mathbf{U}^*_{01}(\kappa_0) \end{pmatrix}, \qquad (\kappa_0 \in \Pi),$$

where

$$\mathbf{U}^*_{01}(\kappa_0) = \mathrm{diag}(1,\ldots,1,\underbrace{\mathbf{1}\{\kappa_0 \in \mathscr{H}_g\}}_{\kappa_0^{th}\text{ place}},1,\ldots,1) \in \mathscr{R}^{J\times J}.$$

Let $\mathbf{h}^*_{00}(\kappa_0,j,k) \in \mathscr{R}^{J\times 1}$ be the lower half of $\mathbf{h}_{00}(\kappa_0,j,k)$ for $\kappa_0,j,k \in \Pi$. Then from the definition of $\mathbf{h}_{00}(\kappa_0,j,k)$ in Sect. 7.4, we have

$$\mathbf{h}^*_{00}(\kappa_0,j,k) = \mathbf{U}^*_{01}(\kappa_0) \sum_{\kappa_1 \in \Pi\setminus\{j\}} p_{\kappa_0\kappa_1}\mathbf{h}^*_{10}(\kappa_1,j,k), \qquad (\kappa_0 \neq j \text{ or } j \in \mathscr{H}_g)$$

and its mth element is given by

$$\mathbf{h}^*_{00}(\kappa_0,j,k)|_m = \begin{cases} \displaystyle\sum_{\kappa_1\in\Pi\setminus\{j\}} p_{\kappa_0\kappa_1}\mathbf{h}^*_{10}(\kappa_1,j,k)|_m, & m \neq \kappa_0 \text{ or } \kappa_0 \in \mathscr{H}_g \\ 0, & m = \kappa_0 \in \mathscr{H}_e \end{cases}$$

for $\kappa_0 \neq j$ or $j \in \mathscr{H}_g$ $(\kappa_0,j,k \in \Pi)$. ($\mathbf{h}^*_{00}(\kappa_0,j,k)|_m = 0$ for all m when $\kappa_0 = j \in \mathscr{H}_e$.)

Furthermore from (7.52), it can be shown that

$$\sum_{\kappa_1 \in \Pi \setminus \{j\}} p_{\kappa_0 \kappa_1} \mathbf{h}_{10}^*(\kappa_1, j, k)|_m = q^0(\kappa_0, j)_m \mathbf{h}_{10}^*(m, j, k)|_m \tag{7.53}$$

for $\kappa_0, j, k \in \Pi$, where

$$q^0(\kappa_0, j)_m = \begin{cases} \sum\limits_{\kappa_1 \in \Pi \setminus \{j,m\}} p_{\kappa_0 \kappa_1} \mathbf{q}(j)_m|_{\kappa_1} + p_{\kappa_0 m}, & m \neq j \\ \sum\limits_{\kappa_1 \in \Pi \setminus \{j\}} p_{\kappa_0 \kappa_1} \mathbf{q}(j)_j|_{\kappa_1}, & m = j. \end{cases} \tag{7.54}$$

Then we have the final expression for $\mathbf{h}_{00}^*$:

$$\mathbf{h}_{00}^*(\kappa_0, j, k)|_m = q^*(\kappa_0, j)_m \mathbf{h}_{10}^*(m, j, k)|_m \tag{7.55}$$

for $\kappa_0, j, k, m \in \Pi$, where

$$q^*(\kappa_0, j)_m = \begin{cases} q^0(\kappa_0, j)_m, & (m \neq \kappa_0 \text{ or } \kappa_0 \in \mathscr{H}_g) \text{ and } (\kappa_0 \neq j \text{ or } j \in \mathscr{H}_g) \\ 0, & \text{otherwise.} \end{cases} \tag{7.56}$$

Then from (7.42) and (7.43), it can be easily shown that the steady-state numbers of customers satisfy the following equations:

$$\tilde{n}_j^k = \begin{cases} \tilde{\varphi}_{hj}^k + \lambda_j \sum\limits_{\kappa_0 \in \Pi} \tilde{g}_{\kappa_0}^{\kappa_0} \mathbf{h}_{00}(\kappa_0, j, k)|_{\kappa_0} \\ \qquad + \lambda_j \sum\limits_{m=1}^{J} \tilde{n}_{qm}(j) \mathbf{h}_{10}^*(m, j, k)|_m, & k \in \Pi \\ \tilde{\varphi}_j^k, & k \in \Pi^s, \end{cases} \tag{7.57}$$

$$\tilde{g}_j^j = \tilde{\psi}_{fj} + \lambda_j \sum_{\kappa_0 \in \Pi} \tilde{g}_{\kappa_0}^{\kappa_0} \mathbf{f}(\kappa_0, j)|_{\kappa_0} + \lambda_j \sum_{\kappa_0 \in \Pi} \sum_{m=1}^{J} \tilde{n}_m^{\kappa_0} \mathbf{f}(\kappa_0, j)|_{J+m}, \tag{7.58}$$

$$\tilde{n}_{qm}(j) = \sum_{\kappa_0 \in \Pi} \tilde{n}_m^{\kappa_0} q^*(\kappa_0, j)_m \tag{7.59}$$

for $k \in \Pi \cup \Pi^s$ and $j, m \in \Pi$, where

$$\begin{aligned} &\tilde{\varphi}_j^k = \lambda_j \sum_{\kappa_0 \in \Pi \cup \Pi^s} \{\tilde{r}^{\kappa_0} \varphi(\kappa_0, j, k) + \tilde{q}^{\kappa_0} h_{01}(\kappa_0, j, k)\}, \\ &\tilde{\boldsymbol{\varphi}}^k = \left(\tilde{\varphi}_j^k : j = 1, \ldots, J\right) \in \mathscr{R}^{1 \times J}, \\ &\tilde{\boldsymbol{\varphi}}_{hj}^k = \tilde{\boldsymbol{\varphi}}_j^k + \lambda_j \sum_{\kappa_0 \in \Pi^s} (\mathbf{0}, \tilde{\varphi}^{\kappa_0}) \mathbf{h}_{00}(\kappa_0, j, k), \qquad (k \in \Pi \text{ for this case}), \\ &\tilde{\psi}_{fj} = \lambda_j \sum_{\kappa_0 \in \Pi \cup \Pi^s} \{\tilde{r}^{\kappa_0} \psi(\kappa_0, j)\} + \lambda_j \sum_{\kappa_0 \in \Pi^s} (\mathbf{0}, \tilde{\varphi}^{\kappa_0}) \mathbf{f}(\kappa_0, j). \end{aligned} \tag{7.60}$$

From the equations, we can construct a successive approximation algorithm for the steady-state values. Note that $\tilde{n}_j^k$ ($k \in \Pi^s,\ j \in \Pi$) can be directly calculated in advance from the known constants.

Algorithm for calculating the steady-state values by successive approximation

1. Set $s = 0$ and the initial values of $\tilde{n}_j^{k(0)}, \tilde{g}_j^{j(0)}, \tilde{n}_{qm}^{(0)}(j)$ for $j,k,m \in \Pi$.
2. Calculate $\tilde{n}_j^{k(s+1)}, \tilde{g}_j^{j(s+1)}, \tilde{n}_{qm}^{(s+1)}(j)$ for $j,k,m \in \Pi$ from the set of equations:

$$\tilde{n}_j^{k(s+1)} = \tilde{\varphi}_{hj}^k + \lambda_j \sum_{\kappa_0 \in \Pi} \tilde{g}_{\kappa_0}^{\kappa_0(s)} \mathbf{h}_{00}(\kappa_0, j, k)|_{\kappa_0} + \lambda_j \sum_{m=1}^{J} \tilde{n}_{qm}^{(s)}(j) \mathbf{h}_{10}^*(m, j, k)|_m,$$

$$\tilde{g}_j^{j(s+1)} = \tilde{\psi}_{fj} + \lambda_j \sum_{\kappa_0 \in \Pi} \tilde{g}_{\kappa_0}^{\kappa_0(s)} \mathbf{f}(\kappa_0, j)|_{\kappa_0} + \lambda_j \sum_{\kappa_0 \in \Pi} \sum_{m=1}^{J} \tilde{n}_m^{\kappa_0(s)} \mathbf{f}(\kappa_0, j)|_{J+m},$$

$$\tilde{n}_{qm}^{(s+1)}(j) = \sum_{\kappa_0 \in \Pi} \tilde{n}_m^{\kappa_0(s+1)} q^*(\kappa_0, j)_m.$$

3. If these values are considered to converge, then stop. Otherwise, let $s \leftarrow s+1$ and go to step 2.

Note. We can show (1) the uniqueness of the solution of (7.42) and (7.43), and (2) the convergence of the values obtained by the successive approximation method to the unique solution (under the assumption that these steady-state average values exist).

7.6.3 Evaluation of Computational Complexity

We now evaluate the computational complexity after the reductions. After applying the third reduction step, in order to derive $\mathbf{h}_{10}(\cdot)$, we are essentially required to solve the J sets of the $O(J)$ linear equations related to the equations $(\mathbf{I} - \mathbf{P}(j))\mathbf{q}'(j)_m = \mathbf{p}(j)_m$, and required to solve the J sets of the $O(J)$ linear equations related to (7.51). And a careful estimation shows that the other calculations require at most $O(J^4)$ numerical operations. Then it can be easily shown that only $O(J^4)$ numerical operations are required in order to calculate the constants $\mathbf{h}_{10}^*(\kappa_0, j, k)$ and $\mathbf{q}(j)_m$ $(\kappa_0, j, k, m \in \Pi)$. Hence if we directly solve the steady-state equations (7.42) and (7.43) by inverting the coefficient matrix after applying the third reduction step, $O(J^6)$ numerical operations are required in order to calculate the mean waiting times for all stations.

Then for the successive approximation of the steady-state values $(\tilde{\mathbf{g}}^k, \tilde{\mathbf{n}}^k)$, it is clear that $O(J^3)$ numerical operations are required in order to calculate the values at each iterative step. And it can be shown that calculations of the other coefficients $(\{\tilde{\varphi}_{hj}^k : j,k \in \Pi\}, \{\tilde{\varphi}_j^k : k \in \Pi \cup \Pi^s,\ j \in \Pi\}, \{\mathbf{h}_{00}(\kappa_0, j, k)|_{\kappa_0} : \kappa_0, j, k \in \Pi\}, \{\tilde{\psi}_{fj} : j \in \Pi\}, \{\mathbf{f}(\kappa_0, j) : \kappa_0, j \in \Pi\}, \{q^*(\kappa_0, j)_m : \kappa_0, j, m \in \Pi\})$ which appear in (7.57)–(7.60) require $O(J^4)$ numerical operations.

Hence if we obtain the mean waiting times for all stations after applying the third reduction step and the successive approximation for the steady-state values, $O(J^4) + O(J^3 N)$ numerical operations are required where $N = N_{J,\rho,\varepsilon}$ is the number

of iterations of the approximation that depends on the number of stations J, the resource utilization ρ, and the required accuracy ε[10].

7.6.4 Comparison of Computational Times by Examples

Now we compare our functional computation method with the buffer occupancy method by actually measuring their running times to compute the average waiting times in the systems with $J = 40$ stations and $J = 80$ stations. Half of the stations take the gated rule and the other stations take the exhaustive rule. In order to make graphs for the running times in each system by changing the resource utilization ρ, the arrival rates are varied. The service times, the switchover times, and the switching probabilities are fixed. The algorithms that adopt the following methods are compared.

Ours 1: Our functional computation method that calculates the steady-state values by directly solving the equations (i.e., inverting their coefficient matrix)

Ours 2: Our functional computation method that calculates the steady-state values by the successive approximation

B.O.: The buffer occupancy method that calculates second moments of the buffer occupancy variables by a successive approximation.

In Figs. 7.1 and 7.2, "Ours 2-1" and "Ours 2-2" denote our second method "Ours 2" with $\varepsilon = 10^{-4}$ and $\varepsilon = 10^{-8}$, respectively,[11] and "B.O.1" and "B.O.2" denote the buffer occupancy method "B.O." with $\varepsilon = 10^{-4}$ and $\varepsilon = 10^{-8}$, respectively. Although the running times of "Ours 1" do not depend on the resource utilization, they are somewhat greater than those of "Ours 2." "Ours 2" takes almost constant running times until the resource utilization reaches about 0.9. It results from the fact that when ρ is less than the value, the number of iterations N is relatively small and the computational complexity of "Ours 2" is approximately $O(J^4)$. When ρ approaches 1, N grows rapidly and its running times also grow rapidly. The numbers of iterations for the approximation methods are given in Tables 7.1 and 7.2. We see from the tables that the numbers of iterations of "B.O." are fairly (10 or more times) greater than those of "Ours 2." This may be caused by the difference between the numbers of variables in the steady-state equations; that is, "Ours 2" has only $O(J^2)$

[10] As noted in Sect. 7.1, the computational complexity of the buffer occupancy method that uses an approximation is $O(J^4 N')$ where N' is the number of its iterations.

[11] When $\left|\sum_{j=1}^{J} \rho_j \bar{w}_j^{(s)} - \sum_{j=1}^{J} \rho_j \bar{w}_j^{(s-1)}\right| < \varepsilon$, the successive approximation methods stop, where $\{\bar{w}_j^{(s)}\}$ is a set of the mean waiting times obtained at their sth iterative step and ε is a required accuracy. The used CPU is the AMD Athlon 64 X 2 4400+ with 4 GB memories, and the programming language is Intel Visual FORTRAN with the IMSL Library.

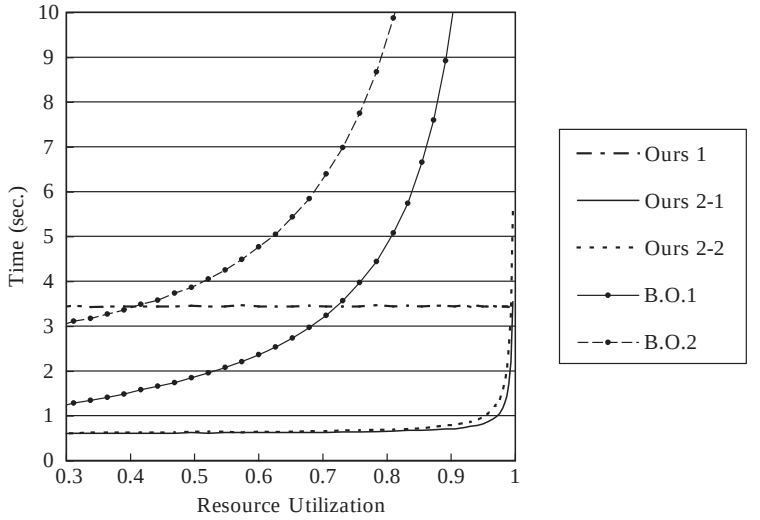


Fig. 7.1 Running times for computing the mean waiting times in the system with $J = 40$.

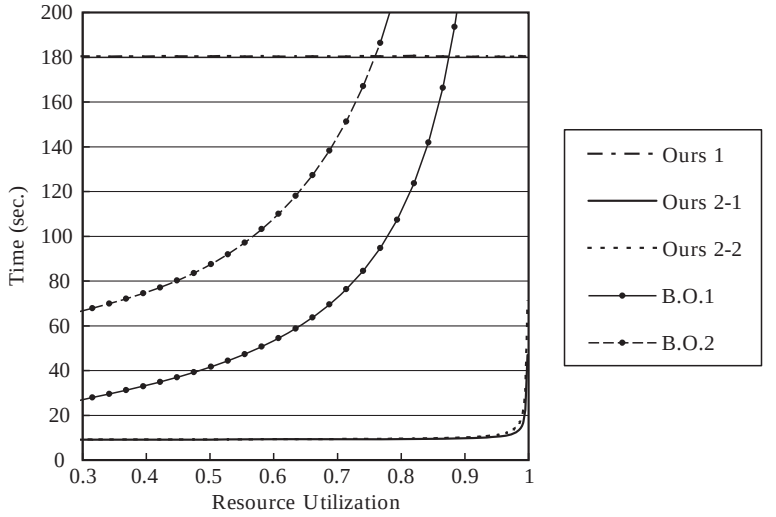


Fig. 7.2 Running times for computing the mean waiting times in the system with $J = 80$.

variables in contrast to "B.O." which has $O(J^3)$ variables. Furthermore for the buffer occupancy method, because $O(J^4)$ operations per iteration are required, its running times are greater than those of "Ours 2." These differences become large as the system is congested (i.e., when ρ is large).

Table 7.1 Numbers of iterations (N and N') for the system with $J = 40$.

	Ours 2 (N)				B.O. (N')			
	Required Accuracies (ε)				Required Accuracies (ε)			
ρ	10^{-2}	10^{-4}	10^{-6}	10^{-8}	10^{-2}	10^{-4}	10^{-6}	10^{-8}
0.3114	5	8	11	14	86	230	399	587
0.5213	9	13	18	23	166	363	566	772
0.7574	19	29	39	49	383	755	1126	1497
0.9057	50	76	102	128	1051	1992	2934	3875
0.9568	110	167	224	281	2339	4389	6440	8490
0.9899	471	714	957	1200	10102	18844	27585	36326

Table 7.2 Numbers of iterations (N and N') for the system with $J = 80$.

	Ours 2 (N)				B.O. (N')			
	Required Accuracies (ε)				Required Accuracies (ε)			
ρ	10^{-2}	10^{-4}	10^{-6}	10^{-8}	10^{-2}	10^{-4}	10^{-6}	10^{-8}
0.3154	6	9	11	14	177	463	786	1152
0.5014	9	13	18	22	318	701	1092	1490
0.7405	19	28	37	46	729	1440	2152	2863
0.8986	49	73	96	119	2006	3812	5617	7423
0.9531	107	157	208	259	4428	8322	12216	16110
0.9919	617	911	1205	1499	25952	48423	70893	93364

7.7 Conclusions

In this chapter we have considered the Markovian polling systems, and have obtained the mean waiting times. It can be shown that the explicit expression for the expected waiting time of a customer conditioned on the system state at its arrival epoch has the linear functional form, which is the representative characteristic of our method. This form results from the linear functional forms of the basic quantities given in Proposition 7.1. And the steady-state average values can be derived from it by simple limiting procedures. It has been shown that the conditional expected waiting times in many types of M/G/1 multiclass queueing systems have the similar linear functional forms. They appear not only in the polling systems [2] but also in the priority systems [24]. Furthermore the conditional expected sojourn times in the systems with customers' feedback also have the linear functional forms [1], [25].

Our functional computation for the mean waiting times in the Markovian polling systems originally requires us to solve $J + 1$ sets of $O(J^2)$ linear equations for the mean waiting times of J stations as opposed to the buffer occupancy method which requires us to solve $O(J^3)$ linear equations. Although our original method requires $O(J^7)$ numerical operations, we can construct the procedure with the successive

approximation for the steady-state values which only requires $O(J^4)+O(J^3N)$ numerical operations where N is the number of its iterations. When we compared our method with the buffer occupancy method by actually computing the mean waiting times, we found that the computation times by our method are less than those by the buffer occupancy method; especially their differences are large when the system is congested.

Besides the above things, there are many advantage of our method [25]. Multiclass queueing models are useful for analyzing the computer communication systems with many datatypes and sources, and more complicated queueing models are necessary in order to derive the performance characteristics in the real systems. Because we can investigate complicated multiclass structures and composite scheduling algorithms by our method, it may stimulate advanced analysis of these systems.

Appendix: Proof of Proposition 7.2

Proof. We prove that the polling equation (7.8) is satisfied by directly substituting the expression for $\hat{H}_j(\mathbf{Y},e,l,k)$ defined by (7.28) into it. Let $\mathbf{Y}=(\kappa_0,r,\mathbf{g},\mathbf{n},L)\in\mathscr{E}$ be the state of the system at time τ_l^e $(l=0,1,\ldots,\ e=1,2,\ldots)$.

Case 1 $(k\in\Pi)$: In the following expressions, the abbreviated condition $(\mathbf{Y},j)_l^e$ means the condition $\mathbf{Y}(\tau_l^e)=\mathbf{Y}$ and $X_S^e(\tau_l^e)=j$ for $l\geq 0,\ e=1,2,\ldots$.

For $(\kappa_0=j,l=0,j\in\mathscr{H}_e)$ or $(\kappa_0=j,l>0,j\in\mathscr{H}_e\cup\mathscr{H}_g)$, it can be easily shown that

$$\hat{H}_j(\mathbf{Y},e,l,k)=0.$$

For $(l=0,\kappa_0\in\Pi,\kappa_0\neq j)$ or $(l=0,\kappa_0=j\in\mathscr{H}_g)$,

$$\begin{aligned}
&H_j^0(\mathbf{Y},e,0,k)+\mathrm{E}[\hat{H}_j(\mathbf{Y}(\tau_1^e),e,1,k)|\mathbf{Y}(\tau_0^e)=\mathbf{Y},X_S^e(\tau_0^e)=j]\\
&=r\varphi^0(\kappa_0,j,k)+(\mathbf{g},\mathbf{n})\mathbf{h}_{00}^0(\kappa_0,j,k)\\
&\quad+\mathrm{E}[(\mathbf{g}(\tau_1^e),\mathbf{n}(\tau_1^e))\mathbf{h}_{10}(\kappa(\tau_1^e),j,k)+h_{11}(\kappa(\tau_1^e),j,k)|(\mathbf{Y},j)_0^e]\\
&=r\varphi^0(\kappa_0,j,k)+(\mathbf{g},\mathbf{n})\mathbf{h}_{00}^0(\kappa_0,j,k)+\sum_{\kappa_1\neq j}p_{\kappa_0\kappa_1}h_{11}(\kappa_1,j,k)\\
&\quad+\sum_{\kappa_1\neq j}p_{\kappa_0\kappa_1}\mathrm{E}[(\mathbf{g}(\tau_1^e),\mathbf{n}(\tau_1^e))|\kappa(\tau_1^e)=\kappa_1,(\mathbf{Y},j)_0^e]\mathbf{h}_{10}(\kappa_1,j,k)\\
&=r\varphi^0(\kappa_0,j,k)+(\mathbf{g},\mathbf{n})\mathbf{h}_{00}^0(\kappa_0,j,k)+\sum_{\kappa_1\neq j}p_{\kappa_0\kappa_1}h_{11}(\kappa_1,j,k)\\
&\quad+\sum_{\kappa_1\neq j}p_{\kappa_0\kappa_1}\{r\upsilon(\kappa_0)+(\mathbf{g},\mathbf{n})\mathbf{U}_0(\kappa_0)+\mathbf{u}_0(j,\kappa_0,\kappa_1)\}\mathbf{h}_{10}(\kappa_1,j,k)\\
&=\hat{H}_j(\mathbf{Y},e,0,k).
\end{aligned}$$

The first equation comes from (7.23) and (7.28). The second equation comes from the definition of the switching probability $p_{\kappa_0\kappa_1}$. The third equation comes from (7.25). The last equation comes from the definitions of the constants (in Sect. 7.4) and (7.28).

For $l = 0$ and $\kappa_0 = (k_0, k_1) \in \Pi^s$,

$$\begin{aligned}
&H_j^0(\mathbf{Y}, e, 0, k) + \mathrm{E}[\hat{H}_j(\mathbf{Y}(\tau_1^e), e, 1, k) | \mathbf{Y}(\tau_0^e) = \mathbf{Y}, X_S^e(\tau_0^e) = j] \\
&= \mathrm{E}[(\mathbf{g}(\tau_1^e), \mathbf{n}(\tau_1^e))\mathbf{h}_{10}(\kappa(\tau_1^e), j, k) + h_{11}(\kappa(\tau_1^e), j, k) | (\mathbf{Y}, j)_0^e] \\
&= \mathrm{E}[(\mathbf{g}(\tau_1^e), \mathbf{n}(\tau_1^e)) | \kappa(\tau_1^e) = k_1, (\mathbf{Y}, j)_0^e]\mathbf{h}_{10}(k_1, j, k) + h_{11}(k_1, j, k) \\
&= \left\{ r\upsilon + (\mathbf{g}, \mathbf{n})\mathbf{U}_0 + (\mathbf{0}, \mathbf{e}_j) \right\} \mathbf{h}_{10}(k_1, j, k) + h_{11}(k_1, j, k) \\
&= \hat{H}_j(\mathbf{Y}, e, 0, k).
\end{aligned}$$

For $l > 0$ and $\kappa_0 \neq j$ ($\kappa_0 \in \Pi$),

$$\begin{aligned}
&H_j^0(\mathbf{Y}, e, l, k) + \mathrm{E}[\hat{H}_j(\mathbf{Y}(\tau_{l+1}^e), e, l+1, k) | \mathbf{Y}(\tau_l^e) = \mathbf{Y}, X_S^e(\tau_l^e) = j] \\
&= (\mathbf{g}, \mathbf{n})\mathbf{h}_{10}^0(\kappa_0, j, k) \\
&\quad + \mathrm{E}[(\mathbf{g}(\tau_{l+1}^e), \mathbf{n}(\tau_{l+1}^e))\mathbf{h}_{10}(\kappa(\tau_{l+1}^e), j, k) + h_{11}(\kappa(\tau_{l+1}^e), j, k) | (\mathbf{Y}, j)_l^e] \\
&= (\mathbf{g}, \mathbf{n})\mathbf{h}_{10}^0(\kappa_0, j, k) + \sum_{\kappa_1 \neq j} p_{\kappa_0\kappa_1} h_{11}(\kappa_1, j, k) \\
&\quad + \sum_{\kappa_1 \neq j} p_{\kappa_0\kappa_1} \mathrm{E}[(\mathbf{g}(\tau_{l+1}^e), \mathbf{n}(\tau_{l+1}^e)) | \kappa(\tau_{l+1}^e) = \kappa_1, (\mathbf{Y}, j)_l^e]\mathbf{h}_{10}(\kappa_1, j, k) \\
&= (\mathbf{g}, \mathbf{n})\mathbf{h}_{10}^0(\kappa_0, j, k) + \sum_{\kappa_1 \neq j} p_{\kappa_0\kappa_1} h_{11}(\kappa_1, j, k) \\
&\quad + \sum_{\kappa_1 \neq j} p_{\kappa_0\kappa_1} \left\{ (\mathbf{g}, \mathbf{n})\mathbf{U}_1(\kappa_0) + \mathbf{u}_1(\kappa_0, \kappa_1) \right\} \mathbf{h}_{10}(\kappa_1, j, k) \\
&= \hat{H}_j(\mathbf{Y}, e, l, k).
\end{aligned}$$

Case 2 ($k \in \Pi^s$): The proof is similar to case 1 and is omitted.
Hence the proof is completed. □

References

1. T. Hirayama, Mean sojourn times in multiclass feedback queues with gated disciplines, *Naval Research Logistics*, vol. 50, no. 7, pp. 719–741, 2003.
2. T. Hirayama, S. J. Hong, and M. M. Krunz, A new approach to analysis of polling systems, *Queueing Systems*, vol. 48, nos. 1–2, pp. 135–158, 2004.
3. M. M. Srinivasan, Nondeterministic polling systems, *Management Science*, vol. 37, no. 6, pp. 667–681, 1991.
4. H. Levy and M. Sidi, Polling systems: Applications, modeling, and optimization, *IEEE Transactions on Communications*, vol. 38, no. 10, pp. 1750–1760, 1990.
5. H. Takagi, Analysis and application of polling models, in: G. Haring, C. Lindemann, and M. Reiser (Eds.), *Performance Evaluation: Origins and Directions*, *Lecture Notes in Computer Science*, vol. 1769, pp. 423–442. Berlin: Springer, 2000.
6. R. B. Cooper, Queues served in cyclic order: Waiting times, *Bell System Technical Journal*, vol. 49, no. 3, pp. 399–413, 1970.

7. R. B. Cooper and G. Murray, Queues served in cyclic order, *Bell System Technical Journal*, vol. 48, no. 3, pp. 675–689, 1969.
8. M. Eisenberg, Queues with periodic service and changeover time, *Operations Research*, vol. 20, no. 2, pp. 440–451, 1972.
9. H. Takagi, Analysis of polling systems with a mixture of exhaustive and gated service disciplines, *Journal of the Operations Research Society of Japan*, vol. 32, no. 4, pp. 450–461, 1989.
10. H. Levy and M. Sidi, Polling systems with simultaneous arrivals, *IEEE Transactions on Communications*, vol. 39, no. 6, pp. 823–827, 1991.
11. M. Sidi, H. Levy and S. W. Fuhrmann, A queueing network with a single cyclically roving server, *Queueing Systems*, vol. 11, nos. 1–2, pp. 121–144, 1992.
12. M. J. Ferguson and Y. J. Aminetzah, Exact results for nonsymmetric token ring systems, *IEEE Transactions on Communications*, vol. 33, no. 3, pp. 223–231, 1985.
13. D. Sarkar and W. I. Zangwill, Expected waiting time for nonsymmetric cyclic queueing systems — Exact results and applications, *Management Science*, vol. 35, no. 12, pp. 1463–1474, 1989.
14. E. M. M. Winands, I. J. B. F. Adan, and G. J. Van Houtum, Mean value analysis for polling systems, *Queueing Systems*, vol. 54, no. 1, pp. 35–44, 2006.
15. H. Takagi, *Analysis of Polling Systems*. Cambridge: MIT Press, 1986.
16. A. G. Konheim, H. Levy, and M. M. Srinivasan, Descendant set: An efficient approach for the analysis of polling systems, *IEEE Transactions on Communications*, vol. 42, nos. 2/3/4, pp. 1245–1253, 1994.
17. M. P. Singh and M. M. Srinivasan, Exact analysis of the state-dependent polling model, *Queueing Systems*, vol. 41, no. 4, pp. 371–399, 2002.
18. O. J. Boxma, Workloads and waiting times in single-server systems with multiple customer classes, *Queueing Systems*, vol. 5, nos. 1–3, pp. 185–214, 1989.
19. R. B. Cooper, S-C. Niu, and M. M. Srinivasan, A decomposition theorem for polling models: The switchover times are effectively additive, *Operations Research*, vol. 44, no. 4, pp. 629–633, 1996.
20. L. Kleinrock and H. Levy, The analysis of random polling systems, *Operations Research*, vol. 36, no. 5, pp. 716–732, 1988.
21. H. Chung, C. K. Un, and W. Y. Jung, Performance analysis of Markovian polling systems with single buffers, *Performance Evaluation*, vol. 19, no. 4, pp. 303–315, 1994.
22. J. E. Baker and I. Rubin, Polling with a general-service order table, *IEEE Transactions on Communications*, vol. 35, no. 3, pp. 283–288, 1987.
23. O. J. Boxma, H. Levy, and J. A. Weststrate, Efficient visit frequencies for polling tables: Minimization of waiting cost, *Queueing Systems*, vol. 9, nos. 1–2, pp. 133–162, 1991.
24. T. Hirayama, Analysis of multiclass M/G/1 queues with a mixture of 1-limited disciplines and gated disciplines, *Journal of the Operations Research Society of Japan*, vol. 42, no. 3, pp. 237–255, 1999.
25. T. Hirayama, Multiclass polling systems with Markovian feedback: Mean sojourn times in gated and exhaustive systems with local priority and FCFS service orders, *Journal of the Operations Research Society of Japan*, vol. 48, no. 3, pp. 226–255, 2005.
26. H. Levy, Delay computation and dynamic behavior of non-symmetric polling systems, *Performance Evaluation*, vol. 10, no. 1, pp. 35–51, 1989.
27. W. Whitt, A review of $L = \lambda W$ and extensions, *Queueing Systems*, vol. 9, no. 3, pp. 235–268, 1991.

Chapter 8
Performance Analysis of a Two-Station MTO/MTS Production System

Kuo-Hwa Chang and Yang-Shu Lu

Abstract We consider a two-station hybrid MTO/MTS production system with random ordinary and specific demands, in which the first station is a MTS system providing the finished standard products for ordinary demands. These finished products also serve as the semi-finished products to specific demands. The second station performs some additional work on the standard products for specific demands. In our system, the MTS system is controlled under the base-stock policy. To evaluate the system, we consider the fill rate of the ordinary demands and the response time of the specific demands. Our objective is to study the relation between base-stock level and the fill rate of the ordinary demands and the response time of the specific demands. We analyze our system by modeling it as an inventory-queue system. Based on these analyses, we can determine the optimal base-stock level numerically under the constraints on the fill rate of the ordinary demands and the response time of the specific demands.

8.1 Introduction

Traditionally, a production system can be distinguished into make-to-order (MTO) or make-to-stock (MTS) systems. MTO products are usually made to customer specifications as nonstandard and custom products, however, MTS products are standard and delivered from inventory (stock). That is, a MTS production stocks the finished products in advance whereas a MTO system starts producing only when it receives orders from the demand. Assembly manufacturing plays a very important role in the

K.-H. Chang
Department of Industrial Engineering, Chung Yuan Christian University, Chung-Li 320, Taiwan
e-mail: kuohwa@cycu.edu.tw

Y.-S. Lu
Department of Industrial Engineering, Chung Yuan Christian University, chung-Li 320, Taiwan
e-mail: g9402404@cycu.edu.tw

W. Yue et al. (eds.), *Advances in Queueing Theory and Network Applications,*

global supply chain of consumer products, such as laptop computers. Assemblers, in addition to fulfilling the ordinary demands for the standard products by adopting MTS production, are often asked to take care of the specific demands for the custom products and to adopt MTO production. In usual cases, ordinary demands are the planned orders and should be satisfied immediately, however, there is a time window for the specific demand.

To the assembler, it is not profitable to maintain a solo MTO production line exclusively for the specific demand. In some case, custom products share almost all the parts of the standard products, therefore, the assembler usually considers embedding the MTO lines into the mainstream MTS lines, which become a hybrid production system. The corresponding design and the control issues for the hybrid lines are important to management.

In this chapter, we assume the custom products can be made by alternating the existing standard ones with little work. We consider a two-station hybrid MTO/MTS production system (see Fig. 8.1) with random ordinary and specific demands, in which the first station (station 1) is a MTS system providing the finished standard products for the ordinary demands. There is a base-stock level for the finished standard products.

These standard products also serve as semi-finished products to the specific demands collected at station 2 where the additional work on the finished standard product is performed to fulfill the corresponding specific demands.

When an ordinary demand arrives at station 1, if there are finished standard products, it will take one of them and leave and, at the same time, this satisfied ordinary order will send a production order to station 1 for a new standard product; if there are no finished products, this ordinary demand will be lost. When a specific order arrives, it will send a request (order) to station 1 for acquiring a finished standard

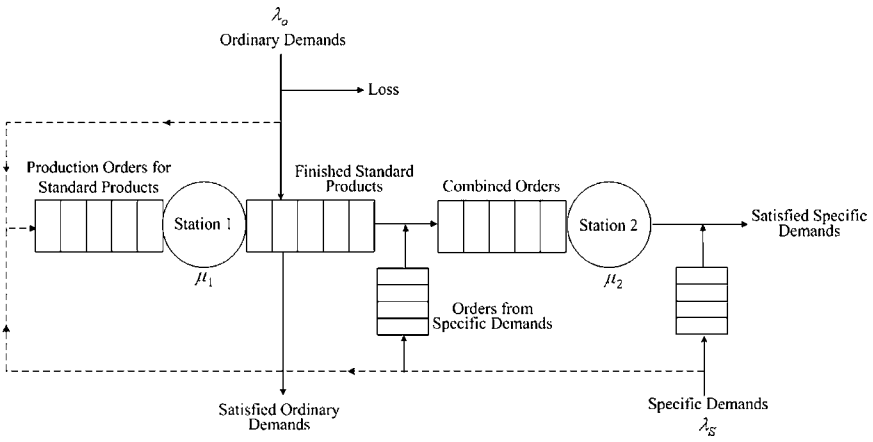


Fig. 8.1 Two-station MTO/MTS hybrid system.

product in stock and, at the same time, station 1 will also send a production order to itself for producing another new standard product; if it obtains a finished standard product, then this order along with the corresponding finished product will become a combined order and will enter station 2 on a First-Come First-Served (FCFS) basis for the additional work.

Among the research on hybrid systems, Soman, van Donk, and Gaalman [1] review the studies on the hybrid MTO/MTS production and mention that such systems can often be seen in the food industry. Krishnamurthy, Suri, and Vernon [2] use simulation to analyze a MTO/MTS hybrid system in which a base-stock controlled MTS production system supplies finished product to multiple MTO production systems. It also compares the performance of MRP and Kanban for a multistage, multiproduct manufacturing system. Adan and Ven der Wal [3] present two single-station systems.

The first system deals with MTS and MTO demands with base-stock control. Production is pre-empted by the MTO demand. The second system deals with the specific demands with base-stock control for the semi-finished products. Production is in two phases. The first phase is to produce semi-finished products and the second phase is to perform the further work on the semi-finished products in stock according to specific demands. Nguyen [4] considers a single-station hybrid production system for multiple MTS orders and multiple MTO orders. MTS orders are satisfied from the inventory controlled by base-stock policy and they are lost if there is no inventory. He models it as a mixed queueing network and approximates the performances under heavy traffic conditions by using the corresponding limiting theorem. Federgruen and Katalan [5] consider a single-station system producing some MTS products and one MTO product.

For the MTS products the base-stock policies with general periodic sequence are considered. By using an M/G/1 model with vacations, the impacts of various priority rules for the MTO products are studied. Carr and Duenyas [6] consider a single-station hybrid production system for the MTS order and MTO order. The MTS orders are satisfied from the finished-product inventory. There is no backorder for the MTS order and unsatisfied MTS orders are lost. They apply admission control on the MTO orders and sequencing on jobs at the workstation. They use the Markov decision process to find an optimal policy to maximize the average profit rate and obtain the corresponding switching curves. Arreola-Risa and DeCroix [7] consider a single-station system producing multiple products with base-stock inventory policies. They study the optimality conditions to decide which products are make-to-stock and which are make-to-order (with base-stock level zero) in order to have the minimum average cost per unit and minimum average cost rate per unit, respectively. Rajagopalan [8] also considers a single-station system for the MTS order and MTO order.

The inventory control policy for the MTS products is a (q,r) policy. Production orders for both MTS and MTO items are served on a FCFS basis. The objective is to partition the MTO/MTS items in order to minimize the inventory costs of MTS products while satisfying the constraint that the percentage of orders of MTO products fulfilled within lead time must be over a prespecified service level. The system

is modeled as an M/G/1 system. The corresponding optimization problem is modeled as a nonlinear integer program and is solved by a heuristic procedure.

To evaluate our system, we consider the fill rate (on the other side, the loss rate) of the ordinary demands and the in-time rates for the specific demands. In-time rates are defined as the probability that the waiting times of specific demands in the system, called the response times, are less than the predetermined lead time. Our system is analyzed by modeling it as an inventory-queue system. For studying the fill rate of the ordinary demands, we consider station 1 separately. We model it as an inventory queue with two classes of demands: ordinary demands and specific demands. By assuming the Markovian property, the limiting probabilities are obtained and the corresponding fill rate under base-stock control policy can also be obtained. For studying the response time for the specific demand, we study the recursive equations for approximating the response times. From these recursive equations, we can express the response times from their preceding demands and, furthermore, we can estimate the approximated distribution of the response time of specific demands. Combining the above analyses, we can further determine the optimal base-stock level under the constraints on the fill rate of the ordinary demands and the in-time rates for the specific demands according to some cost structure. We call the requirements on the fill rate for ordinary demands and the in-time rate for specific demands the corresponding required qualities of services.

The remainder of this chapter is organized as follows. In Sect. 8.2, we present the inventory-queue model of our hybrid system. Our model is analyzed and the closed-form expressions for the fill rate and the distribution of the response times are obtained. In Sect. 8.3, we verify our approximations obtained in Sect. 8.2 and present some numerical examples. We conclude our study in Sect. 8.4.

8.2 Model Description

We consider a two-station hybrid production system in which the ordinary demands arrive at station 1 according to a Poisson process with rate λ_o and specific demands arrive at station 2 according to a Poisson process with λ_s. We assume the exponential service times at each station with respective rates μ_1 and μ_2. Station 1 (MTS system) is controlled under the base-stock policy with base-stock level S.

Let B_1 be the number of production orders for standard products in the queue or under processing at station 1; B_2 be the number of orders from specific demands at the end of station 1; B_3 be the number of specific demands waiting in the demand queue at the end of station 2; N_1 be the number of finished standard products in stock at station 1; and N_2 be the number of combined orders in the queue or under processing at station 2. Note that only one of N_1 and B_2 can be positive. We have the following relations.

Proposition 8.1.

$$B_3 = B_2 + N_2, \tag{8.1}$$

$$B_1 + N_1 - B_2 = S. \tag{8.2}$$

Proof. Relation (8.1) is well-known for a base-stock system. In fact, (8.2) is true for a base-stock system with only one kind of demand. Before we prove (8.2) by induction, we discuss the changes on B_1, B_2, and N_2 after any state transition. We first consider the case when $N_1 > 0$ ($B_2 = 0$). In this case, if a demand arrives, whether it is a specific demand or an ordinary demand, N_1 will be decreased by 1 and B_1 will be increased by 1, however, B_2 is still zero and (8.2) still holds; if a standard product is produced, then B_1 will be decreased by 1 but N_1 will be increased by 1.

Now we consider the case when $N_1 = 0$ ($B_2 \geq 0$). In this case, there will be no arriving ordinary demands that can be satisfied. If a specific demand arrives, both B_1 and B_2 will be increased by 1; if a standard product is produced and $B_2 = 0$ then B_1 will be decreased by 1 but N_1 will be increased by 1; if a standard product is produced and $B_2 > 0$ then B_1 will be decreased by 1 but B_2 will decrease by 1, and N_1 is still zero. All the changes mentioned above still make (8.2) hold.

We now prove (8.2) by induction on state transitions. Initially, $N_1 = S$, $B_1 = 0$, and $B_2 = 0$. After the first transition, (8.2) still holds from the assertion for the case $N_1 > 0$. Suppose that, after the kth transition (8.2) holds, then (8.2) will still hold after the $(k+1)$st transition based on the above assertions.

For the fill rate of the ordinary demands, we consider the subsystem corresponding to station 1. Let the state be (m,n) where m denotes the number of finished standard products at station 1 and n denotes the number of orders from specific demands in stock at the end of station 1. That is, $m = N_1$ and $n = B_2$. The possible states are actually $(m,0)$ where $0 \leq m \leq S$ and $(0,n)$ for all $n \geq 0$. The corresponding transition rate diagram is shown in Fig. 8.2. Note that if l is the number of production orders for the standard products in the queue or under processing, then, from (8.2), we have $l = n - m + S$. Our objective here is to find the fill rate, denoted by P_f, for the ordinary demands and the corresponding effective arrival rate, denoted by λ_e, where $\lambda_e = P_f \lambda_o$.

Define $P(m,n)$ to be the limiting probability of state (m,n); then the balance equations are as follows:

$$\begin{aligned}
(\lambda_s + \lambda_o) P(S,0) &= \mu_1 P(S-1,0), \\
(\lambda_s + \lambda_o + \mu_1) P(m,0) &= \mu_1 P(m-1,0) + (\lambda_s + \lambda_o) P(m+1,0), \quad 1 \leq m \leq S-1, \\
(\lambda_s + \mu_1) P(0,0) &= (\lambda_s + \lambda_o) P(1,0) + \mu_1 P(0,1), \\
(\lambda_s + \mu_1) P(0,n) &= \lambda_s P(0,n-1) + \mu_1 P(0,n+1), \quad 1 \leq n \leq \infty.
\end{aligned}$$

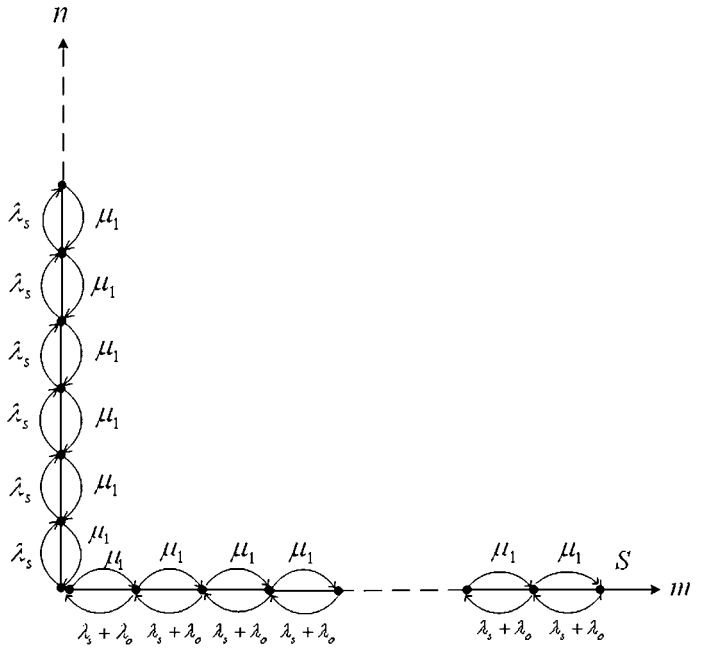


Fig. 8.2 State transition diagram for a two-station MTO/MTS hybrid system.

We have the further expressions for any $P(m,n)$.

$$P(m,0) = \left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)^{S-m} P(S,0), \qquad 0 \le m \le S-1,$$

$$P(0,n) = \left(\frac{\lambda_s}{\mu_1}\right)^{n} \left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)^{S} P(S,0), \qquad 1 \le n \le \infty.$$

By the law of total probabilities,

$$\sum_{m=0}^{S} P(m,0) + \sum_{n=1}^{\infty} P(0,n) = 1. \tag{8.3}$$

If $\lambda_s/\mu_1 < 1$, then limiting probabilities exist and

$$P(S,0) = \left\{ \left(\frac{1-\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)^{S+1}}{1-\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)} \right) + \left(\frac{\frac{\lambda_s}{\mu_1}\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)^{S}}{1-\frac{\lambda_s}{\mu_1}} \right) \right\}^{-1}.$$

Let L_{N_1} denote the expected number of finished standard products in stock and L_{B_2} be the expected number of orders from specific orders at the end of station 1; then P_f, L_{N_1}, and L_{B_2} can be obtained as follows:

$$\begin{aligned} P_f &= \sum_{m=1}^{S} P(m,0) \\ &= \left(\frac{1-\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)^S}{1-\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)} \right) \left\{ \left(\frac{1-\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)^{S+1}}{1-\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)} \right) + \left(\frac{\frac{\lambda_s}{\mu_1}\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)^S}{1-\frac{\lambda_s}{\mu_1}} \right) \right\}^{-1}, \end{aligned} \tag{8.4}$$

$$\begin{aligned} L_{N_1} &= \sum_{m=0}^{S} mP(m,0) \\ &= P(S,0)\left(1\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)^{S-1} + 2\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)^{S-2} + \cdots + (S-1)\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right) + S \right) \\ &= \frac{\left(\frac{1-\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)^S}{1-\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)} - \frac{S}{\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)} \right)}{\left(1-\frac{1}{\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)} \right) \left\{ \left(\frac{1-\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)^{S+1}}{1-\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)} \right) + \left(\frac{\frac{\lambda_s}{\mu_1}\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)^S}{1-\frac{\lambda_s}{\mu_1}} \right) \right\}}, \end{aligned} \tag{8.5}$$

$$\begin{aligned} L_{B_2} &= \sum_{n=0}^{\infty} nP(0,n) \\ &= \frac{\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)^S \left\{ \frac{\left(\frac{\lambda_s}{\mu_1}\right)}{\left(1-\left(\frac{\lambda_s}{\mu_1}\right)\right)^2} \right\}}{\left\{ \left(\frac{1-\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)^{S+1}}{1-\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)} \right) + \left(\frac{\frac{\lambda_s}{\mu_1}\left(\frac{\lambda_s+\lambda_o}{\mu_1}\right)^S}{1-\frac{\lambda_s}{\mu_1}} \right) \right\}}. \end{aligned} \tag{8.6}$$

Because we want to find the base-stock levels where the predetermined qualities of services can be satisfied, we should look at the limiting behaviors of the fill rate, L_{N_1} and L_{B_2} as S goes to infinity. We need to study them in two cases: $\lambda_s + \lambda_o < \mu_1$ and $\lambda_s + \lambda_o \geq \mu_1$. After some algebra, the limits are found and presented in the following theorem.

Proposition 8.2. *As* $S \to \infty$,

(a) If $\lambda_s + \lambda_o < \mu_1$, *then*

$$P_f \to 1, \tag{8.7}$$

$$L_{N_1} \to \infty, \tag{8.8}$$

$$L_{B_2} \to 0. \tag{8.9}$$

(b) If $\lambda_s + \lambda_o \geq \mu_1$, *then*

$$P_f \to \frac{\mu_1 - \lambda_s}{\lambda_o}, \tag{8.10}$$

$$L_{N_1} \to \frac{\dfrac{\lambda_s + \lambda_o}{\mu_1}}{\left(\dfrac{\lambda_s + \lambda_o}{\mu_1} - 1\right)\left(\left(\dfrac{\lambda_s + \lambda_o}{\mu_1}\right) + \left(\dfrac{\lambda_s}{\mu_1 - \lambda_s}\right)\left(\dfrac{\lambda_s + \lambda_o}{\mu_1} - 1\right)\right)}, \tag{8.11}$$

$$L_{B_2} \to \frac{\dfrac{\lambda_s}{\mu_1}\left(\dfrac{\lambda_s + \lambda_o}{\mu_1} - 1\right)}{\left(\left(\dfrac{\lambda_s + \lambda_o}{\mu_1}\right) - \left(\dfrac{\lambda_s}{\mu_1}\right)\right)\left(1 - \left(\dfrac{\lambda_s}{\mu_1}\right)\right)}. \tag{8.12}$$

Proof. The proofs for the results in part (a) are straightforward and they are omitted. For part (b), we only prove the convergence of (8.10). The proofs on the convergences of the other two can be conducted in the similar way. The term on the right side of (8.4) can be rewritten in terms of $\rho = (\lambda_o + \lambda_s)/\mu_1$ and $\rho_s = \lambda_s/\mu_1$ as follows:

$$\frac{(1-\rho^S)/(1-\rho)}{(1-\rho^{S+1})/(1-\rho) + \rho_s \rho^S/(1-\rho_s)}. \tag{8.13}$$

After applying l'Hôpital's rule, it can be shown that (8.13) converges to

$$\frac{-1/(1-\rho)}{-\rho/(1-\rho) + \rho_s/(1-\rho_s)} \tag{8.14}$$

which can be simplified to $(\mu_1 - \lambda_s)/\lambda_o$.

According to Proposition 8.2, when the capacity of the workstation is large enough to handle all the traffic, the fill rate will converge to 1 and the expected number of the order from specific demands at the end of station 1 will converge to zero as the base-stock level increases to infinity. This implies that, in the case of $\lambda_s + \lambda_o < \mu_1$, we are able to find a base-stock level to satisfy the predetermined service qualities. When the capacity of the workstation is not enough to handle all the traffic, all of these three converge to constants as we increase the base-stock level. Equation (8.10) implies λ_e converges to $\mu_1 - \lambda_s$. Note that the specific demands will eventually be served. This means that the maximal capacity that the system can offer to the ordinary demands is the residual capacity, $\mu_1 - \lambda_s$. In this case, $(\mu_1 - \lambda_s)/\lambda_o$

can be considered as the upper bound of the fill rate and it can be used to check the feasibility of the system. Also note that, in this case, although both L_{N_1} and L_{B_2} converge to constants, from (8.2) the expected number of production orders for standard products in front of station 1 will go to infinity as the base-stock increases to infinity.

Let L_{B_1} be the expected number of production orders for standard products in front of station 1. From Proposition 8.1 and 8.2 we have the following limiting results for L_{B_1} for the case $\lambda_s + \lambda_o < \mu_1$. Note that L_{B_1} will diverge when $\lambda_s + \lambda_o \geq \mu_1$.

Proposition 8.3. *If* $\lambda_s + \lambda_o < \mu_1$,

$$L_{B_1} \to \frac{(\lambda_s + \lambda_o)/\mu_1}{1 - (\lambda_s + \lambda_o)/\mu_1}, \quad as\ S \to \infty. \tag{8.15}$$

It is intuitive that the right term in (8.15) is the expected number of customers in the system in an M/M/1 queue because all the ordinary demands will be satisfied as the stock level becomes large and the arrival rates of the production orders from the ordinary demands will be λ_o. For studying the response time of the specific demand, we consider the case when $\lambda_s + \lambda_o < \mu_1$.

We first express the respective response times at both stations by the recursive equations. Note that when a specific order arrives, it will wait for its custom product by sending an order (request) to the inventory of station 1 for a finished product, and, at the same time, station 1 will also send a production order to itself for a standard product.

Let $\{A_n, n = 1, 2, ...\}$ be the arrival process of the specific demands, where A_n denotes the arrival time of the nth specific demand. Let U_n be the interarrival time between the nth and $(n-1)$st arrivals; then, by our assumptions, U_ns are i.i.d. exponential random variables with rate λ_s. Note that $\{A_n, n = 1, 2, ...\}$ is also the arrival process of orders at the end of station 1. Let $\{A'_n, n = 1, 2, ...\}$ be the arrival process of production orders for the standard products in front of station 1. Note that these orders can be initiated by either specific demands or satisfied ordinary demands. Let U'_n be the interarrival time between the nth and $(n-1)$st arrivals and we approximate U'_ns as i.i.d. exponential random variables with rate $\lambda_s + \lambda_e$. Suppose that there are already d satisfied ordinary demands that left the system when the nth specific order arrives; then $A_n = A'_{n+d}$.

If the response time at station 1 of the nth specific order is positive, then it means that when the nth specific order arrives, there are no finished standard products available and it will wait for the product made by the $n + d - S$ production orders for the standard products. And, before it obtains this standard product, there will be no other ordinary demands that can be satisfied. Therefore, in this case, the response time of the nth specific demand at station 1, denoted by R^1_n, is

$$\begin{aligned} R^1_n &= A'_{n+d-S} + W'_{n+d-S} - A'_{n+d} \\ &= W'_{n+d-S} - \sum_{k=n+d-S+1}^{n+d} U'_n, \end{aligned} \tag{8.16}$$

where W_n' is the waiting time in the system of the nth production order for a standard product at station 1. We approximate the underlying queueing system of W_n's by an M/M/1 queue with arrival rate $\lambda_s + \lambda_e$ and service rate μ_1. Also note that $\sum_{k=n+d-S+1}^{n+d} U_n'$ is distributed as a gamma distribution with parameters S and $\lambda_s + \lambda_e$. Therefore, for $t > 0$, the density function for the response time, denoted by $f_{R^1}(t)$, can be obtained as

$$\begin{aligned} f_{R^1}(t) &= \int_{v=0}^{\infty} \frac{(\lambda_s+\lambda_e)^S v^{S-1}}{(S-1)!} e^{-(\lambda_s+\lambda_e)v} (\mu_1 - (\lambda_s+\lambda_e)) \, \mathrm{e}^{-(\mu_1-(\lambda_s+\lambda_e))(v+t)} dv \\ &= \left(\frac{\lambda_s+\lambda_e}{\mu_1}\right)^S (\mu_1 - (\lambda_s+\lambda_e)) \, \mathrm{e}^{-(\mu_1-(\lambda_s+\lambda_e))t}, \qquad t > 0. \end{aligned}$$

Furthermore, the probability that the response time is zero, denoted by $P(R^1 = 0)$, is equal to

$$1 - \left(\frac{\lambda_s+\lambda_e}{\mu_1}\right)^S. \tag{8.17}$$

Let R_n^2 denote the response time of the nth specific demand at station 2 and W_n^2 denote the waiting time in system of the nth combined order at station 2; then

$$R_n^2 = R_n^1 + W_n^2.$$

The arrival process of the combined orders to station 2 is not a Poisson process, thus we consider the queueing system corresponding to the combined orders at station 2 as a GI/M/1 queue. The waiting time in system of a combined order, denoted by W^2, has the density

$$f_{W^2}(t) = \mu_2(1-\alpha)e^{-\mu_2(1-\alpha)t}, \qquad t \geq 0,$$

where α is a solution of $\alpha = F^*(\mu_2(1-\alpha))$ and F^* is the Laplace transform of the interarrival time of a combined order to station 2 (see Kulkarni [9]). Because departures from station 1 may be triggered by ordinary demands or specific demands and only the departing specific demands will enter station 2, we approximate the arrival process to station 2 as the departure process of an M/M/1 base-stock inventory-queue with arrival rate λ_s and service rate μ_1. Form Buzacott, Price, and Shanthikumar [10], we have

$$\begin{aligned} F^*(\tau) = \left(1 - (\lambda_s/\mu_1)^{S+1}\right) \frac{\lambda_s}{\lambda_s+\tau} + (\lambda_s/\mu_1)^{S-1} \frac{\mu_1}{\mu_1+\tau} - (\lambda_s/\mu_1)^{S-1} \\ \cdot \left(1 - (\lambda_s/\mu_1)^2\right) \frac{\lambda_s+\mu_1}{\lambda_s+\mu_1+\tau}. \end{aligned}$$

The density of the response time of a specific demand is then

$$f_{R^2}(t) = \int_{v=0}^{t} f_{R^1}(v) f_{W^2}(t-v) \, dv + P(R^1 = 0) f_{W^2}(t).$$

We assume the independence of the response time at station 1 and the waiting time in the system at station 2. After some algebra we have, for $t > 0$,

$$\begin{aligned} f_{R^2}(t) = & \left(1 - \left(\frac{\lambda_s + \lambda_e}{\mu_1}\right)^S \left(\frac{\mu_2(1-\alpha)}{\mu_2(1-\alpha) - \mu_1 + (\lambda_s + \lambda_e)}\right)\right) \\ & \cdot (\mu_2(1-\alpha)) e^{-\mu_2(1-\alpha)t} + \left(\frac{\lambda_s + \lambda_e}{\mu_1}\right)^S (\mu_1 - (\lambda_s + \lambda_e)) \\ & \cdot \frac{\mu_2(1-\alpha)}{\mu_2(1-\alpha) - \mu_1 + (\lambda_s + \lambda_e)} e^{-(\mu_1 - (\lambda_s + \lambda_e))t} \end{aligned} \tag{8.18}$$

and the expected response times of specific demands

$$\begin{aligned} \mathrm{E}[R^2] &= \int_0^\infty t f_{R^2}(t) dt \\ &= \frac{1 - \left(\frac{\lambda_s + \lambda_e}{\mu_1}\right)^S \left(\frac{\mu_2(1-\alpha)}{\mu_2(1-\alpha) - \mu_1 + (\lambda_s + \lambda_e)}\right)}{(\mu_2(1-\alpha))} \\ &\quad + \frac{\left(\frac{\lambda_s + \lambda_e}{\mu_1}\right)^S \left(\frac{\mu_2(1-\alpha)}{\mu_2(1-\alpha) - \mu_1 + (\lambda_s + \lambda_e)}\right)}{\mu_1 - (\lambda_s + \lambda_e)}. \end{aligned}$$

Finally, we have a closed form for the c.d.f of R^2 as follows:

$$\begin{aligned} F_{R^2}(u) &= \int_0^u f_{R^2}(t) dt \\ &= \left(1 - \left(\frac{\lambda_s + \lambda_e}{\mu_1}\right)^S \left(\frac{\mu_2(1-\alpha)}{\mu_2(1-\alpha) - \mu_1 + (\lambda_s + \lambda_e)}\right)\right) \left(1 - e^{-\mu_2(1-\alpha)u}\right) \\ &\quad + \left(\frac{\lambda_s + \lambda_e}{\mu_1}\right)^S \left(\frac{\mu_2(1-\alpha)}{\mu_2(1-\alpha) - \mu_1 + (\lambda_s + \lambda_e)}\right) \\ &\quad \cdot \left(1 - e^{-(\mu_1 - (\lambda_s + \lambda_e))u}\right). \end{aligned} \tag{8.19}$$

Note that if T is our maximal lead time for a specific demand, then $F_{R^2}(T)$ will be the corresponding in-time rate.

8.3 Numerical Results

In this section, we conduct numerical studies to verify our results and we are then interested in finding an optimal base-stock level to minimize the total cost subject to the requirements on the fill rate and in-time rate. Based on our results of the

closed-form expressions for the fill rate (8.4) and in-time rate (8.19), we first verify our results by comparing our results and the results from simulations through examples (Examples 8.1 and 8.2). Note (8.4) can be expressed in terms of $\rho_s = \lambda_s/\mu_1$ and $\rho_o = \lambda_o/\mu_1$. In the following example, we test our fill rate results with the results from simulations based on various ρ_s and ρ_o.

Example 8.1. We consider four cases with $\rho_s = 0.5$ and 0.3 and $\rho_0 = 0.2$ and 0.4 under three base-stock levels, 1, 2, and 3, to verify our results on the fill rate, P_f. The comparison results are shown in Table 8.1. Our results (indicated by "Approx.") and those obtained from simulations (indicated by "Sim.") are very close to each other.

In the next example, we verify our approximations (indicated by "Approx.") on in-time rates with the results from simulations (indicated by "Sim.").

Example 8.2. Let $\lambda_o = 0.05$, $\lambda_s = 0.02$, $\mu_1 = 0.1$, and $\mu_2 = 0.075$. We assume that the requested lead time of the specific demand is 70. The comparisons between the results obtained from our approximations for the in-time rate and the expected response time $\mathrm{E}[R^2]$ on $S = 1, 2, 3, 4, 5$, and 6 are shown in Table 8.2.

In the following two examples, we verify our limiting results on P_f, L_{N_1}, and L_{B_2} obtained from Proposition 8.2. According to Proposition 8.2, we discuss this matter in two cases: $\lambda_s + \lambda_o < \mu_1$ in Example 8.3 and $\lambda_s + \lambda_o \geq \mu_1$ in Example 8.4.

Example 8.3 ($\lambda_s + \lambda_o < \mu_1$). Let $\lambda_o = 5$, $\lambda_s = 3$, $\mu_1 = 10$, and $\mu_2 = 11$. The results on various base-stock levels S are shown in Table 8.3. As we can see, P_f converges to 1; L_{N_1} is getting large and L_{B_2} converges to zero.

Table 8.1 Comparison results on fill rates P_f of ordinary demands on various ρ_s and ρ_o.

ρ_o	S		ρ_s		ρ_o	S		ρ_s	
			0.5	0.3				0.5	0.3
0.2	1	Sim.	0.4153	0.5850	0.4	1	Sim.	0.3573	0.4999
		Approx.	0.4167	0.5833			Approx.	0.3571	0.5000
	2	Sim.	0.6342	0.8082		2	Sim.	0.5400	0.7078
		Approx.	0.6343	0.8077			Approx.	0.5398	0.7083
	3	Sim.	0.7670	0.9073		3	Sim.	0.6505	0.8176
		Approx.	0.7615	0.9074			Approx.	0.6502	0.8172

Table 8.2 Comparison results on in-time rates and mean response times.

In-Time Rate			$\mathrm{E}[R^2]$		
S	Sim.	Approx.	S	Sim.	Approx.
1	0.941	0.940	1	26.539	26.565
2	0.943	0.938	2	25.371	25.533
3	0.944	0.942	3	24.338	24.340
4	0.951	0.948	4	22.954	23.138
5	0.953	0.953	5	22.046	22.051
6	0.962	0.960	6	21.040	21.132

Table 8.3 P_f, L_{N_1}, and L_{B_2} on various base-stock levels ($\lambda_o = 5$, $\lambda_s = 3$, $\mu_1 = 10$, $\mu_2 = 11$).

S	P_f	L_{N_1}	L_{B_2}	L_{B_1}	S	P_f	L_{N_1}	L_{B_2}	L_{B_1}
1	0.466	0.466	0.228	0.762	11	0.973	7.823	0.011	3.188
2	0.663	1.031	0.144	1.113	12	0.979	8.702	0.008	3.306
3	0.769	1.652	0.098	1.446	13	0.983	9.596	0.007	3.411
4	0.834	2.316	0.070	1.754	14	0.987	10.506	0.005	3.499
5	0.877	3.016	0.052	2.036	15	0.989	11.427	0.004	3.577
6	0.907	3.750	0.039	2.289	16	0.991	12.361	0.003	3.642
7	0.929	4.515	0.030	2.515	17	0.993	13.304	0.002	3.698
8	0.945	5.307	0.023	2.716	18	0.994	14.255	0.002	3.747
9	0.957	6.124	0.018	2.894	40	0.996	15.214	0.002	3.996
10	0.966	6.963	0.014	3.051	60	0.996	16.179	0.001	4.000

Table 8.4 P_f, L_{N_1}, and L_{B_2} on various base-stock levels ($\lambda_o = 20$, $\lambda_s = 7$, $\mu_1 = 10$, $\mu_2 = 11$).

S	P_f	L_{N_1}	L_{B_2}	L_{B_1}	S	P_f	L_{N_1}	L_{B_2}	L_{B_1}
1	0.100	0.100	2.100	3.000	11	0.150	0.238	1.983	12.745
2	0.132	0.167	2.025	3.858	12	0.150	0.238	1.983	13.745
3	0.143	0.204	1.998	4.794	13	0.150	0.238	1.983	14.745
4	0.147	0.223	1.988	5.765	14	0.150	0.238	1.983	15.745
5	0.149	0.231	1.985	6.754	15	0.150	0.238	1.983	16.745
6	0.149	0.235	1.984	7.749	16	0.150	0.238	1.983	17.745
7	0.149	0.237	1.983	8.746	17	0.150	0.238	1.983	18.745
8	0.150	0.237	1.983	9.746	18	0.150	0.238	1.983	19.745
9	0.150	0.238	1.983	10.745	19	0.150	0.238	1.983	20.745
10	0.150	0.238	1.983	11.745	20	0.150	0.238	1.983	21.745

Example 8.4 ($\lambda_s + \lambda_o \geq \mu_1$). Let $\lambda_o = 20$, $\lambda_s = 7$, $\mu_1 = 10$, and $\mu_2 = 11$. The situations with various base-stock levels S are in Table 8.4. P_f, L_{N_1}, and L_{B_2} all converge to the same constants as estimated in Proposition 8.2.

After verifying our estimations on the fill rate and in-time rate, in the next example we implement our results in finding the feasible base-stock levels where both requirements on the fill rate and in-time rate can be satisfied. In this example, we consider the case when $\lambda_s + \lambda_o < \mu_1$.

Example 8.5. Consider a system with $\lambda_o = 9$, $\lambda_s = 4$, $\mu_1 = 16$, and $\mu_2 = 15$. Suppose that the fill rate is required to be at least 0.9 and the in-time rate (with the required lead time 0.5) at least 0.95. We first try to find the base-stock levels where these qualities of services can be satisfied. The results on various base-stock levels are shown in Table 8.5. We can see that these qualities of services are satisfied if S is greater than or equal to 6.

Now, we apply some cost structure by defining the following costs. Let C_1 denote the penalty cost for each unsatisfied ordinary demand; Let C_2 denote the penalty cost for each unsatisfied specific demand and u be the maximal allowable lead time. Let

Table 8.5 Fill rates and in-time rates on various S ($\lambda_o = 9$, $\lambda_s = 4$, $\mu_1 = 16$, and $\mu_2 = 15$).

S	P_f	In-Time Rate	S	P_f	In-Time Rate
1	0.4800	0.9683	11	0.9724	0.9733
2	0.6731	0.9560	12	0.9779	0.9767
3	0.7757	0.9505	13	0.9822	0.9796
4	0.8381	0.9487	14	0.9857	0.9822
5	0.8795	0.9501	15	0.9885	0.9845
6	0.9083	0.9530	16	0.9907	0.9864
7	0.9291	0.9571	17	0.9925	0.9880
8	0.9446	0.9613	18	0.9939	0.9894
9	0.9564	0.9655	19	0.9950	0.9905
10	0.9654	0.9696	20	0.9960	0.9915

Table 8.6 TCs on various base-stock levels ($\lambda_o = 9$, $\lambda_s = 4$, $\mu_1 = 16$, $\mu_2 = 15$, $C_1 = \$5$, $C_2 = \$15$, and $C_3 = \$1$).

S	TC	S	TC
1	25.782	11	10.539
2	18.394	12	10.947
3	14.723	13	11.455
4	12.672	14	12.036
5	11.415	15	12.680
6	10.662	16	13.385
7	10.225	17	14.138
8*	10.049	18	14.929
9	10.067	19	15.759
10	10.236	20	16.611

C_3 denote the inventory cost rate per each finished standard product in stock. Then, the total cost rate, TC, is expressed as

$$TC = \left(1 - P_f\right) \lambda_o C_1 + (1 - F_{R^2}(u)) \lambda_s C_2 + LC_3. \tag{8.20}$$

Following Example 8.5, we are interested in finding an optimal base-stock level minimizing the total cost subject to the requirements on the fill rate and in-time rate.

Example 8.6. We consider the same case of $\lambda_o = 9$, $\lambda_s = 4$, $\mu_1 = 16$, and $\mu_2 = 15$ with $C_1 = \$5$, $C_2 = \$15$, and $C_3 = \$1$. Suppose the qualities of service are that the fill rate must be at least 0.9 and the in-time rate (with the required lead time 0.5) must be at least 0.95. The TCs on various base-stock levels S are shown in Table 8.6 and the corresponding figure is Fig. 8.3. From Example 8.5, we know that the feasible base-stock levels are those greater than or equal to 6. Among these feasible levels, we then obtain the optimal base-stock level at $S = 8$ with minimum total cost rate 10.049.

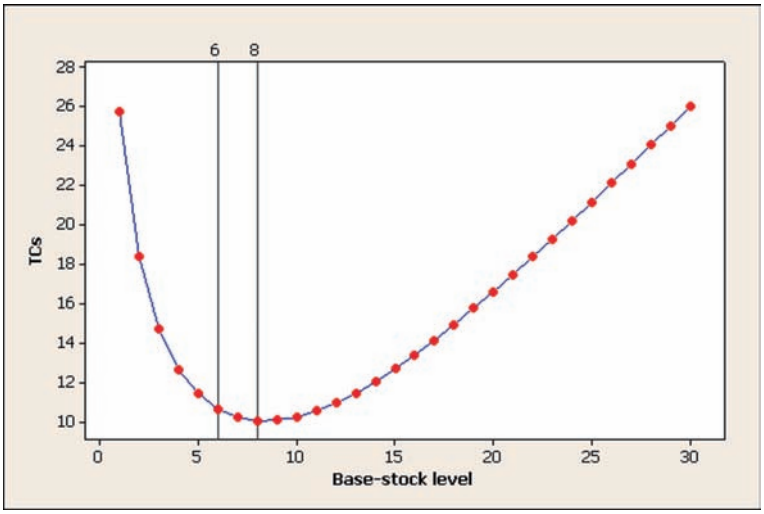


Fig. 8.3 TCs on various base-stock levels ($\lambda_o = 9$, $\lambda_s = 4$, $\mu_1 = 16$, $\mu_2 = 15$, $C_1 = \$5$, $C_2 = \$15$, and $C_3 = \$1$).

8.4 Conclusions

In this chapter, we consider a two-station MTO/MTS hybrid production system dealing with ordinary and specific demands. We are interested in determining the fill rate of ordinary demands and response times of specific demands. By assuming the Markovian model, for station 1, we give the closed-form for the fill rate and some limiting results as the base-stock level increases, however, because of the intractability in analyzing station 2, we approximate station 2 as a GI/M/1 queue. The corresponding closed-form for the approximated in-time rate is obtained. These results of the fill rate and in-time rate can assist management in determining the optimal base-stock level efficiently. In future study, we may consider a multistation system for both the process producing the standard products and the process performing the additional work for the custom work.

Acknowledgments This work was supported by the National Science Council of Taiwan, ROC under the grant Contract NSC-95-2221-E-033-032.

References

1. C. A. Soman, D. P. van Donk, and G. Gaalman, Combined make-to-order and make-to-stock in a food production system, *International Journal of Production Economics*, vol. 90, pp. 223–235, 2004.

2. A. Krishnamurthy, R. Suri, and M. Vernon, Re-examining the performance of MRP and kanban material control strategies for multi-product flexible manufacturing systems, *The International Journal of Flexible Manufacturing Systems*, vol. 16, pp. 123–150, 2004.
3. I. J. B. F. Adan and J. ven der Wal, Combining make to order and make to stock, *OR Spektrum*, vol. 20, pp. 73–81, 1998.
4. V. Nguyen, A multiclass hybrid production center in heavy traffic, *Operations Research*, vol. 46, Supp. 3, pp. S13–S25, 1998.
5. A. Federguruen and Z. Katalan, Impact of adding a make-to-order item to a make-tostock production system, *Management Science*, vol. 45, no. 7, pp. 980–994, 1999.
6. S. Carr and I. Duenyas, Optimal admission control and sequencing in a Make-to-Stock/Make-to-Order production system, *Operations Research*, vol. 48, no. 5, pp. 709–720, 2000.
7. A. Arreola-Risa and G. A. DeCroix, Make-to-order versus make-to-stock in a production-inventory system with general production times, *IIE Transactions*, vol. 30, no. 8, pp. 705–713, 1998.
8. S. Rajagopalan, Make to order or make to stock: model and application, *Management Science*, vol. 48, no. 2, pp. 241–256, 2002.
9. V. G. Kulkarni, *Modeling, Analysis, Design, and Control of Stochastic Systems*. New York: Springer, 1999.
10. J. A. Buzacott, S. M. Price, and J. G. Shanthikumar, Service level in multistage MRP and base stock controlled production systems, in: G. Fandel, T. Gulledge, A. Jones (Ed.), *New Directions for Operations for Operations Research in Manufacturing System*, pp. 445–464. New York: Springer, 1992.

Chapter 9
Analysis of an M/M/*c*/N Queueing System with Balking, Reneging, and Synchronous Vacations

Dequan Yue and Wuyi Yue

Abstract In this chapter, we present an analysis for an M/M/*c*/N queueing system with simultaneous balking, reneging, and synchronous vacations of servers. By using the blocked matrix method, we obtain the steady-state probability vector presented by the inverses of two matrices. The computing of the inverses of the two matrices is discussed. Then, we calculate the steady-state probabilities by using the elements of the inverses of the two matrices. We also derive the conditional stationary distribution of the queue length and waiting time.

9.1 Introduction

Many practical queueing systems, especially those with balking and reneging, have been widely applied to many real-life problems such as situations involving impatient telephone switchboard customers, hospital emergency rooms' handling of critical patients, and perishable goods storage inventory systems. Balking and reneging are not only common phenomena in queues arising in daily activities, but also in telecommunication networks and in various machine repair models.

Ke [1] gave an example of the occurrence of balking in the operational model of WWW servers. An interesting example of the occurrence of balking and reneging in air defense systems was given in Ancker and Gafarian [2]. For other examples of articles that address queueing systems which use balking and reneging, interested readers may refer to [1]– [3], and the references therein.

D. Yue
College of Sciences, Yanshan University, Qinhuangdao 066004, China
e-mail: ydq@ysu.edu.cn

W. Yue
Department of Intelligence and Informatics, Konan University, Kobe 658-8501, Japan
e-mail: yue@konan-u.ac.jp

W. Yue et al. (eds.), *Advances in Queueing Theory and Network Applications,*

Haghighi, Medhi, and Mohanty [4] derived the steady-state probabilities for multiserver M/M/c queues with balking and reneging. Abou-El-Ata and Hariri [5] analyzed multiserver M/M/c/N queues where balking and reneging were applied and derived the steady-state probabilities. Wang and Chang [6] extended this work to study an M/M/c/N queue with balking, reneging, and server breakdowns. They derived the steady-state probabilities in matrix form and developed a cost model to determine the optimal number of servers.

In many real-world queueing systems, servers may become unavailable for a random period of time when there are no customers waiting in line at a service completion instant. This random period of server absence, often called a server vacation, can represent the time when the server is performing some secondary task. Single-server queueing models with vacations have been studied by many researchers and have been found to be applicable in analyzing numerous real-world queueing situations, such as flexible manufacturing systems, service systems, and telecommunication systems. Several excellent surveys on these vacation models have been done by Doshi [7], [8] and Takagi [9].

Multiple-server vacation models are more flexible and applicable in practice than their single-server counterparts. However, there are only a few studies on multiple-server vacation models in the vacation model literature due to the complexity of the systems. The M/M/c queue with exponentially distributed vacations was first studied by Levy and Yechiali [10]. In the system of [10], all the servers take a vacation together when the system is completely empty. Because all these servers take vacations simultaneously, these vacations are called "synchronous vacations".

Tian, Li, and Cao [11] modeled the M/M/c vacation systems of [10] as a quasi birth-and-death (QBD) process, and presented a more detailed analysis. They proved several conditional stochastic decomposition results for the queue length and the customer waiting time. Recently, Zhang and Tian [12] extended the model presented in [11] by studying an M/M/c queueing system with synchronous vacations of partial servers. In the system of [12], some servers take vacations when they become idle and other servers are always available for serving arriving customers. They call this type of model the "partial server vacation model".

It may be remarked here that all the studies on multiple-server vacation models mentioned above assume availability of infinite buffer space in front of the servers. However, finite buffer queues are more common in certain practical applications. Yue, Yue, and Sun [13] considered the balking and reneging phenomena in a finite buffer M/M/c/N queueing system with the same vacation policy as in [12]. They obtained the steady-state probability vector presented by the inverses of three matrices. However, they did not obtain the explicit expressions for the inverses of these three matrices.

In this chapter, we consider a special case of the partial-server vacation model in [13]. We study a finite buffer M/M/c/N queueing system with balking, reneging, and the same synchronous vacation policy as in [11]. The Markov chain underlying the queueing system in this chapter is a level-dependent quasi birth-and-death (LDQBD) process. The matrix-geometric solution method applied in [11] and [12]

cannot be used to obtain the stationary probabilities of the system in this chapter. The prevailing method applied to obtain the stationary probabilities of a LDQBD process is to develop some approximations to diminish the level dependence at higher levels. However, in this chapter, we present a different approach to obtain the stationary probabilities of the system.

The rest of this chapter is organized as follows. In Sect. 9.2, we give a description of the queueing model. In Sect. 9.3, we derive the steady-state equations and obtain the steady-state probability vector presented by the inverses of two matrices with the blocked matrix method. We also discuss the computing of the inverses of the two matrices. Then, we calculate the steady-state probabilities by using the elements of the inverses of the two matrices. In Sect. 9.4, we derive the conditional stationary distribution of the queue length and waiting time. Conclusions are given in Sect. 9.5.

9.2 System Model

In this chapter, we consider a finite buffer M/M/*c*/N queueing system with balking, reneging, and synchronous vacations in all servers. The system capacity is finite N. The assumptions of the system model are as follows:

(1) Customers arrive according to a Poisson process with arrival rate λ. There are c servers in the system. The service time for each server is assumed to be distributed according to an exponential distribution with service rate μ.
(2) If some servers are busy, and some servers are idle, then a customer who on arrival joins the system will be serviced immediately. If all servers are either busy or taking a vacation, then a customer who on arrival finds n customers in the system, either decides to enter the queue with probability b_n or balks with probability $1-b_n$, $0 \le b_{n+1} \le b_n < 1$, $0 \le n \le N-1$, $b_N = 0$.
(3) All servers take synchronous vacations when the system is completely empty at a service completion instant. At a vacation completion instant, if the system is still empty, all the servers take another vacation together; otherwise, they return to serve the queue. The vacation time is assumed to be exponentially distributed with mean $1/\eta$.
(4) After joining the queue, in the case where all the servers are occupied each customer will wait a certain length of time T_r for service to begin before he gets impatient and leaves the queue without receiving service. This time T_r is assumed to be distributed according to an exponential distribution with mean $1/\alpha$.
(5) The service order is assumed to be on a First-Come First-Served (FCFS) basis and the interarrival times, service times, and vacations are mutually independent.

9.3 Steady-State Probability

In this section, we first develop steady-state probability equations by using the Markov process. Then, we derive the steady-state probabilities by using the blocked matrix method.

9.3.1 Steady-State Equations

Let $L(t)$ be the number of customers in the system at time t and let

$$J(t) = \begin{cases} 0, & \text{servers are on vacation at time } t \\ 1, & \text{servers are not on vacation at time } t. \end{cases}$$

Then, $\{L(t), J(t)\}$ is a Markov process with state space:

$$\Omega = \{(i,0) : i = 0,1,\ldots,N\} \cup \{(i,1) : i = 1,2,\ldots,N\}.$$

The steady-state probabilities of the system are defined as follows:

$$P_0(n) = \lim_{t\to\infty} P\{L(t) = n, J(t) = 0\}, \qquad n = 0,1,\ldots,N,$$
$$P_1(n) = \lim_{t\to\infty} P\{L(t) = n, J(t) = 1\}, \qquad n = 1,2,\ldots,N.$$

By applying the Markov process theory, we can obtain the following set of steady-state probability equations:

$$\begin{aligned}
s_1 P_1(1) + v_1 P_0(1) &= u_0 P_0(0), \\
u_{n-1} P_0(n-1) + v_{n+1} P_0(n+1) &= w_n P_0(n), \qquad n = 1,2,\ldots,N-1, \\
u_{N-1} P_0(N-1) &= w_N P_0(N), \\
\eta P_0(1) + s_2 P_1(2) &= (s_1 + t_1) P_1(1), \\
\eta P_0(n) + t_{n-1} P_1(n-1) + s_{n+1} P_1(n+1) &= (s_n + t_n) P_1(n), \qquad n = 2,3,\ldots,N-1, \\
\eta P_0(N) + t_{N-1} P_1(N-1) &= s_N P_1(N), \\
\sum_{n=0}^{N} P_0(n) + \sum_{n=1}^{N} P_1(n) &= 1,
\end{aligned}$$

where

$$u_i = \lambda b_i, \qquad i = 0,1,\ldots,N-1,$$

$$v_i = i\alpha, \qquad i = 1,2,\ldots,N,$$

$$w_i = \begin{cases} v_i + \eta + u_i, & i = 1,2,\ldots,N-1 \\ \eta + v_N, & i = N, \end{cases}$$

$$s_i = \begin{cases} i\mu, & i = 1,2,\ldots,c \\ c\mu + (i-c)\alpha, & i = c+1,c+2,\ldots,N, \end{cases}$$

$$t_i = \begin{cases} \lambda, & i = 1,2,\ldots,c-1 \\ \lambda b_i, & i = c,c+1,\ldots,N-1. \end{cases}$$

9.3.2 Matrix Solution

In the following, we derive the steady-state probabilities by using the blocked matrix method. Let

$$\mathbf{P} = (P_0(0), P_0(1), \ldots, P_0(N), P_1(1), P_1(2), \ldots, P_1(N))$$

be the steady-state probability vector. Then, the steady-state probability equations above can be rewritten in matrix form as follows:

$$\begin{cases} \mathbf{PQ} = 0 \\ \mathbf{Pe} = 1, \end{cases} \tag{9.1}$$

where $\mathbf{e} = (1,1,\ldots,1)^T$ is a $(2N+1)\times 1$ vector, and the transition rate matrix $\mathbf{Q}$ of the Markov process has the blocked matrix structure:

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} & \mathbf{Q}_{13} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} & \mathbf{Q}_{23} \\ \mathbf{Q}_{31} & \mathbf{Q}_{32} & \mathbf{Q}_{33} \end{pmatrix}.$$

Each matrix $\mathbf{Q}_{lk}$ $(l,k = 1,2,3)$ is given as follows:

$$\mathbf{Q}_{11} = (-u_0, v_1, 0, \ldots, 0)^T, \qquad \mathbf{Q}_{31} = (s_1, 0, \ldots, 0)^T,$$
$$\mathbf{Q}_{22} = (0, 0, \ldots, v_N, -w_N), \qquad \mathbf{Q}_{23} = (0, 0, \ldots, 0, \eta),$$

$$\mathbf{Q}_{12} = \begin{pmatrix} u_0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -w_1 & u_1 & 0 & \cdots & 0 & 0 & 0 \\ v_2 & -w_2 & u_2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -w_{N-2} & u_{N-2} & 0 \\ 0 & 0 & 0 & \cdots & v_{N-1} & -w_{N-1} & u_{N-1} \end{pmatrix},$$

$$\mathbf{Q}_{13} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ \eta & 0 & 0 & \cdots & 0 & 0 \\ 0 & \eta & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \eta & 0 \end{pmatrix},$$

$$\mathbf{Q}_{33} = \begin{pmatrix} -(s_1+t_1) & t_1 & 0 & \cdots & 0 & 0 & 0 \\ s_2 & -(s_2+t_2) & t_2 & \cdots & 0 & 0 & 0 \\ 0 & s_3 & -(s_3+t_3) & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & s_{N-1} & -(s_{N-1}+t_{N-1}) & t_{N-1} \\ 0 & 0 & 0 & \cdots & 0 & s_N & -s_N \end{pmatrix},$$

where $\mathbf{Q}_{11}$ and $\mathbf{Q}_{31}$ are $N \times 1$ vectors, $\mathbf{Q}_{12}$, $\mathbf{Q}_{13}$, and $\mathbf{Q}_{33}$ are $N \times N$ matrices, $\mathbf{Q}_{22}$ and $\mathbf{Q}_{23}$ are $1 \times N$ vectors, $\mathbf{Q}_{21} = 0$ is a constant, and $\mathbf{Q}_{32} = \mathbf{0}$ is an $N \times N$ matrix.

The four submatrices $\mathbf{Q}_{11}$, $\mathbf{Q}_{12}$, $\mathbf{Q}_{21}$, and $\mathbf{Q}_{22}$ give the transition rates during the vacation period. For example, the submatrix $\mathbf{Q}_{12}$ gives the transition rates from vacation state $(0,i)$ to vacation state $(0,j)$, $i = 0,1,\ldots,N-1$, $j = 1,2,\ldots,N$. The two submatrices $\mathbf{Q}_{13}$ and $\mathbf{Q}_{23}$ give the transition rates from a vacation state to a busy state. For example, the submatrix $\mathbf{Q}_{13}$ gives the transition rates from vacation state $(0,i)$ to busy state $(1,j)$, $i = 0,1,\ldots,N-1$, $j = 1,2,\ldots,N$. The two submatrices $\mathbf{Q}_{31}$ and $\mathbf{Q}_{32}$ give the transition rates from a busy state to a vacation state. The submatrix $\mathbf{Q}_{33}$ gives the transition rates during the busy period.

In order to solve (9.1) by using the blocked matrix method, we consider computing the inverses of the matrices $\mathbf{Q}_{12}$ and $\mathbf{Q}_{33}$.

Let c_{ij} be the (ij) element of the inverse matrix $\mathbf{Q}_{12}^{-1}$, i, $j = 1,2,\ldots,N$. Let d_{ij} be the (ij) element of the inverse matrix $\mathbf{Q}_{33}^{-1}$, i, $j = 1,2,\ldots,N$. We then have the following lemmas.

Lemma 9.1. *The matrix $\mathbf{Q}_{12}$ is invertible. For $j = 1,2,\ldots,N$, the elements of the inverse matrix $\mathbf{Q}_{12}^{-1}$ are given by*

$$c_{ij} = \begin{cases} 0, & i = 1,2,\ldots,j-1 \\ \dfrac{1}{u_{j-1}}, & i = j \\ k_{ij}\dfrac{1}{u_{j-1}}, & i = j+1, j+2,\ldots,N, \end{cases} \tag{9.2}$$

where k_{ij} is given by the following recursive relations

$$k_{ij} = \frac{w_{i-1}}{u_{i-1}}k_{i-1j} - \frac{v_{i-1}}{u_{i-1}}k_{i-2j}, \qquad i = j+1, j+2,\ldots,N, \tag{9.3}$$

where $k_{jj} = 1$ and $k_{j-1j} = 0$.

Proof. See Appendix. □

Remark 1. For the special case where $\alpha = 0$ (i.e., no reneging occurs in the system) the closed-form expression for the $\mathbf{Q}_{12}^{-1}$ can be obtained from Lemma 9.1. Let $\alpha = 0$ in Lemma 9.1; then we have the following recursive relation:

$$k_{ij} = \frac{w_{i-1}}{u_{i-1}}k_{i-1j}, \qquad i = j+1, j+2,\ldots,N.$$

Hence, we get the closed-form expression for k_{ij} as follows:

$$k_{ij} = \frac{w_{i-1}w_{i-2}\cdots w_j}{u_{i-1}u_{i-2}\cdots u_j}, \qquad i = j+1, j+2,\ldots,N.$$

Lemma 9.2. *The matrix $\mathbf{Q}_{33}$ is invertible. For $j = 1,2,\ldots,N$, the elements of the inverse matrix $\mathbf{Q}_{33}^{-1}$ are given by*

$$d_{ij} = \begin{cases} -\displaystyle\sum_{k=1}^{i} \frac{t_k t_{k+1}\ldots t_{j-1}}{s_k s_{k+1}\ldots s_j}, & i = 1,2,\ldots,j-1 \\ -\displaystyle\sum_{k=1}^{j-1} \frac{t_k t_{k+1}\ldots t_{j-1}}{s_k s_{k+1}\cdots s_j} - \frac{1}{s_j}, & i = j, j+1,\ldots,N. \end{cases} \tag{9.4}$$

The empty summation $\sum_{k=1}^{0}$ is defined to be zero.

Proof. See Appendix. □

In the following, we derive the steady-state probabilities from (9.1). To accommodate the partitioned blocked structure of $\mathbf{Q}$, we partition the steady-state probability vector into segments accordingly as follows:

$$\mathbf{P} = (\mathbf{P}_0, P_0(N), \mathbf{P}_1),$$

where

$$\mathbf{P}_0 = (P_0(0), P_0(1), \ldots, P_0(N-1)),$$

$$\mathbf{P}_1 = (P_1(1), P_1(2), \ldots, P_1(N)).$$

Theorem 9.1. *The segments of the steady-state probability vector are given by*

$$\mathbf{P}_0 = -P_0(N)\mathbf{Q}_{22}\mathbf{Q}_{12}^{-1}, \tag{9.5}$$

$$\mathbf{P}_1 = -P_0(N)(\mathbf{Q}_{23} - \mathbf{Q}_{22}\mathbf{Q}_{12}^{-1}\mathbf{Q}_{13})\mathbf{Q}_{33}^{-1}, \tag{9.6}$$

where

$$P_0(N) = \left\{1 - \mathbf{Q}_{22}\mathbf{Q}_{12}^{-1}\mathbf{e}_N - (\mathbf{Q}_{23} - \mathbf{Q}_{22}\mathbf{Q}_{12}^{-1}\mathbf{Q}_{13})\mathbf{Q}_{33}^{-1}\mathbf{e}_N\right\}^{-1} \tag{9.7}$$

and $\mathbf{e}_N = (1,1,\ldots,1)^T$ *is an* $N \times 1$ *vector.*

Proof. Based on the partitions of the vector $\mathbf{P}$, (9.1) can be rewritten as

$$\mathbf{P}_0\mathbf{Q}_{11} + \mathbf{P}_1\mathbf{Q}_{31} = 0, \tag{9.8}$$

$$\mathbf{P}_0\mathbf{Q}_{12} + P_0(N)\mathbf{Q}_{22} = 0, \tag{9.9}$$

$$\mathbf{P}_0\mathbf{Q}_{13} + P_0(N)\mathbf{Q}_{23} + \mathbf{P}_1\mathbf{Q}_{33} = 0, \tag{9.10}$$

$$\mathbf{P}_0\mathbf{e}_N + P_0(N) + \mathbf{P}_1\mathbf{e}_N = 1. \tag{9.11}$$

From Lemma 9.1 and (9.9), we have

$$\mathbf{P}_0 = -P_0(N)\mathbf{Q}_{22}\mathbf{Q}_{12}^{-1}. \tag{9.12}$$

Substituting (9.12) into (9.10), from Lemma 9.2, we have

$$\mathbf{P}_1 = -P_0(N)(\mathbf{Q}_{23} - \mathbf{Q}_{22}\mathbf{Q}_{12}^{-1}\mathbf{Q}_{13})\mathbf{Q}_{33}^{-1}, \tag{9.13}$$

where $P_0(N)$ can be obtained as the expression given in (9.7) by substituting (9.12) and (9.13) into (9.11). This completes the proof of Theorem 9.1. □

Theorem 9.2. *The steady-state probabilities are given by*

$$P_0(j) = \frac{-\beta_{j+1}}{\triangle}, \qquad j = 0,1,\ldots,N-1, \tag{9.14}$$

$$P_0(N) = -\frac{1}{\triangle}, \tag{9.15}$$

$$P_1(j) = -\frac{\eta}{\triangle}\left(d_{Nj} - \sum_{i=1}^{N-1} d_{ij}\beta_{i+1}\right), \qquad j = 1,2,\ldots,N, \tag{9.16}$$

where

$$\Delta = 1 - \sum_{j=1}^{N}\beta_j - \eta\sum_{j=1}^{N}\left(d_{Nj} - \sum_{i=1}^{N-1} d_{ij}\beta_{i+1}\right), \tag{9.17}$$

$$\beta_j = \begin{cases} v_N c_{N-1\,j} - w_N c_{Nj}, & j = 1,2,\ldots,N-1 \\ -w_N c_{NN}, & j = N, \end{cases} \tag{9.18}$$

c_{ij} and d_{ij} are given by Lemma 9.1 and Lemma 9.2.

Proof. Define

$$\mathbf{Q}_{22}\mathbf{Q}_{12}^{-1} = (\beta_1, \beta_2, \ldots, \beta_N). \tag{9.19}$$

Then, from Lemma 9.1, β_j $(j = 1,2,\ldots,N)$ can be obtained as the expression given in (9.18). Note that

$$\mathbf{Q}_{22}\mathbf{Q}_{12}^{-1}\mathbf{Q}_{13} = \eta(\beta_2, \beta_3, \ldots, \beta_N, 0);$$

we have

$$\mathbf{Q}_{23} - \mathbf{Q}_{22}\mathbf{Q}_{12}^{-1}\mathbf{Q}_{13} = -\eta(\beta_2, \beta_3, \ldots, \beta_N, -1). \tag{9.20}$$

Then, from Theorem 9.1, (9.19) and (9.20), we can derive (9.14)–(9.17). This completes the proof of Theorem 9.2. □

9.3.3 Some Special Cases

In the following, we present some special cases of our model. Some of them are existing models in the literature.

(1) If $\eta = \infty$ (i.e., the servers do not take vacations) and

$$b_n = \begin{cases} 1, & 0 \le n \le c \\ \dfrac{\beta\left(1 - \dfrac{1}{N}(n-c+1)\right)}{(n-c+2)^m}, & c \le n \le N, \end{cases}$$

then our model becomes the model studied by Abou-El-Ata and Hariri [5]: M/M/*c*/N queue with balking and reneging.

(2) If $\alpha = 0$ (i.e., customers do not renege), then our model becomes the model M/M/*c*/N queue with balking and synchronous vacation of all servers.

(3) If $N = \infty$, $\alpha = 0$, and $b_i = 1$, $i = 0, 1, \ldots$ (i.e., customers do not balk or renege), then our model becomes the model studied by Tian et al. [11]: M/M/*c*/∞ queue with synchronous vacation of all servers.

(4) If $c = 1$, then our model becomes the model studied by Yue et al. [14]: M/M/1/N queue with balking, reneging, and multiple vacations.

9.4 Conditional Distributions of Queue Length and Waiting Time

In an M/M/c multiple-server vacation system, Tian et al. [11] investigated the conditional stationary distribution of the queue length and waiting time under the condition when all servers are busy. They presented in such a system a conditional stochastic decomposition property for steady-state queue length and waiting time. In this section, we derive the conditional stationary distribution of the queue length and waiting time for the system studied in this chapter.

Let $P(Q_c = j)$, $j = 0, 1, \ldots, N-c$, represent the conditional stationary distribution of the queue length given that all servers are busy. It is given in the following theorem.

Theorem 9.3. *The conditional stationary distribution of the queue length is given by*

$$P(Q_c = j) = \frac{d_{Nj+c} - \sum_{i=1}^{N-1} d_{ij+c}\beta_{i+1}}{\sum_{j=c}^{N}\left(d_{Nj} - \sum_{i=1}^{N-1} d_{ij}\beta_{i+1}\right)}, \qquad j = 0, 1, \ldots, N-c, \tag{9.21}$$

where d_{ij} and β_j are given in Lemma 9.2 and (9.18), respectively.

Proof. From Theorem 9.2, the probability that all servers are busy is

$$\sum_{j=c}^{N} P_1(j) = -\frac{\eta}{\Delta}\sum_{j=c}^{N}\left(d_{Nj} - \sum_{i=1}^{N-1} d_{ij}\beta_{i+1}\right). \tag{9.22}$$

Note that

$$P(Q_c = j) = \frac{P_1(j+c)}{\sum_{j=c}^{N} P_1(j)}, \qquad j = 0, 1, \ldots, N-c \tag{9.23}$$

and substituting the probability given in (9.22) and the probability $P_1(j)$ given by Theorem 9.2 into (9.23), we can get the conditional distribution of the queue length given by (9.21). □

In the following, we consider the conditional distribution of the waiting time under the condition that all servers are busy when a customer on arrival joins the queue.

Let B_j represent the event that there are j customers in front of the new customer who on arrival joins the queue, and all the servers are busy. Under the assumption B_j, the c customers are in service and the other $j-c$ customers are waiting for service. Let T_j be the time remaining until the number of customers j diminishes by $j-1$ because of the completion of a customer's service or a customer's reneging, $j = c$,

$c+1, \ldots, N-1$. Because both the service time and the waiting time of a customer before he reneges are exponentially distributed, T_j is exponentially distributed with the distribution function given by

$$H_j(t) = 1 - e^{-\theta_j t}, \qquad t \geq 0, \qquad j = c, c+1, \ldots, N-1 \tag{9.24}$$

and the Laplace-Stieltjes transformation (LST) given by

$$H_j^*(s) = \frac{\theta_j}{\theta_j + s}, \quad s \geq 0, \qquad j = c, c+1, \ldots, N-1, \tag{9.25}$$

where $\theta_j = c\mu + (j-c)\alpha$, $j = c, c+1, \ldots, N-1$. It is easy to see that the random variables $T_c, T_{c+1}, \ldots, T_{N-1}$ are mutually independent because of the "no memory" property of the exponential distribution.

Let $\gamma_j = P(T_r > T_j + T_{j-1} + \cdots + T_c)$ and $\Phi_j(t) = P(T_j + T_{j-1} + \cdots + T_c \leq t)$, $j = c, c+1, \ldots, N-1$. Then, γ_j is the probability that the new customer on arrival joins the queue and waits in the queue until he acquires service under the condition B_j. We then have the following lemma.

Lemma 9.3.

$$\gamma_j = \frac{c\mu}{c\mu + (j+1-c)\alpha}, \qquad j = c, c+1, \ldots, N-1 \tag{9.26}$$

and

$$\Phi_j(t) = 1 - \sum_{k=c}^{j} \delta_{jk} e^{-\delta_{jk} t}, \qquad j = c, c+1, \ldots, N-1, \quad t \geq 0, \tag{9.27}$$

where

$$\delta_{jk} = \prod_{i=c, i \neq k}^{j} \frac{\theta_i}{\theta_i - \theta_k}, \quad k = c, c+1, \ldots, j, \quad j = c, c+1, \ldots, N-1. \tag{9.28}$$

Proof.

$$\begin{aligned} \gamma_j &= P(T_r > T_j + T_{j-1} + \cdots + T_c) \\ &= P(T_r > T_j) P(T_r - T_j > T_{j-1} + T_{j-2} + \cdots + T_c | T_r > T_j) \\ &= P(T_r > T_j) P(\tilde{T}_r > T_{j-1} + T_{j-2} + \cdots + T_c), \quad j = c, c+1, \ldots, N-1, \end{aligned} \tag{9.29}$$

where $\tilde{T}_r = [T_r - T_j | T_r > T_j]$ has the same exponential distribution as T_r because of the "no memory" property of the exponential distribution. It is easy to see that

$$P(T_r > T_j) = \frac{\theta_j}{\theta_j + \alpha}. \tag{9.30}$$

Hence, by the recursive relation of (9.29), we get the first result of Lemma 9.3.

Note that the random variables T_c, $T_{c+1},\ldots,T_{N-1}$ are mutually independent. $\Phi_j(t)$ has the LST as follows:

$$\Phi_j^*(s)=\prod_{k=c}^{j}H_k^*(s). \tag{9.31}$$

Substituting (9.25) into (9.31), we get

$$\begin{aligned}\Phi_j^*(s)&=\prod_{k=c}^{j}\frac{\theta_k}{\theta_k+s}\\&=\sum_{k=c}^{j}\delta_{jk}\frac{\theta_k}{\theta_k+s},\qquad j=c,c+1,\ldots,N-1.\end{aligned} \tag{9.32}$$

Taking the reverse of the LST for the two sides of (9.32), we get the second result of Lemma 9.3. □

Let $W_c(t)$ represent the distribution of the conditional waiting time given that all the servers are busy when a customer on arrival joins the queue. Let q_j be the stationary probability that there are j customers in the system under the condition that all the servers are busy when a customer on arrival joins the queue. Note that $b_jP_1(j)$ represents the probability that there are j customers in the system when a customer on arrival joins the queue. It is easy to see that

$$q_j=\frac{b_jP_1(j)}{\sum_{j=c}^{N-1}b_jP_1(j)},\qquad j=c,c+1,\ldots,N-1, \tag{9.33}$$

where $P_1(j)$ is given by Theorem 9.2.

Next, we have the following theorem.

Theorem 9.4. *The distribution of the conditional waiting time is given by*

$$W_c(t)=1-\sum_{j=c}^{N-1}q_j\gamma_j\sum_{k=c}^{j}\delta_{jk}e^{-\delta_{jk}t}-\sum_{j=c}^{N-1}q_j(1-\gamma_j)e^{-\alpha t}, \tag{9.34}$$

where γ_j, δ_{jk}, and q_j are given by (9.26), (9.28), and (9.33), respectively.

Proof. The conditional waiting time has the following distribution:

$$W_c(t)=\sum_{j=c}^{N-1}q_jP(W\le t|B_j), \tag{9.35}$$

where W represents the waiting time and B_j represents the event that there are j customers in front of the new customer who on arrival joins the queue, and all the servers are busy. Let F_1 and F_2 be the events that the customer either reneges or does not renege when the customer on arrival joins the queue, respectively. Then, we have

$$P(W \le t|B_j) = P(F_1|B_j)P(W \le t|B_j,F_1) + P(F_2|B_j)P(W \le t|B_j,F_2)$$
$$= (1-\gamma_j)(1-e^{-\alpha t}) + \gamma_j \Phi_j(t). \tag{9.36}$$

Thus, by Lemma 9.3, we get the result of Theorem 9.4. □

Remark 2. Based on Theorem 9.2, we can obtain some other performance measures such as the expected number of customers in the system, the expected number of servers that are busy, the average rate of customer loss due to impatience, and so on. The stationary distribution of waiting time can also be obtained from conditioning on every state $(i, j) \in \Omega$. However, these performance measures and the stationary distribution have very complex expressions. Hence, we have omitted the details from this discussion.

9.5 Conclusions

In this chapter, we studied a finite buffer M/M/*c*/N queueing system with balking, reneging, and the synchronous vacations of all servers. By using the blocked-matrix method, we obtained the steady-state probabilities by using the elements of the inverses of two matrices and derived the conditional stationary distribution of the queue length and waiting time.

Tian et al. [11] and Zhang and Tian [12] proved several conditional stochastic decomposition results for the queue length and customer waiting time. These results can be used to compare the M/M/*c* vacation system with its classical M/M/*c* queueing system. Due to the complexity of the formulas, at present, we have not investigated the conditional stochastic decomposition for the queue length and customer waiting time for the model in this chapter.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China (No. 70671088) and the Natural Science Foundation of Hebei Province (No. A2004000185), China, and was supported in part by GRANT-IN-AID FOR SCIENTIFIC RESEARCH (No. 19500070) and MEXT.ORC (2004-2008), Japan.

Appendix

Proof of Lemma 9.1. Let $\mathbf{X}_j = (c_{1j}, c_{2j}, \ldots, c_{Nj})^T$, $j = 1,2,\ldots,N$, be the jth column vector of the inverse matrix $\mathbf{Q}_{12}^{-1}$, and let $\varepsilon_j = (0,\ldots,1,\ldots,0)^T$ be the jth unit column vector; then we have

$$\mathbf{Q}_{12}\mathbf{X}_j = \varepsilon_j, \qquad j = 1,2,\ldots,N. \tag{9.37}$$

For $j = 1, 2, \ldots, N$, (9.37) can be rewritten as the following set of equations,

$$v_{i-1}c_{i-2j} - w_{i-1}c_{i-1j} + u_{i-1}c_{ij} = 0, \qquad i \neq j, \quad i = 1, 2, \ldots, N, \tag{9.38}$$

$$v_{i-1}c_{i-2j} - w_{i-1}c_{i-1j} + u_{i-1}c_{ij} = 1, \qquad i = j, \tag{9.39}$$

where c_{0j} and c_{-1j} are defined to be zero. Repeating the use of (9.38) gives

$$c_{ij} = 0, \qquad i = 1, 2, \ldots, j-1. \tag{9.40}$$

Substituting (9.40) into (9.39) yields

$$c_{jj} = \frac{1}{u_{j-1}}. \tag{9.41}$$

From (9.38), we have

$$c_{ij} = \frac{w_{i-1}}{u_{i-1}}c_{i-1j} - \frac{v_{i-1}}{u_{i-1}}c_{i-2j}, \qquad i = j+1, \quad j+2, \ldots, N. \tag{9.42}$$

In (9.42), we let

$$c_{ij} = k_{ij}\frac{1}{u_{j-1}}, \qquad i = j+1, \quad j+2, \ldots, N, \tag{9.43}$$

and substitute (9.43) into (9.42), so we get the recursive relations given by (9.3) for k_{ij}. This completes the proof of Lemma 9.1. □

Proof of Lemma 9.2. Let $Y_j = (d_{1j}, d_{2j}, \ldots, d_{Nj})$ be the jth column vector of the inverse matrix $\mathbf{Q}_{33}^{-1}$, then we have

$$\mathbf{Q}_{33}\mathbf{Y}_j = \boldsymbol{\varepsilon}_j, \qquad j = 1, 2, \ldots, N. \tag{9.44}$$

For $j = 1, 2, \ldots, N$, (9.44) can be rewritten as the following set of equations:

$$-s_1 d_{1j} - t_1(d_{1j} - d_{2j}) = 0, \tag{9.45}$$

$$s_i(d_{i-1j} - d_{ij}) - t_i(d_{ij} - d_{i+1j}) = 0, \qquad i = 1, 2, \ldots, N-1, \quad i \neq j, \tag{9.46}$$

$$s_j(d_{j-1j} - d_{jj}) - t_j(d_{jj} - d_{j+1j}) = 1, \tag{9.47}$$

$$s_N(d_{N-1j} - d_{Nj}) = 0. \tag{9.48}$$

Equation (9.46) can be rewritten as the following recursive relation:

$$d_{i-1j} - d_{ij} = \frac{t_i}{s_i}(d_{ij} - d_{i+1j}), \qquad i = 1, 2, \ldots, N-1, \quad i \neq j. \tag{9.49}$$

From (9.48) and (9.49), we get

$$d_{ij} = d_{jj}, \qquad i = j+1, \quad j+2, \ldots, N. \tag{9.50}$$

In (9.50), we let $i = j+1$ and then substitute it into (9.47), so we get

$$d_{j-1\,j} - d_{jj} = \frac{1}{s_j}. \tag{9.51}$$

Using (9.51) and repeating the use of the recursive relation (9.49) gives

$$d_{i-1\,j} - d_{ij} = \frac{t_i t_{i+1} \cdots t_{j-1}}{s_i s_{i+1} \cdots s_j}, \qquad i = 2, 3, \ldots, j-1. \tag{9.52}$$

In (9.52), we let $i = 2$ and then substitute it into (9.45), so we get

$$d_{1\,j} = -\frac{t_1 t_2 \cdots t_{j-1}}{s_1 s_2 \cdots s_j}. \tag{9.53}$$

Note that

$$d_{ij} = d_{1\,j} - \sum_{k=2}^{i} (d_{k-1\,j} - d_{kj}), \qquad i = 2, 3, \ldots, j-1, \tag{9.54}$$

and then substituting (9.52) and (9.53) into (9.54), we get

$$d_{ij} = -\sum_{k=1}^{i} \frac{t_k t_{k+1} \cdots t_{j-1}}{s_k s_{k+1} \cdots s_j}, \qquad i = 2, 3, \ldots, j-1. \tag{9.55}$$

In (9.55), we let $i = j-1$ and then substitute it into (9.51) and use (9.50), so we get the results of Lemma 9.2. □

References

1. J. Ke, Operating characteristic analysis on the M^X/G/1 system with a variant vacation policy and balking, *Applied Mathematical Modelling*, vol. 31, pp. 1321–1337, 2007.
2. C. Ancker, Jr. and A. Gafarian, Queueing with impatient customers who leave at random, *Journal of Industry Engineering*, vol. XIII, pp. 84–90, 1962.
3. A. Shawky, The machine interference model: M/M/C/K/N with balking, reneging and spares, *Opsearch*, vol. 37, pp. 25–35, 2000.
4. A. M. Haghighi, J. Medhi, and S. G. Mohanty, On multi-server Markovian queueing system with balking and reneging, *Computers and Operations Research*, vol. 19, pp. 421–424, 1992.
5. M. Abou-El-Ata and A. Hariri, The M/M/C/N queue with balking and reneging, *Computers and Operations Research*, vol. 19, pp. 713–716, 1992.
6. K-H. Wang and Y-C. Chang, Cost analysis of a finite M/M/R queueing system with balking, reneging and server breakdowns, *Mathematical Methods of Operations Research*, vol. 56, pp. 169–180, 2002.
7. B. Doshi, Single server queues with vacation: a survey, *Queueing System*, vol. 1, pp. 29–66, 1986.
8. B. Doshi, Single server queues with vacations, in: H. Takagi (Ed.), *Stochastic Analysis of Computer and Communication Systems*, pp. 217–265. The Netherlands: North-Holland, 1990.
9. H. Takagi, *Queueing Analysis, A Foundation of Performance Evaluation, Volume 1: Vacation and Priority Systems*. Amsterdam: Elsevier, 1991.

10. Y. Levy and U. Yechiali, An M/M/*c* queue with server's vacations, *INFOR*, vol. 14, pp. 153–163, 1976.
11. N. Tian, Q. Li, and J. Cao, Conditional stochastic decomposition in M/M/*c* queue with server vacations, *Stochastic Models*, vol. 15, pp. 367–377, 1999.
12. Z. Zhang and N. Tian, Analysis on queueing systems with synchronous vacations of partial servers, *Performance Evaluation*, vol. 52, pp. 269–282, 2003.
13. D. Yue, W. Yue, and Y. Sun, Performance analysis of an M/M/*c*/N queueing system with balking, reneging and synchronous vacations of partial servers, in *Proc. 6th International Symposium on Operations Research and Its Applications*, pp. 128–143, 2006.
14. D. Yue, Y. Zhang, and W. Yue, Optimal performance analysis of an M/M/1/N queue system with balking, reneging and server vacation, *International Journal of Pure and Applied Mathematics*, vol. 28, pp. 101–115, 2006.

Chapter 10
Analysis of Mixed Loss-Delay M/M/*m*/K Queueing Systems with State-Dependent Arrival Rates

Yoshinori Ozaki and Hideaki Takagi

Abstract An M/M/*m* queue with mixed loss and delay calls was analyzed by J. W. Cohen half a century ago (1956) where the two types of calls had identical constant arrival and service rates. It is straightforward to extend his analysis to an M/M/*m*/K queue. In this chapter, we further generalize the model such that the call arrival rates can depend on the number of calls present in the system at the arrival time. This model includes the balking and the finite population size models as special cases. We present a method of calculating the blocking probability for loss calls as well as the distribution of the waiting time for accepted delay calls. We solve a set of linear simultaneous equations for the state probabilities by numerical computation. The effects of loss calls on the mean waiting time of delayed calls are discussed based on the numerical results.

10.1 Introduction

In the traditional basic modeling of teletraffic engineering, an M/M/*m* loss system is used as a model of circuit-switched traffic leading to the Erlang-B formula for the blocking probability [1, p. 106]. An M/M/*m* delay system with an infinite waiting room is used as a model of packet-switched traffic leading to the Erlang-C formula for the waiting probability [1, p. 103]. Such models are actually used in the methodology for the spectrum requirement calculation for the International

Y. Ozaki
Graduate School of Systems and Information Engineering, University of Tsukuba, Ibaraki 305-8573, Japan
e-mail: ozaki30@sk.tsukuba.ac.jp

H. Takagi
Graduate School of Systems and Information Engineering, University of Tsukuba, Ibaraki 305-8573, Japan
e-mail: takagi@sk.tsukuba.ac.jp

W. Yue et al. (eds.), *Advances in Queueing Theory and Network Applications,*

Mobile Telecommunication-2000 (IMT-2000) in third-generation wireless communication systems [2]. Cohen [3] analyzed an M/M/m queueing system with mixed loss and delay calls with different arrival rates and identical service rates (see [9], pp. 304–305).

The mixed loss–delay system could be used as a model for the performance evaluation of a communication channel shared by circuit-switched traffic and packet-switched traffic. Cohen's analysis was recently extended to an M/M/m/K queueing system with a finite waiting room by Takagi [5], who derives explicit formulas for the blocking probability of loss calls, the blocking probability of delay calls, and the waiting time distribution of delay calls.

In this chapter, we consider a mixed loss–delay M/M/m/K queueing system in which the arrival rates of loss and delay calls can depend on the number of those calls in the system at their arrival times and the constant service rates can be different between the loss and delay calls. More specifically, when there are j loss calls and k delay calls in the system, the two types of calls arrive in an independent Poisson process with rates $\lambda_1(j,k)$ and $\lambda_2(j,k)$, respectively. Their service times are independent of each other and exponentially distributed with constant rates μ_1 and μ_2, respectively. The number of servers is denoted by m. The loss calls are lost if all servers are busy when they arrive. The delay calls wait in the waiting room unless the total number of calls present in the system exceeds K when they arrive. Namely, K is the capacity of the system including m calls in service ($m \leq K$). The assumption of state-dependent arrival rates allows us to handle a wide range of customer arrival processes. An example is the balking such that the arrival rate decreases as the number of customers present in the system increases. Another example is a queue with a finite population of customers. Figure 10.1 shows a schema of our system.

The rest of the chapter is organized as follows. In Sect. 10.2, we present a set of linear simultaneous equations for the equilibrium state probabilities. These equations are assumed to be solved numerically. In Sect. 10.3, we calculate the blocking

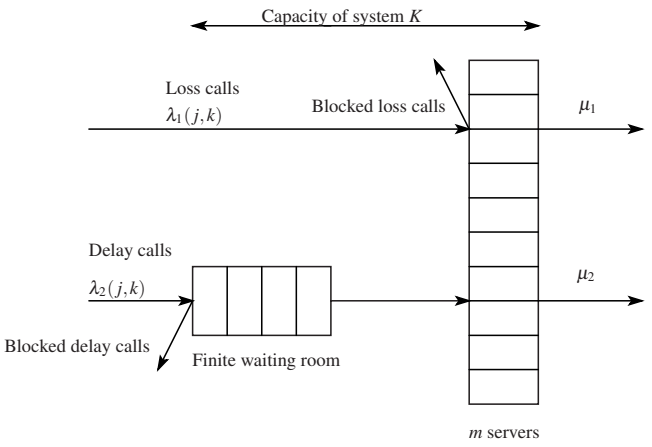


Fig. 10.1 Mixed loss–delay M/M/m/K queueing system.

probabilities for both loss and delay calls, the waiting and nonwaiting probabilities, as well as the waiting time distribution for accepted delay calls. Numerical examples are shown in Sect. 10.4. We conclude in Sect. 10.5 with a summary of present work and a plan for future study.

10.2 Equilibrium State Probability Equations

Let us denote the equilibrium state probability by

$$P_{j,k} := P\{\text{The number of the calls of loss system in the system} = j,\ \text{The number of the calls of delay system in the system} = k\},\quad 0 \le j \le m, 0 \le j+k \le K. \quad (10.1)$$

The number of states is

$$(K+1)(m+1) - \frac{m(m+1)}{2} = (m+1)\left(K+1-\frac{m}{2}\right).$$

Figure 10.2 shows the state transition rate diagram for the mixed loss–delay M/M/*m*/K system we analyze now.

Considering the number of loss and delay calls present in the system simultaneously, we can write the balance equations for the equilibrium state probabilities as follows:

First, we consider the empty state $(0,0)$. The system goes out of this state when a call arrives, and comes into this state when the service finishes at state $(1,0)$ and $(0,1)$. Thus we have

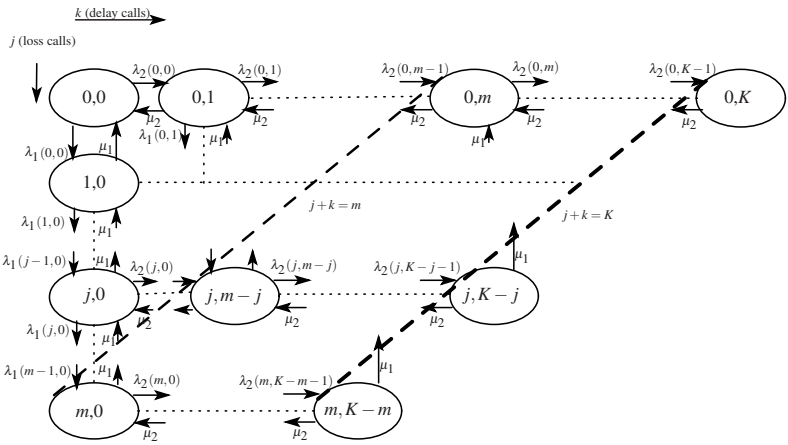


Fig. 10.2 State transition rate diagram for the mixed loss–delay M/M/*m*/K system.

$$[\lambda_1(0,0)+\lambda_2(0,0)]P_{0,0}=\mu_1 P_{1,0}+\mu_2 P_{0,1}. \tag{10.2}$$

Second, we consider the state (j,k) such that $0\le j\le m-1$, $1\le j+k\le m-1$ in which there are calls being served and some free servers, and find

$$[\lambda_1(0,k)+\lambda_2(0,k)+k\mu_2]P_{0,k}=\lambda_2(0,k-1)P_{0,k-1}+\mu_1 P_{1,k}+(k+1)\mu_2 P_{j,1},\quad 1\le k\le m-1, \tag{10.3}$$

$$[\lambda_1(j,k)+\lambda_2(j,k)+j\mu_1+k\mu_2]P_{j,k}=\lambda_1(j-1,k)P_{j-1,k}+\lambda_2(j,k-1)P_{j,k-1}+(j+1)\mu_1 P_{j+1,k}+(k+1)\mu_2 P_{j,k+1},\quad 1\le j\le m-1,\ 1\le k\le m-1,\ 2\le j+k\le m-1, \tag{10.4}$$

$$[\lambda_1(j,0)+\lambda_2(j,0)+j\mu_1]P_{j,0}=\lambda_1(j-1,0)P_{j-1,0}+(j+1)\mu_1 P_{j+1,0}+\mu_2 P_{j,1},\quad 1\le j\le m-1. \tag{10.5}$$

Third, we consider the state (j,k) such that $0\le j\le m$, $j+k=m$ in which all servers are busy and all waiting positions are available for delay calls, and find

$$[\lambda_2(0,m)+m\mu_2]P_{0,m}=\lambda_2(0,m-1)P_{0,m-1}+\mu_1 P_{1,m}+m\mu_2 P_{0,m+1}, \tag{10.6}$$

$$[\lambda_2(j,k)+j\mu_1+k\mu_2]P_{j,k}=\lambda_1(j-1,k)P_{j-1,k}+\lambda_2(j,k-1)P_{j,k-1}+(j+1)\mu_1 P_{j+1,k}+(m-j)\mu_2 P_{j,k+1},\quad 1\le j\le m-1,\ 1\le k\le m-1,\ j+k=m, \tag{10.7}$$

$$[\lambda_2(m,0)+m\mu_1]P_{m,0}=\lambda_1(m-1,0)P_{m-1,0}. \tag{10.8}$$

Fourth, we consider the state (j,k) such that $0\le j\le m$, $m+1\le j+k\le K-1$ in which all servers are busy and there is at least one waiting position available for a delay call, and find

$$[\lambda_2(0,k)+m\mu_2]P_{0,k}=\lambda_2(0,k-1)P_{0,k-1}+\mu_1 P_{1,k}+m\mu_2 P_{0,k+1},\quad m+1\le k\le K-1, \tag{10.9}$$

$$[\lambda_2(j,k)+j\mu_1+(m-j)\mu_2]P_{j,k}=\lambda_2(j,k-1)P_{j,k-1}+(j+1)\mu_1 P_{j+1,k}+(m-j)\mu_2 P_{j,k+1},\quad 1\le j\le m-1,\ m+1\le j+k\le K-1, \tag{10.10}$$

$$[\lambda_2(m,k)+m\mu_1]P_{m,k}=\lambda_2(m,k-1)P_{m,k-1},\qquad 1\le k\le K-m-1. \tag{10.11}$$

Finally, we consider the state (j,k) such that $0 \le j \le m,\ j+k=K$ in which all servers are busy and all waiting positions are occupied, and find

$$m\mu_2 P_{0,K} = \lambda_2(0,K-1)P_{0,K-1}, \tag{10.12}$$

$$[j\mu_1 + (m-j)\mu_2]P_{j,k} = \lambda_2(j,k-1)P_{j,k-1}, \quad 1 \le j \le m-1,\ j+k=K, \tag{10.13}$$

$$m\mu_1 P_{m,K-m} = \lambda_2(m,K-m-1)P_{m,K-m-1}. \tag{10.14}$$

The total number of equations is given by

$$1+(m-1)+\frac{(m-1)(m-2)}{2}+(m-1)+1+(m-1)+1+(K-m-1)$$
$$+\,(K-m-1)(m-1)+(K-m-1)+1+(m-1)+1=(m+1)\left(K+1-\frac{m}{2}\right),$$

which equals the number of all states. One of the equations is redundant. The normalization condition is given by

$$\sum_{j=0}^{m}\sum_{k=0}^{K-j} P_{j,k} = 1. \tag{10.15}$$

Hence we have a set of linear simultaneous equations with respect to the unknowns $\{P_{j,k};\ 0 \le j \le m,\ 0 \le j+k \le K\}$. It is assumed that they are solved numerically.

10.3 Analysis of Blocking Probability and Waiting Time

We are now in a position to calculate the blocking probability of loss calls, the blocking probability of delay calls, the waiting and nonwaiting probabilities of accepted delay calls, and the waiting time distribution of accepted delay calls.

10.3.1 Blocking Probability of Loss Calls

Loss calls are blocked if all servers are busy upon their arrival. If the population of loss calls is infinite, the blocked loss calls are simply lost for good. If the population of loss calls is finite, the blocked loss calls are assumed to return to their source without being served.

Let us consider a long time τ. The mean number of loss calls that arrive in τ is given by product of the arrival rate $\lambda_1(j,k)$ of loss calls and the time interval $P_{j,k}\tau$ in which the system is in state (j,k) during τ summed over all possible states as follows:

$$\sum_{j=0}^{m}\sum_{k=0}^{K-j}\lambda_1(j,k)P_{j,k}\tau. \tag{10.16}$$

The mean number of loss calls blocked during τ is given by the product of the arrival rate $\lambda_1(j,k)$ of loss calls and the time interval $P_{j,k}\tau$ summed over all states in which all servers are busy:

$$\sum_{j=0}^{m}\sum_{k=m-j}^{K-j}\lambda_1(j,k)P_{j,k}\tau. \tag{10.17}$$

Thus the blocking probability P_B of loss calls is given by the ratio of the above two equations:

$$P_B=\frac{\displaystyle\sum_{j=0}^{m}\sum_{k=m-j}^{K-j}\lambda_1(j,k)P_{j,k}}{\displaystyle\sum_{j=0}^{m}\sum_{k=0}^{K-j}\lambda_1(j,k)P_{j,k}}. \tag{10.18}$$

10.3.2 Blocking Probability of Delay Calls

Delay calls are blocked if all servers are busy and all waiting positions are occupied upon their arrival. If the population of delay calls is infinite, the blocked delay calls are simply lost. If the population of delay calls is finite, the blocked delay calls are assumed to return to their source without being served. The mean number of arrivals of delay calls during time τ is given by

$$\sum_{j=0}^{m}\sum_{k=0}^{K-j}\lambda_2(j,k)P_{j,k}\tau. \tag{10.19}$$

The mean number of delay calls blocked during τ is given by the product of the arrival rate $\lambda_2(j,k)$ of delay calls and the time interval $P_{j,K-j}\tau$ summed over all states $0\le j\le m$ in which all servers are busy and all waiting positions are occupied:

$$\sum_{j=0}^{m}\lambda_2(j,K-j)P_{j,K-j}\tau. \tag{10.20}$$

Thus the blocking probability P'_B of delay calls is given by the ratio of the two:

$$P'_B=\frac{\displaystyle\sum_{j=0}^{m}\lambda_2(j,K-j)P_{j,K-j}}{\displaystyle\sum_{j=0}^{m}\sum_{k=0}^{K-j}\lambda_2(j,k)P_{j,k}}. \tag{10.21}$$

10.3.3 Waiting and Nonwaiting Probabilities of Accepted Delay Calls

We now consider the delay calls that are accepted upon arrival. The mean number of delay calls accepted during τ is given by the product of the arrival rate $\lambda_2(j,k)$ of delay calls and the time interval $P_{j,k}\tau$ summed over all states $0 \le j \le m,\ 0 \le j+k \le K-1$ in which there is at least one waiting position available:

$$\sum_{j=0}^{m} \sum_{k=0}^{K-j-1} \lambda_2(j,k) P_{j,k} \tau. \tag{10.22}$$

Therefore the probability that there are j loss calls and k delay calls present in the system immediately before the arrival of an arbitrary delay call that is to be accepted is given by

$$\hat{P}_{j,k} = \frac{\lambda_2(j,k) P_{j,k}}{\sum_{j=0}^{m} \sum_{k=0}^{K-j-1} \lambda_2(j,k) P_{j,k}}, \qquad 0 \le j \le m, \qquad 0 \le j+k \le K-1. \tag{10.23}$$

Let us denote by W the waiting time of an accepted delay call. The probability that accepted delay calls do not wait is given by the probability that there is at least one server available upon their arrival:

$$P\{W=0\} = \sum_{j=0}^{m-1} \sum_{k=0}^{m-j-1} \hat{P}_{j,k} = \frac{\sum_{j=0}^{m-1} \sum_{k=0}^{m-j-1} \lambda_2(j,k) P_{j,k}}{\sum_{j=0}^{m} \sum_{k=0}^{K-j-1} \lambda_2(j,k) P_{j,k}}. \tag{10.24}$$

The probability that accepted delay calls wait is given by the probability that all the servers are busy but that there is at least one waiting position available upon their arrival:

$$P\{W>0\} = \sum_{j=0}^{m} \sum_{k=m-j}^{K-j-1} \hat{P}_{j,k} = \frac{\sum_{j=0}^{m} \sum_{k=m-j}^{K-j-1} \lambda_2(j,k) P_{j,k}}{\sum_{j=0}^{m} \sum_{k=0}^{K-j-1} \lambda_2(j,k) P_{j,k}}. \tag{10.25}$$

10.3.4 Waiting Time Distribution of Accepted Delay Calls

Let us denote by $R^*_{j,k,j+k-m}(s)$ the Laplace–Stieltjes transform (LST) of the distribution function (DF) of the waiting time of a delay call that arrives when there are j

loss calls and k delay calls in the system, where $j+k \geq m$. This is the time until the total number of calls in the system decreases to $m-1$, at which point the service to that call is started. Then the LST of the DF for the waiting time W (>0) of accepted delay calls that are to wait is given by

$$F_W^*(s \mid W>0) = \frac{\displaystyle\sum_{j=0}^{m}\sum_{k=m-j}^{K-j-1} \lambda_2(j,k) P_{j,k} R^*_{j,k,j+k-m}(s)}{\displaystyle\sum_{j=0}^{m}\sum_{k=m-j}^{K-j-1} \lambda_2(j,k) P_{j,k}}. \tag{10.26}$$

The LST of the DF for the waiting time W (≥ 0) of all accepted delay calls is given by

$$\begin{aligned} F_W^*(s) &= P\{W=0\} + F_W^*(s \mid W>0) P\{W>0\} \\ &= \frac{\displaystyle\sum_{j=0}^{m}\left[\sum_{k=0}^{m-j-1} \lambda_2(j,k) P_{j,k} + \sum_{k=m-j}^{K-j-1} \lambda_2(j,k) P_{j,k} R^*_{j,k,j+k-m}(s)\right]}{\displaystyle\sum_{j=0}^{m}\sum_{k=0}^{K-j-1} \lambda_2(j,k) P_{j,k}}. \end{aligned} \tag{10.27}$$

We can obtain $R^*_{j,k,j+k-m}(s)$ $(m \leq j+k \leq K-1)$ as follows. Note that the third subscript of $R^*_{j,k,j+k-m}(s)$ denotes the number of calls present in the waiting room. We start with

$$R^*_{j,k,0}(s) = r_{j,k}(s) + \hat{r}_{j,k}(s) \tag{10.28}$$

for $j+k=m$, where

$$r_{j,k}(s) = \frac{j\mu_1}{j\mu_1+k\mu_2} \times \frac{j\mu_1+k\mu_2}{s+j\mu_1+k\mu_2} = \frac{j\mu_1}{s+j\mu_1+k\mu_2} \tag{10.29}$$

is the LST of the DF for the transition time from state (j,k) to state $(j-1,k)$, and

$$\hat{r}_{j,k}(s) = \frac{k\mu_2}{j\mu_1+k\mu_2} \times \frac{j\mu_1+k\mu_2}{s+j\mu_1+k\mu_2} = \frac{k\mu_2}{s+j\mu_1+k\mu_2} \tag{10.30}$$

is the LST of the DF for the transition time from state (j,k) to state $(j,k-1)$. For $j+k=m+l$, we have

$$R^*_{j,k,l}(s) = r_{j,k}(s) R^*_{j-1,k,l-1}(s) + \hat{r}_{j,k}(s) R^*_{j,k-1,l-1}(s). \tag{10.31}$$

Therefore, we can calculate $R^*_{j,k,l}(s)$ recursively for $l=1,2,\ldots,K-m-1$ by starting with $R^*_{j,k,0}(s)$ given in (10.28).

10.4 Numerical Examples

Using the method of analysis given in Sect. 10.3, we present numerical examples of the blocking probabilities for loss and delay calls and the mean waiting time for accepted delay calls. The latter can be obtained from the LST of the DF for the waiting time given in (10.27). We consider the cases of fixed arrival rates, balking of delay calls, and the finite population size.

10.4.1 Equilibrium State Probabilities

Let us first confirm that our generalization in the above yields the same results as the analysis in [5] for the M/M/*m*/K queue with constant arrival rates and identical service rates. To do so numerically, we consider the mixed loss–delay M/M/3/5 queue with $\lambda_1 = 2$, $\lambda_2 = 3$, and $\mu_1 = \mu_2 = 3$. Table 10.1 shows the equilibrium state probabilities we have computed with the above method. We have confirmed that these values are identical with those calculated by using the formulas in [5].

10.4.2 Blocking Probabilities of Loss and Delay Calls

We now consider the M/M/*m*/K queues with constant arrival rates in the case in which the service rates are different for loss and delay calls. Figure 10.3 shows the blocking probabilities of loss and delay calls in the M/M/4/7 queue with $\mu_1 = 2$, $\mu_2 = 1$, $\lambda_1 = 0.005$ for $0 \leq \lambda_2 \leq 20$. As the arrival rate of delay calls increases, both blocking probabilities increase. The blocking probability of loss calls increases faster than that of delay calls.

Table 10.1 Equilibrium state probabilties in the mixed loss–delay M/M/3/5 queue with $\lambda_1 = 2$, $\lambda_2 = 3$, and $\mu_1 = \mu_2 = 3$.

	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
$j=0$	0.535698	0.201639	0.038431	0.005252	0.000706	0.000088
$j=1$	0.133172	0.049915	0.009509	0.001196	0.000150	
$j=2$	0.016282	0.005830	0.000688	0.000086		
$j=3$	0.001206	0.000134	0.000017			

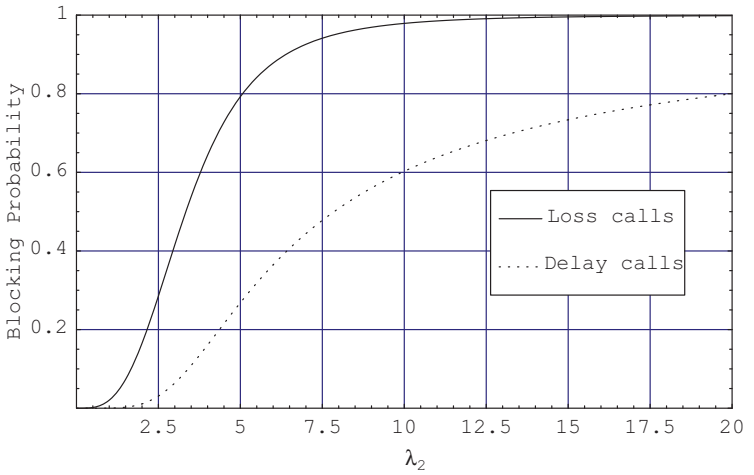


Fig. 10.3 Blocking probabilities of loss and delay calls in the M/M/4/7 queue with fixed arrival and service rates ($\mu_1 = 2, \mu_2 = 1$, $\lambda_1 = 0.005$, and $0 \leq \lambda_2 \leq 20$).

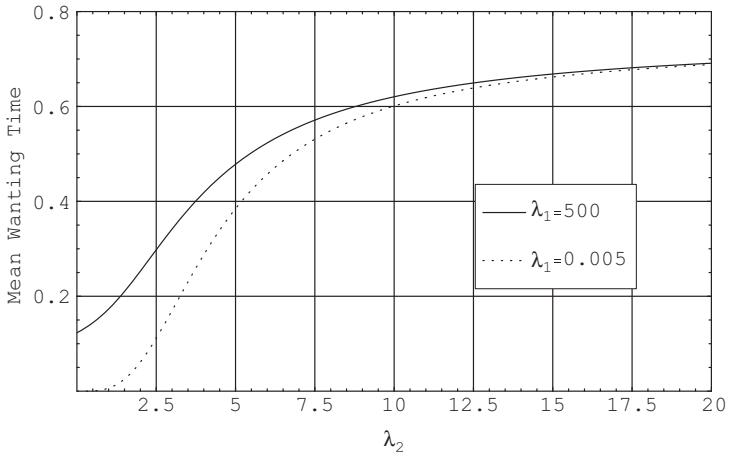


Fig. 10.4 Mean waiting time of delay calls in the M/M/4/7 queue with fixed arrival and service rates ($\mu_1 = 2,\ \mu_2 = 1,\ \lambda_1 = \{500, 0.005\}$, and $0 \leq \lambda_2 \leq 20$).

10.4.3 Mean Waiting Time

We evaluate the mean waiting times of accepted delay calls for several cases of state-dependent arrival rates in the M/M/4/7 queue.

1. *Fixed Arrival Rates*

Figure 10.4 shows numerical examples of the mean waiting time of delay calls when $\mu_1 = 2, \mu_2 = 1$, $\lambda_1 = \{500, 0.005\}$ for $0 \leq \lambda_2 \leq 20$. The mean waiting time of delay

calls increases as their arrival rate λ_2 increases. When λ_2 is small, the mean waiting time increases quickly. When λ_2 is large, the mean waiting time increases slowly. We can also observe the effects of sharing the servers with loss calls on the mean waiting time.

2. *Balking*

Balking in the arrival process means that the arrival rate of calls decreases as the number of calls present in the system increases. We consider three models of balking for delay calls in which their arrival rates $\lambda_2(j,k)$ for $j+k>m$ are given as follows:

$$\text{Model 1}: \lambda_2(j,k) = \nu_2 \left(\frac{K-j-k}{K-m}\right)^{\alpha}, \qquad 0 \le \nu_2 \le 20,$$

$$\text{Model 2}: \lambda_2(j,k) = \frac{\nu_2}{(j+k-m+1)^{\alpha}}, \qquad 0 \le \nu_2 \le 20,$$

$$\text{Model 3}: \lambda_2(j,k) = \nu_2 e^{-\alpha(j+k-m)}, \qquad 0 \le \nu_2 \le 20,$$

where $\alpha > 0$. It is assumed that $\lambda_2(j,k) = \nu_2$ for $0 \le j+k \le m$ in the three models. Model 1 is the case in which the arrival rate of delay calls decreases in power law with the occupancy ratio of waiting positions. Model 2 is the case in which the arrival rate of delay calls decreases in power law with the number of occupied waiting positions. Model 3 is the case in which the arrival rate of delay calls decreases exponentially with the number of occupied waiting positions. See Fig. 10.5 for dependence of $\lambda_2(j,k)$ on the total number of calls, $j+k$, present in the system.

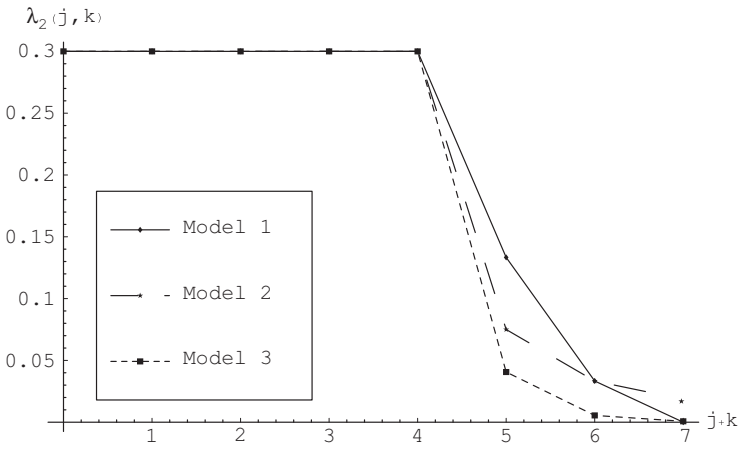


Fig. 10.5 Three models of the arrival rate of delay calls with balking ($m=4,\ K=7,\ \alpha=2,\ \nu_2=0.3$).

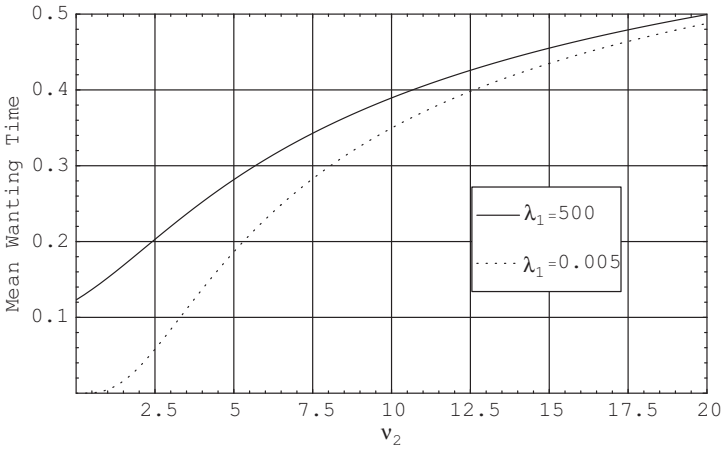


Fig. 10.6 Mean waiting time of delay calls in the M/M/4/7 queue with balking of model 1 ($\mu_1 = 2,\ \mu_2 = 1,\ \lambda_1 = \{500,\ 0.005\},\ \alpha = 2$, and $0 \leq \nu_2 \leq 20$).

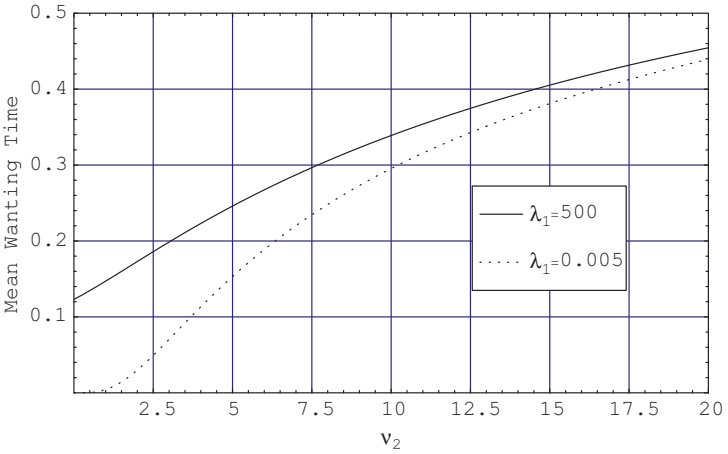


Fig. 10.7 Mean waiting time of delay calls in the M/M/4/7 queue with balking model 2 ($\mu_1 = 2,\ \mu_2 = 1,\ \lambda_1 = \{500,\ 0.005\},\ \alpha = 2$, and $0 \leq \nu_2 \leq 20$).

In Figs. 10.6–10.8, we plot the mean waiting time of delay calls with balking for models 1–3, respectively, by assuming $\mu_1 = 2,\ \mu_2 = 1,\ \alpha = 2,\ \lambda_1 = \{500, 0.005\}$ for $0 \leq \nu_2 \leq 20$.

3. *Finite Population Size*

M/M/*m*/K queues with finite population of loss and delay calls can be handled with our model of state-dependent arrival rates by assuming that

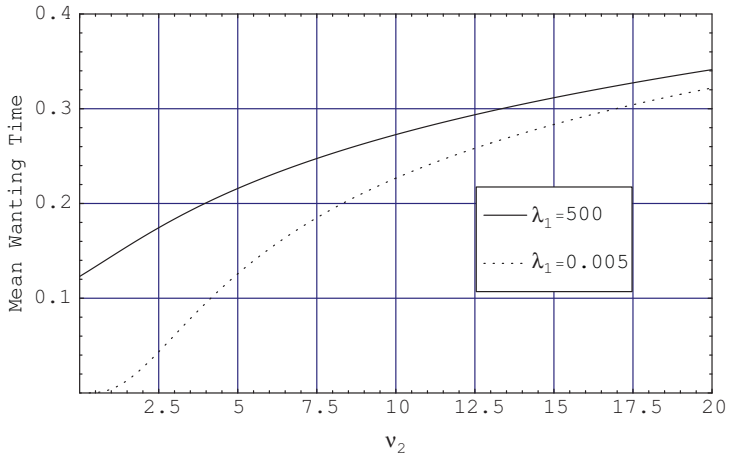


Fig. 10.8 Mean waiting time of delay calls in the M/M/4/7 queue with balking model 3 ($\mu_1 = 2,\ \mu_2 = 1,\ \lambda_1 = \{500,\ 0.005\},\ \alpha = 2$, and $0 \le \nu_2 \le 20$).

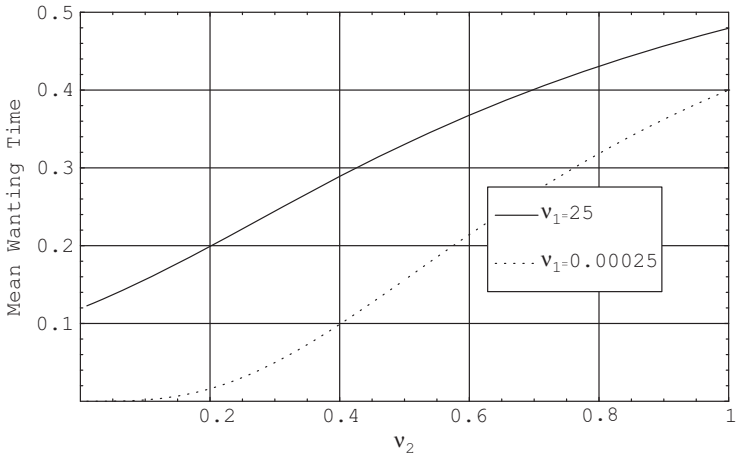


Fig. 10.9 Mean waiting time of delay calls in the M/M/4/7 queue with finite population ($\mu_1 = 2,\ \mu_2 = 1,\ n_1 = n_2 = 20,\ \alpha = 2,\ \nu_1 = \{25,\ 0.00025\}$, and $0 \le \nu_2 \le 1$).

$$\lambda_1(j,k) = (n_1 - j)\nu_1, \qquad 0 \le j \le n_1,$$

$$\lambda_2(j,k) = (n_2 - k)\nu_2, \qquad 0 \le k \le n_2,$$

where n_1 and n_2 are the fixed total numbers of loss and delay calls, respectively. The call arrivals then form pseudo-Poisson processes.

In Fig. 10.9, we show the mean waiting time of delay calls in the finite population model with $\mu_1 = 2,\ \mu_2 = 1,\ n_1 = n_2 = 20,\ \alpha = 2,\ \nu_1 = \{25,\ 0.00025\}$ for $0 \le \nu_2 \le 1$.

10.5 Concluding Remarks

In this chapter, we have shown the analysis of a mixed loss–delay M/M/*m*/K queueing system with state-dependent arrival rates and different constant service rates. We have first presented a set of linear simultaneous equations for the equilibrium state probabilities and the normalization condition. We have then evaluated the blocking probabilities for loss and delay calls and the mean waiting time for accepted delay calls.

For numerical examples, we have considered the cases of fixed arrival rates, balking of delay calls, and finite population size in the M/M/4/7 queueing system. In these examples, we have observed how the mean waiting time of accepted delay calls increases as their arrival rate increases when they share the servers with loss calls.

It is our future work to extend the model to allow multiple classes of both loss and delay calls with some scheduling discipline among them. Such a model would be closer to the channel sharing by circuit- and packet-switched traffic in the next-generation wireless communication systems.

References

1. L. Kleinrock, *Qeueing Systems, Volume 1: Theory*. New York: John Wiley & Sons, 1975.
2. ITU-R, Methodology for the calculation of IMT-2000 terrestrial spectrum requirements, Recommendation ITU-R M.1390, 1999.
3. J. W. Cohen, Certain delay problems for a full availability trunk group loaded by two traffic sources, *Communication News*, vol. 16, no. 3, pp. 105–113, 1956.
4. T. L. Saaty, *Elements of Queueing Theory with Applications*. New York: McGraw-Hill, 1981. Republished by New York: Dover Publications, 1983.
5. H. Takagi, Explicit delay distribution in First-Come First-Served M/M/*m*/K and M/M/*m*/K/*n* queues and a mixed loss-delay system, in *Proc. Asia-Pacific Symposium on Queuing Theory and Its Application to Telecommunication Networks*, pp. 1–11, 2006. *International Journal of Pure and Applied Mathematics*, vol. 40, no. 2, pp. 185–200, 2007.

Chapter 11
Asymptotic Behavior of Loss Rate for Feedback Finite Fluid Queue with Downward Jumps

Yutaka Sakuma and Masakiyo Miyazawa

Abstract We consider a feedback fluid queue with a finite buffer and downward jumps, where the net flow rate and the jump size for the buffer content are controlled by a background Markov chain with a finite state space. The feedback means that the transition structure of the background Markov chain may change when the buffer content becomes empty or full. In this chapter, we show that the loss rate for this fluid queue decays exponentially as the buffer size gets large under a negative drift condition.

11.1 Introduction

We are concerned with a fluid queue, which consists of a server and a buffer. When the input rate of the fluid flow exceeds the processing capacity of the server, the unprocessed fluid is stored in the buffer. The input and output rates of the fluid flow depend on the state of a background process. Assume that the background process is a continuous time Markov chain on a finite state space. Fluid queues have been applied in many situations, for example, real-world systems such as petroleum and chemical industries, performance analysis of high-speed data networks, designing computer systems, and so on. Anick, Mitra, and Sondhi [1] study the data-handling switch in a computer network by using a fluid queue with an infinite buffer. Bonald [2] and van Foreest, Mandjes, and Scheinhardt [3] study the performance evaluation of TCP/IP (see e.g., [4]) by using fluid queues. Specifically, [3] studies the feature of *Additive Increase/Multiplicative Decrease* in TCP/IP

Y. Sakuma
Department of Information Sciences, Tokyo University of Sciences, Chiba 278-8510, Japan
e-mail: sakuma-y@rs.noda.tus.ac.jp

M. Miyazawa
Department of Information Sciences, Tokyo University of Sciences, Chiba 278-8510, Japan
e-mail: miyazawa@rs.noda.tus.ac.jp

W. Yue et al. (eds.), *Advances in Queueing Theory and Network Applications,*

by using a feedback fluid queue with a finite buffer. The feedback means that the transition structure of the background Markov chain may change when the buffer content becomes empty or full. da Silva Soares and Latouche [5] show that the stationary density of the buffer content for the feedback fluid queue with a finite buffer is expressed as a linear combination of two exponential matrices, by using the matrix analytic method. We call this fluid queue a feedback finite fluid queue (FFFQ, for short).

In this chapter, we extend the FFFQ in such a way that an accumulated net fluid flow may have downward jumps (i.e., instantaneous draining) when the background state changes. The downward jump is motivated by the following observations. Consider a bottleneck router connected to TCP sources in the Internet. IP packets arriving from the TCP sources enter the buffer of the router, and wait to be served. The router sends IP packets to output links according to a routing table, which may be updated in a timely manner. Assume that IP packets and the output links have various sizes and capacities, respectively. When IP packets of small sizes are transferred to the output link with low capacity, the buffer content of the router slowly decreases. On the other hand, when IP packets of large sizes are transferred to the output link with high capacity, the buffer content rapidly decreases, which may be regarded as downward jumps.

We aim to consider an asymptotic behavior of the loss rate $\ell_{\mathrm{Loss}}^{(b)}$ for the FFFQ with downward jumps as the buffer size b goes to infinity. Under a negative drift condition, we show that there exist positive constants c and α such that

$$\lim_{b\to\infty} e^{\alpha b}\ell_{\mathrm{Loss}}^{(b)} = c$$

and we obtain α as the solution of a certain equation (see Theorem 11.1 in Sect. 11.5 of this chapter). Note that Asmussen and Pihlsgård [6] study an asymptotic behavior of a Levy process with two reflecting boundaries, and generalize it to a Markov-modulated Levy process. They show that the loss rate decays exponentially as the one of the boundaries goes to infinity. However, their model does not have the feedback mechanism.

In this chapter, we heavily use the results in [7] and [8]. In [7], a Markov-modulated fluid queue with an infinite buffer and downward jumps is studied. And the stationary distribution of the buffer content is given by a matrix-exponential form, which is one of the key observations for our study (see Theorem 3.1 of [8]). In [8], a Markov-modulated fluid queue upward jumps is studied. For this model, the hitting probability for an upper level does not have the matrix-exponential form, because this process is not skip-free in the upward direction. The hitting probability is also the key observation for our study because of the two-sided reflections of our model. Ramaswami [9] studies a fluid flow model by using a quasi birth-and-death (QBD, for short) process, and as mentioned in [5] and [10], the FFFQ has a close connection with a finite-level QBD process. In this sense, if there is no jump, our results are related to those in [11], where a many-server queue with a finite buffer is modeled by a finite-level QBD process. The loss probability for this queueing model decays geometrically as the buffer size goes to infinity under a negative drift condition.

This chapter is composed of six sections. In Sect. 11.2, we introduce a Markov additive process with downward jumps. In Sect. 11.3, we put a reflecting boundary at level 0 for this additive process. Then we get a feedback fluid queue with an infinite buffer and downward jumps. In Sect. 11.4, we further put a reflecting boundary at level b, and get the FFFQ with downward jumps. In Sect. 11.5, we give the asymptotic behavior of the loss rate for the FFFQ with downward jumps. In Sect. 11.6, we provide some numerical results. Finally, conclusions are drawn in Sect. 11.7.

11.2 MAP (Markov Additive Process) with Downward Jumps

When the two boundaries of the FFFQ with downward jumps are removed, we get a Markov additive process (MAP, for short) with downward jumps. So we first consider the MAP with downward jumps and its hitting probability for an upper level in this section. The additive process and its hitting probabilities play key roles in the subsequent sections. Before proceeding, we first introduce some notations for matrices and vectors, which are used throughout the chapter. Denote an identity matrix, a unit vector, and a zero vector by I, $\mathbf{1}$, and $\mathbf{0}$, respectively, where their sizes can be identified in the contexts where they appear. For vector $\mathbf{a}$, let $\Delta_{\mathbf{a}}$ be the diagonal matrix whose (i,i)th element is the ith element of the vector $\mathbf{a}$. Denote the (i,j)th element of matrix A by $[A]_{ij}$, and the ith element of vector $\mathbf{a}$ by $[\mathbf{a}]_i$ unless stated otherwise. Let A^{T} be the transposition of matrix A.

Let $M(t)$ be a continuous-time Markov chain (CTMC, for short) with a finite state space $\mathscr{S}$. The transition rate matrix of $M(t)$ is decomposed into two $\mathscr{S}\times\mathscr{S}$ matrices C and D, where C is an ML-matrix and D is a nonnegative matrix such that $(C+D)\mathbf{1}=\mathbf{0}$. Throughout the chapter, assume that $C+D$ is irreducible. Then we have a stationary distribution $\boldsymbol{\pi}$ for $C+D$; that is, $\boldsymbol{\pi}(C+D)=\mathbf{0}$ and $\boldsymbol{\pi}\mathbf{1}=1$. Let $\mathbf{r}=(r(i);i\in\mathscr{S})$, where $r(\cdot)$ is a real-valued function defined on $\mathscr{S}$. Define an additive process $X(t)$ driven by $M(t)$ as follows:

(i) When $M(t)=i(\in\mathscr{S})$, $X(t)$ changes at rate $r(i)$; that is, $(d/dt)X(t)=r(M(t))$.
(ii) When $M(t)$ changes from i to j by $[C]_{ij}$, the changing rate of $X(t)$ changes from $r(i)$ to $r(j)$.
(iii) When $M(t)$ changes from i to j by $[D]_{ij}$, the changing rate of $X(t)$ changes from $r(i)$ to $r(j)$, and $X(t)$ jumps down with a jump size subject to a distribution F_{ij}.

The two-dimensional CTMC $(X(t),M(t))$ with a state space $(-\infty,\infty)\times\mathscr{S}$ is called a MAP with downward jumps, or simply called MAP (see [7]). We call the first component the level process or sometimes the additive component, and call the second component the background process.

For simplicity, assume that r takes nonzero values. Divide the state space $\mathscr{S}$ into two disjoint subsets $\mathscr{S}^-$ and $\mathscr{S}^+$, where $\mathscr{S}^-=\{i\in\mathscr{S}|r(i)<0\}$ and $\mathscr{S}^+=\{i\in\mathscr{S}|r(i)>0\}$. To avoid the trivial case, assume that neither $\mathscr{S}^-$ nor $\mathscr{S}^+$ is a null set. Then we partition $\boldsymbol{\pi}$, $\mathbf{r}$, C, and D according to $\mathscr{S}^-$ and $\mathscr{S}^+$ as

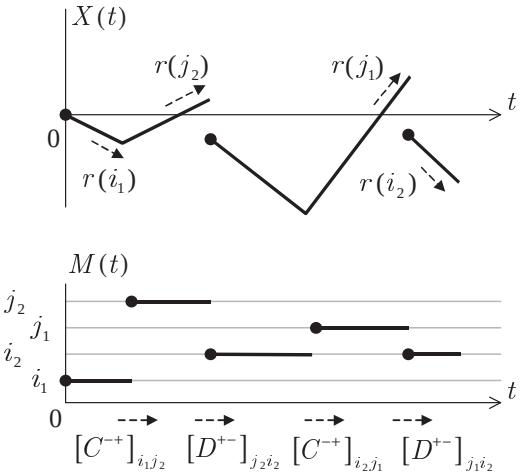


Fig. 11.1 MAP with downward jumps, where $\mathscr{S}^- = \{i_1, i_2\}$, $\mathscr{S}^+ = \{j_1, j_2\}$.

$$\boldsymbol{\pi} = (\boldsymbol{\pi}^-, \boldsymbol{\pi}^+), \qquad \mathbf{r} = (\mathbf{r}^-, \mathbf{r}^+), \qquad \begin{pmatrix} C^{--} & C^{-+} \\ C^{+-} & C^{++} \end{pmatrix}, \qquad \begin{pmatrix} D^{--} & D^{-+} \\ D^{+-} & D^{++} \end{pmatrix}$$

(see Fig. 11.1).

For $x \geq 0$, let τ_x^+ be a first hitting time when the level process hits x; that is, $\tau_x^+ = \inf\{t > 0; X(t) \geq x\}$. For $x \geq 0$, define the $\mathscr{S} \times \mathscr{S}^+$ matrix $R^{\bullet +}(x)$ whose (i, j)th element is given by

$$[R^{\bullet +}(x)]_{ij} = \mathrm{P}(M(\tau_x^+) = j | X(0) = 0, M(0) = i),$$

which is a first hitting probability for an upper-level x with a background state j, starting from level 0 with a background state i. Divide $R^{\bullet +}(x)$ into blocks $R^{-+}(x)$ and $R^{++}(x)$ according to $\mathscr{S}^-$ and $\mathscr{S}^+$. Throughout the chapter, assume the following condition,

$$\mathrm{E}[X(1) - X(0)] < 0, \tag{11.1}$$

which is referred to as a negative drift condition. For $x \geq 0$, define the $\mathscr{S} \times \mathscr{S}$ matrix $D(x) = ([D]_{ij} F_{ij}(x); i, j \in \mathscr{S})$. Then $R^{\bullet +}(x)$ has the matrix exponential form.

Proposition 11.1. (Theorem 3.1 of [8]) Under the negative drift condition (11.1), there exist the $\mathscr{S}^+ \times \mathscr{S}^+$ defective transition rate matrix U and the $\mathscr{S}^- \times \mathscr{S}^+$ substochastic matrix R_0 satisfying

$$\begin{pmatrix} R_0 \\ I \end{pmatrix} U = \Delta_{\mathbf{r}}^{-1} \left\{ C \begin{pmatrix} R_0 \\ I \end{pmatrix} + \int_0^\infty D(du) \begin{pmatrix} R_0 \\ I \end{pmatrix} \exp(uU) \right\}. \tag{11.2}$$

And we have

$$\begin{pmatrix} R^{-+}(x) \\ R^{++}(x) \end{pmatrix} = \begin{pmatrix} R_0 \\ I \end{pmatrix} \exp(xU), \qquad x \geq 0. \tag{11.3}$$

Remark 11.1. R_0 and U are computed by the following recursion formula:

$$\begin{aligned}
&U_{[0]} = C^{++}, \qquad R_{[0]} = 0, \\
&U_{[n+1]} = \Delta_{\mathbf{r}^+}^{-1} \left(C^{++} + C^{+-} R_{[n]} + \int_0^\infty (D^{+-}(du) R_{[n]} + D^{++}(du)) \exp(uU_{[n]}) \right), \\
&R_{[n+1]} = -\Delta_{\mathbf{r}^-}^{-1} \Bigg(C^{-+} + (-\eta \Delta_{\mathbf{r}^-} + C^{--}) R_{[n]} \\
&\qquad + \int_0^\infty (D^{--}(du) R_{[n]} + D^{-+}(du)) \, \exp(uU_{[n]}) \Bigg) (\eta I - U_{[n]})^{-1}, \qquad x \geq 0,
\end{aligned}$$

where $\eta (>0)$ is chosen so that $-\eta \Delta_{\mathbf{r}^-} + C^{--}$ is a nonnegative matrix (see [8] and [12]).

We next introduce the two hitting times,

$$\tau_0^- = \inf\{t > 0 | X(t) \leq 0\}, \qquad \tau_b^+ = \inf\{t > 0 | X(t) \geq b\},$$

where $b > 0$. Let ${}_0A_{0b}^{++}$ be the $\mathscr{S}^+ \times \mathscr{S}^+$ matrix whose (i, j)th element is given by

$$[{}_0A_{0b}^{++}]_{ij} = \mathrm{P}(M(\tau_b^+) = j, \tau_b^+ < \tau_0^- | X(0) = 0, M(0) = i).$$

This is the hitting probability that the level process hits level b with a background state $j \in \mathscr{S}^+$ before it goes below level 0, starting from level 0 with a background state $i \in \mathscr{S}^+$. Let P_{00}^{++} be the $\mathscr{S}^+ \times \mathscr{S}^+$ matrix whose (i, j)th element is the probability that $X(t)$ returns to level 0 with a background state $j \in \mathscr{S}^+$, starting from level 0 with a background state $i \in \mathscr{S}^+$. That is, the (i, j)th element of P_{00}^{++} is given by

$$[P_{00}^{++}]_{ij} = \mathrm{P}(M(\zeta_0^+) = j | X(0) = 0, M(0) = i),$$

where $\zeta_0^+ = \inf\{t > 0; X(t-) < 0 < X(t+)\}$. The following result plays a key role in our main result. We defer its proof to the Appendix.

Lemma 11.1. Let $-\alpha (< 0)$ be the Perron–Frobenius (P-F, for short) eigenvalue of the defective transition rate matrix U, and $\mathbf{q}^+$ be the corresponding positive right eigenvector; that is, $U\mathbf{q}^+ = -\alpha \mathbf{q}^+$. Under the negative drift condition (11.1), we have

$$\lim_{b \to \infty} e^{\alpha b} {}_0A_{0b}^{++} = (I - P_{00}^{++}) \mathbf{q}^+ \mathbf{u}^+ \Delta_{\mathbf{q}^+}^{-1},$$

where $\mathbf{u}^+$ is the stationary distribution of the nondefective transition rate matrix $\Delta_{\mathbf{q}^+}^{-1}(\alpha I + U)\Delta_{\mathbf{q}^+}$. Furthermore, $-\alpha$ is obtained as the solution of

$$\chi(z) = 0, \tag{11.4}$$

where $\chi(z)$ is the P-F eigenvalue of the following ML-matrix,

$$C + \int_0^\infty D(du)\exp(zu) - z\Delta_{\mathbf{r}}.$$

By Proposition 11.1, the hitting probability for an upper level has the matrix-exponential form. In general, the hitting probability for a lower level does not have a similar form because of the downward jumps. However, from [7], we know that the hitting probability for a lower level is obtained by integrating the matrix-exponential form. In what follows, we present this result. For $x > 0$, let $H^{+\bullet}(x) = (H^{+-}(x), H^{++}(x))$ be the $\mathscr{S}^+ \times \mathscr{S}$ matrix whose (i,j)th element is given by

$$[H^{+\bullet}(x)]_{ij} = \mathrm{P}(M(\tau_0^-) = j, X(\tau_0^-) \in (-x,0) | X(0) = 0, M(0) = i)$$

which is the first hitting probability for a lower level with a jump. Let H_0^{+-} be the $\mathscr{S}^+ \times \mathscr{S}^-$ matrix whose (i,j)th element is given by

$$[H_0^{+-}]_{ij} = \mathrm{P}(M(\tau_0^-) = j, X(\tau_0^-) = 0 | X(0) = 0, M(0) = i)$$

which is the first hitting probability for a lower level without jump. These hitting probabilities for a lower level are given as follows.

Proposition 11.2. (Lemma 3.1 of [7]) The hitting probability for a lower level with a jump is given by

$$(H^{+-}(x), H^{++}(x)) = \Delta_{\mathbf{r}^+}^{-1}\Delta_{\pi^+}^{-1}\left\{\int_0^x ds \int_s^\infty \Delta_\pi \tilde{D}(dy) \begin{pmatrix} \tilde{R}_0 \exp((y-s)\tilde{U}) \\ \exp((y-s)\tilde{U}) \end{pmatrix}\right\}^{\mathrm{T}},$$

where $\tilde{D}(y) = \Delta_\pi^{-1} D(y)^{\mathrm{T}} \Delta_\pi$. $\tilde{R}_0$ and $\tilde{U}$ are the $\mathscr{S}^- \times \mathscr{S}^+$ and $\mathscr{S}^+ \times \mathscr{S}^+$ matrices, respectively, satisfying

$$\begin{pmatrix} \tilde{R}_0 \\ I \end{pmatrix} \tilde{U} = \Delta_{\mathbf{r}}^{-1}\left\{\tilde{C}\begin{pmatrix} \tilde{R}_0 \\ I \end{pmatrix} + \int_0^\infty \tilde{D}(dy)\begin{pmatrix} \tilde{R}_0 \\ I \end{pmatrix}\exp(y\tilde{U})\right\},$$

where $\tilde{C} = \Delta_\pi^{-1} C^{\mathrm{T}} \Delta_\pi$. On the other hand, the hitting probability without jump is given by

$$H_0^{+-} = \Delta_{\mathbf{r}^+}^{-1}\Delta_{\pi^+}^{-1}\int_0^\infty ds \int_0^\infty \left\{\Delta_{\pi^-} K^{--}(s)\tilde{W}^{-\bullet}(dy)\begin{pmatrix} \tilde{R}_0 \\ I \end{pmatrix}\exp((s+y)\tilde{U})\right\}^{\mathrm{T}},$$

where $\tilde{W}^{-\bullet}(y)$ is the $\mathscr{S}^- \times \mathscr{S}$ matrix defined by $\Delta_{\pi^-}^{-1} W^{\bullet-}(y)^{\mathrm{T}} \Delta_\pi$. $W^{\bullet-}(y)$ is the $\mathscr{S} \times \mathscr{S}^-$ matrix whose (i,j)th element is given by

$$1_{\{i\neq j\}}[C]_{ij}\delta(y)+[D(y)]_{ij},$$

where 1_A is the indicator function for event A and $\delta(y)$ is the Dirac distribution which has a unit mass at the origin. $K^{--}(s)$ is the $\mathscr{S}^-\times\mathscr{S}^-$ diagonal matrix whose (k,k)th element is given by $\exp(c(k)s/r(k))$, where

$$c(k)=-[C]_{kk}.$$

Remark 11.2. $\tilde{R}_0$ and $\tilde{U}$ are obtained by a similar formula as noted in Remark 11.1.

By the negative drift condition (11.1), the following $\mathscr{S}^+\times\mathscr{S}$ matrix

$$(H_0^{+-}+H^{+-},H^{++}) \tag{11.5}$$

is stochastic, where $H^{+u}=\lim_{x\to\infty}H^{+u}(x)$ for $u=\pm 1$.

11.3 FIFQ (Feedback Infinite Fluid Queue) with Downward Jumps

In this section, we set a boundary to the MAP $(X(t),M(t))$ so that the additive component is reflected at level 0. Consider a two-dimensional CTMC $(Y(t),J(t))$ with a state space $[0,\infty)\times\mathscr{S}$, where its transition structure is given as follows (see Fig. 11.2).

(i) While $Y(t)>0$, $(Y(t),J(t))$ has the same transition structure as the MAP $(X(t),M(t))$.
(ii) When $Y(t)$ hits level 0, the transition rate matrix of $J(t)$ immediately changes to another $\mathscr{S}^-\times\mathscr{S}$ matrix:

$$\underline{C}=\left(\underline{C}^{--}\ \underline{C}^{-+}\right),$$

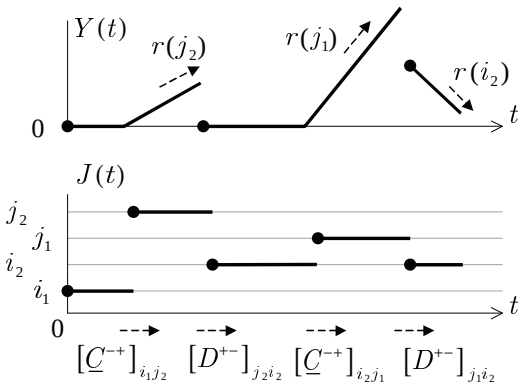


Fig. 11.2 FIFQ with downward jumps, where $\mathscr{S}^-=\{i_1,i_2\}$, and $\mathscr{S}^+=\{j_1,j_2\}$.

where $\underline{C}^{--}$ is the $\mathscr{S}^- \times \mathscr{S}^-$ ML-matrix, $\underline{C}^{-+}$ is the $\mathscr{S}^- \times \mathscr{S}^+$ nonnegative matrix, and $\underline{C}\mathbf{1} = \mathbf{0}$. This modification for the transition structure of the background process $J(t)$ is referred to as feedback. There are two types of the hitting level 0.

(iia) If it occurs due to C^{--} or D^{--}, $Y(t)$ stays in level 0 until $J(t)$ changes due to $\underline{C}^{-+}$. After $J(t)$ changes due to $\underline{C}^{-+}$, $Y(t)$ goes up from level 0. Then $(Y(t),J(t))$ again has the same transition structure as (i).

(iib) If it occurs due to D^{-+}, $Y(t)$ immediately goes up from level 0. Then $(Y(t),J(t))$ again has the same transition structure as (i).

We introduce a nonnegative (resp., positive) valued function r_{in} (resp., r_{out}) defined on $\mathscr{S}$. Assume that there is a fluid input (resp., output) at rate $r_{\text{in}}(i)$ (resp., $r_{\text{out}}(i)$) when $J(t) = i(\in \mathscr{S})$. In this chapter, the net flow rate r is given by the difference of the input and output rates; that is,

$$r = r_{\text{in}} - r_{\text{out}}.$$

Then $(Y(t),J(t))$ is referred to as a feedback infinite fluid queue (FIFQ, for short) with downward jumps, or simply referred to as FIFQ.

By the negative drift condition (11.1), there exists a stationary distribution for $(Y(t),J(t))$.

Proposition 11.3. (Theorem 4.1 of [8]) For $x > 0$, we have

$$\mathrm{P}(Y > x, J = i) = \begin{cases} [\Delta_{\pi^-} R_0 \exp(xU)\mathbf{1}]_i, & i \in \mathscr{S}^- \\ [\Delta_{\pi^+} \exp(xU)\mathbf{1}]_i, & i \in \mathscr{S}^+, \end{cases}$$

where (Y,J) means $(Y(t),J(t))$ in steady state.

Let $\mathbf{p}^-$ be the $\mathscr{S}^-$-dimensional row vector whose ith element is given by

$$[\mathbf{p}^-]_i = \mathrm{P}(Y = 0, J = i).$$

By censoring $(Y(t),J(t))$ at subspace $\{0\} \times \mathscr{S}^-$, $\mathbf{p}^-$ is obtained as a stationary measure for $Q_{00} = \underline{C}^{--} + \underline{C}^{-+}(I - H^{++})^{-1}(H_0^{+-} + H^{+-})$; that is,

$$\mathbf{p}^- Q_{00} = \mathbf{0}.$$

Note that Q_{00} is a nondefective transition rate matrix by the negative drift condition (11.1). For $x > 0$, let $\mathbf{p}(x)$ be the $\mathscr{S}$- dimensional row vector whose ith element is given by

$$[\mathbf{p}(x)]_i = \mathrm{P}(Y > x, J = i).$$

Then $\mathbf{p}^-$ is normalized so that $\mathbf{p}^-\mathbf{1} + \mathbf{p}(0)\mathbf{1} = 1$; that is,

$$\mathbf{p}^-\mathbf{1} + \pi^- R_0 \mathbf{1} + \pi^+ \mathbf{1} = 1 \tag{11.6}$$

by Proposition 11.3.

11.4 FFFQ (Feedback Finite Fluid Queue) with Downward Jumps

In this section, we put another boundary to the FIFQ with downward jumps $(Y(t),J(t))$ in such a way that the additive component is also reflected at level b, where $b>0$. We further assume that the transition structure of the background process may change at level b. Denote this reflected additive process by $(Y^{(b)}(t),J^{(b)}(t))$, which is a two-dimensional CTMC with a state space $[0,b]\times\mathscr{S}$. The background process $J^{(b)}(t)$ has the following three types of transition structures depending on the level $Y^{(b)}(t)$ (see Fig. 11.3).

(i) While $Y^{(b)}(t)$ stays in $(0,b)$, $(Y^{(b)}(t),J^{(b)}(t))$ has the same transition structure as FIFQ $(Y(t),J(t))$; that is, $J^{(b)}(t)$ is a CTMC with transition rate matrix $C+D$.

(ii) When $Y^{(b)}(t)$ hits level b, the transition rate matrix of $J^{(b)}(t)$ immediately changes to another $\mathscr{S}^+\times\mathscr{S}$ matrix:

$$\overline{C}+\overline{D}=\left(\overline{C}^{+-}\ \overline{C}^{++}\right)+\left(\overline{D}^{+-}\ \overline{D}^{++}\right),$$

where $\overline{C}^{++}$ is the $\mathscr{S}^+\times\mathscr{S}^+$ ML-matrix, $\overline{C}^{+-}$, $\overline{D}^{+-}$ and $\overline{D}^{++}$ are the $\mathscr{S}^+\times\mathscr{S}^-$, $\mathscr{S}^+\times\mathscr{S}^-$, and $\mathscr{S}^+\times\mathscr{S}^+$ nonnegative matrices, respectively, and $(\overline{C}+\overline{D})\mathbf{1}=\mathbf{0}$. After $Y^{(b)}(t)$ hits level b, there can be the following three cases.

(iia) If $J^{(b)}(t)$ changes due to $\overline{C}^{++}$, $Y^{(b)}(t)$ stays in level b.

(iib) If $J^{(b)}(t)$ changes due to $\overline{C}^{+-}$, $Y^{(b)}(t)$ goes below level b and $(Y^{(b)}(t),J^{(b)}(t))$ again has the same transition structure as (i).

(iic) If $J^{(b)}(t)$ changes due to $\overline{D}$, $Y^{(b)}(t)$ jumps down below level b. The jump size is distributed subject to $\overline{D}(x)$, where $\overline{D}(x)$ is the $\mathscr{S}^+\times\mathscr{S}$ matrix whose

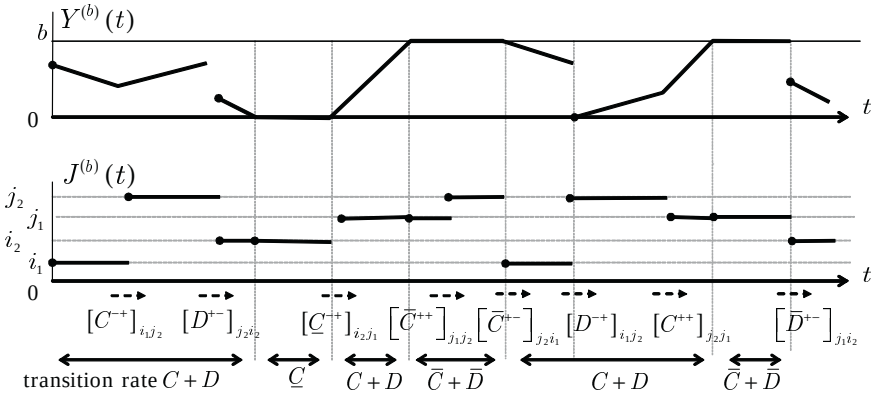


Fig. 11.3 FFFQ with downward jumps, where $\mathscr{S}^-=\{i_1,i_2\}$, $\mathscr{S}^+=\{j_1,j_2\}$.

(i,j)th element is $[\overline{D}]_{ij}G_{ij}(x)$, where $G_{ij}(x)$ is a distribution function. When the jump size is less than b, $(Y^{(b)}(t),J^{(b)}(t))$ has the same transition structure as (i). Otherwise, $Y^{(b)}(t)$ hits level 0 and $(Y^{(b)}(t),J^{(b)}(t))$ has the same transition structure as (iii).

(iii) When $Y^{(b)}(t)$ hits level 0, the transition rate matrix of $J^{(b)}(t)$ changes to $\underline{C} = (\underline{C}^{--},\underline{C}^{-+})$. There are two types of the hitting level 0.

(iiia) If it occurs due to C^{--}, D^{--}, or $\overline{D}^{+-}$, $Y^{(b)}(t)$ stays in level 0 until $J^{(b)}(t)$ changes due to $\underline{C}^{-+}$. After $J^{(b)}(t)$ changes due to $\underline{C}^{-+}$, $Y^{(b)}(t)$ goes up from level 0. Then $(Y^{(b)}(t),J^{(b)}(t))$ again has the same transition structure as (i).

(iiib) If it occurs due to D^{-+} or $\overline{D}^{++}$, $Y^{(b)}(t)$ immediately goes up from 0. Then $(Y^{(b)}(t),J^{(b)}(t))$ again has the same transition structure as (i).

This reflected additive process $(Y^{(b)}(t),J^{(b)}(t))$ is referred to as a feedback finite fluid queue (FFFQ, for short) with downward jumps, or simply referred to as FFFQ.

11.5 Asymptotic Behavior of Loss Rate for FFFQ with Downward Jumps

In this section, we study the asymptotic behavior of the loss rate $\ell_{\rm Loss}^{(b)}$ for the FFFQ with downward jumps $(Y^{(b)}(t),J^{(b)}(t))$ as the buffer size b gets large. Let $\mathbf{p}^{(b)}$ be the $\mathscr{S}$-dimensional probability vector whose ith element is given by

$$[\mathbf{p}^{(b)}]_i = \begin{cases} \mathrm{P}(Y^{(b)}=0,J^{(b)}=i), & i\in\mathscr{S}^- \\ \mathrm{P}(Y^{(b)}=b,J^{(b)}=i), & i\in\mathscr{S}^+, \end{cases}$$

where $(Y^{(b)},J^{(b)})$ means $(Y^{(b)}(t),J^{(b)}(t))$ in steady state. We partition $\mathbf{p}^{(b)}$ according to $\mathscr{S}^-$ and $\mathscr{S}^+$ such that $\mathbf{p}^{(b)} = (\mathbf{p}^{(b)-},\mathbf{p}^{(b)+})$. Then the loss rate $\ell_{\rm Loss}^{(b)}$ is given by

$$\ell_{\rm Loss}^{(b)} = \mathbf{p}^{(b)+}\mathbf{r}_{\rm in}^{+}, \tag{11.7}$$

where $\mathbf{r}_{\rm in}^{+} = (r_{\rm in}(i);i\in\mathscr{S}^+)$. So it is sufficient to consider the asymptotic behavior of $\mathbf{p}^{(b)+}$ as b gets large.

Note that $\mathbf{p}^{(b)}$ is a stationary measure for $(Y^{(b)}(t),J^{(b)}(t))$ censoring at subspace

$$\mathscr{S}_{\{0,b\}} = (\{0\}\times\mathscr{S}^-)\cup(\{b\}\times\mathscr{S}^+).$$

We introduce the following two hitting probabilities (see Fig. 11.4). Let ${}_b\Psi_{x0}^{\bullet-}$ be the $\mathscr{S}\times\mathscr{S}^-$ matrix for $x\in[0,b]$ whose (i,j)th element is given by

$$[{}_b\Psi_{x0}^{\bullet-}]_{ij} = \mathrm{P}(J^{(b)}(\tau_0^{(b)-})=j,\tau_0^{(b)-}<\tau_b^{(b)+}|Y^{(b)}(0)=x,J^{(b)}(0)=i)$$

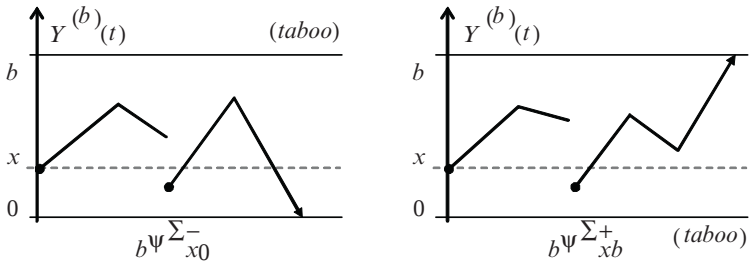


Fig. 11.4 Hitting probabilities for FFFQ with downward jumps.

and ${}_0\Psi^{\bullet+}_{xb}$ be the $\mathscr{S}\times\mathscr{S}^+$ matrix for $x\in[0,b]$ whose (i,j)th element is given by

$$[{}_0\Psi^{\bullet+}_{xb}]_{ij}=\mathrm{P}(J^{(b)}(\tau_b^{(b)+})=j,\tau_b^{(b)+}<\tau_0^{(b)-}|Y^{(b)}(0)=x,J^{(b)}(0)=i),$$

where $\tau_0^{(b)-}=\inf\{t>0|Y^{(b)}(t)=0,J^{(b)}(t)\in\mathscr{S}^-\}$ and $\tau_b^{(b)+}=\inf\{t>0\mid Y^{(b)}(t)=b,J^{(b)}(t)\in\mathscr{S}^+\}$. Partition ${}_b\Psi^{\bullet-}_{x0}$ and ${}_0\Psi^{\bullet+}_{xb}$ into blocks according to $\mathscr{S}^-$ and $\mathscr{S}^+$ such that

$${}_b\Psi^{\bullet-}_{x0}=\begin{pmatrix}{}_b\Psi^{--}_{x0}\\ {}_b\Psi^{+-}_{x0}\end{pmatrix},\qquad {}_0\Psi^{\bullet+}_{xb}=\begin{pmatrix}{}_0\Psi^{-+}_{xb}\\ {}_0\Psi^{++}_{xb}\end{pmatrix}.$$

Then the transition rate matrix for the censored process at subspace $\mathscr{S}_{\{0,b\}}$ is given by

$$Q=\begin{pmatrix}Q^{(b)}_{00} & Q^{(b)}_{0b}\\ Q^{(b)}_{b0} & Q^{(b)}_{bb}\end{pmatrix},$$

where the each submatrix is given by

$$\begin{aligned}
Q^{(b)}_{00}&=\underline{C}^{--}+\underline{C}^{-+}{}_b\Psi^{+-}_{00},\qquad Q^{(b)}_{0b}=\underline{C}^{-+}{}_0\Psi^{++}_{0b},\\
Q^{(b)}_{b0}&=\overline{C}^{+-}{}_b\Psi^{--}_{b0}+\int_0^b\overline{D}^{+-}(dx){}_b\Psi^{--}_{(b-x)0}+\int_b^\infty\overline{D}^{+-}(dx)\\
&\quad+\int_0^b\overline{D}^{++}(dx){}_b\Psi^{+-}_{(b-x)0}+\int_b^\infty\overline{D}^{++}(dx){}_b\Psi^{+-}_{00},\\
Q^{(b)}_{bb}&=\overline{C}^{++}+\overline{C}^{+-}{}_0\Psi^{-+}_{bb}+\int_0^b\overline{D}^{++}(dx){}_0\Psi^{++}_{(b-x)b}+\int_b^\infty\overline{D}^{++}(dx){}_0\Psi^{++}_{0b}\\
&\quad+\int_0^b\overline{D}^{+-}(dx){}_0\Psi^{-+}_{(b-x)b}.
\end{aligned}$$

Then $\mathbf{p}^{(b)-}$ and $\mathbf{p}^{(b)+}$ satisfy

$$\mathbf{p}^{(b)-}Q^{(b)}_{00}+\mathbf{p}^{(b)+}Q^{(b)}_{b0}=\mathbf{0},\qquad \mathbf{p}^{(b)-}Q^{(b)}_{0b}+\mathbf{p}^{(b)+}Q^{(b)}_{bb}=\mathbf{0}. \tag{11.8}$$

By the negative drift condition (11.1), we have

$$\lim_{b\to\infty} {}_0\Psi_{0b}^{++} = 0, \qquad \lim_{b\to\infty} Q_{bb}^{(b)} = \hat{Q}_{00},$$

where $\hat{Q}_{00}$ is a defective transition rate matrix. Because $\lim_{b\to\infty} Q_{0b}^{(b)} = 0$, we have

$$\lim_{b\to\infty} \mathbf{p}^{(b)+} = \mathbf{0}.$$

Furthermore, we have

$$\lim_{b\to\infty} {}_b\Psi_{00}^{+-} = (I - H^{++})^{-1}(H_0^{+-} + H^{+-}),$$

because the effect of level b disappears as b gets large. Hence, (11.8) and the above observations imply that

$$\lim_{b\to\infty} \mathbf{p}^{(b)-} = \mathbf{p}^{-},$$

which is a stationary measure for Q_{00} with the normalizing condition (11.6).

By the second equation of (11.8), we have

$$\mathbf{p}^{(b)+}(-Q_{bb}^{(b)}) = \mathbf{p}^{(b)-}\underline{C}^{-+}{}_0\Psi_{0b}^{++}, \tag{11.9}$$

which implies that the asymptotic behavior of $\mathbf{p}^{(b)+}$ is determined by that of ${}_0\Psi_{0b}^{++}$. From Proposition 11.1, Proposition 11.2, and Lemma 11.1, we arrive at the main result. Its proof is deferred to the appendix.

Theorem 11.1. The asymptotic behavior of the loss rate is given by

$$\lim_{b\to\infty} e^{\alpha b}\ell_{\text{Loss}}^{(b)} = \mathbf{p}^{-}\underline{C}^{-+}(I - H^{++})^{-1}(I - P_{00}^{++})(\mathbf{q}^{+}\mathbf{u}^{+}\Delta_{\mathbf{q}^{+}}^{-1})(-\hat{Q}_{00})^{-1}\mathbf{r}_{\text{in}}^{+},$$

where $-\alpha$ is the solution of (11.4), and $\hat{Q}_{00}$, $P_{00}^{++}\mathbf{q}^{+}$ are obtained as follows:

(I) $\hat{Q}_{00} = \overline{C}^{++} + \overline{C}^{+-}R_0 + \int_0^\infty \overline{D}^{++}(dx)\exp(xU) + \int_0^\infty \overline{D}^{+-}(dx)R_0\exp(xU)$.

(II) $P_{00}^{++}\mathbf{q}^{+}$ is given by

$$\left\{ H_0^{+-}R_0 + \Delta_{\mathbf{r}^{+}}^{-1}\Delta_{\pi^{+}}^{-1}\int_0^\infty \left(\begin{pmatrix}\tilde{R}_0\\ I\end{pmatrix}\exp(y\tilde{U})V(y)\right)^{\mathrm{T}}\Delta_{\pi}D(dy)\begin{pmatrix}R_0\\ I\end{pmatrix}\right\}\mathbf{q}^{+},$$

where

$$V(y) = \Delta_{\tilde{\mathbf{q}}^{+}}(\mathbf{1}\boldsymbol{\kappa} - \hat{U})^{-1}\left(\exp(-y\hat{U}) + y\mathbf{1}\boldsymbol{\kappa} - I\right)\Delta_{\tilde{\mathbf{q}}^{+}}^{-1}$$

and $\tilde{\mathbf{q}}^{+}$ is the P-F right eigenvector for $\tilde{U}$ with the P-F eigenvalue $-\alpha$. $\boldsymbol{\kappa}$ is the stationary distribution for $\hat{U} = \Delta_{\tilde{\mathbf{q}}^{+}}^{-1}(\alpha I + \tilde{U})\Delta_{\tilde{\mathbf{q}}^{+}}$.

11.6 Numerical Examples

We provide some numerical examples for the FFFQ with downward jumps by computing the positive constants α and c such that

$$\lim_{b\to\infty} e^{\alpha b}\ell_{\text{Loss}}^{(b)} = c.$$

This indicates that we may approximate the loss rate by $ce^{-\alpha b}$. Suppose that the jump sizes are deterministic for each possible transition. That is, let B be the $\mathscr{S}\times\mathscr{S}$ matrix, whose (i,j)th element denotes the jump size of the buffer when the background state changes from i to j. Then $D(x)$ is given by

$$[D(x)]_{ij} = [D]_{ij}1_{\{x=[B]_{ij}\}}.$$

Similarly, let $\overline{B}^{+-}$ (resp., $\overline{B}^{++}$) be the $\mathscr{S}^+\times\mathscr{S}^-$ (resp., $\mathscr{S}^+\times\mathscr{S}^+$) matrix, whose (i,j)th element denotes the jump size when the background state changes from i to j at level b. Then $\overline{D}^{+-}(x)$ and $\overline{D}^{++}(x)$ are given by

$$[\overline{D}^{+-}(x)]_{ij} = [\overline{D}^{+-}]_{ij}1_{\{x=[\overline{B}^{+-}]_{ij}\}}, \qquad [\overline{D}^{++}(x)]_{ij} = [\overline{D}^{++}]_{ij}1_{\{x=[\overline{B}^{++}]_{ij}\}}.$$

Assume the following parameter settings.

$$\mathscr{S}^-=\{0,1\}, \qquad \mathscr{S}^+=\{2\},$$

$$\begin{pmatrix} r_{\text{in}}(0)\\ r_{\text{in}}(1)\end{pmatrix} = \begin{pmatrix} 5.5\\ 8.0\\ 10.0\end{pmatrix}, \qquad \begin{pmatrix} r_{\text{out}}(0)\\ r_{\text{out}}(1)\\ r_{\text{out}}(2)\end{pmatrix} = \begin{pmatrix} 6.0\\ 8.7\\ 6.0\end{pmatrix},$$

$$C = \begin{pmatrix} -4.6 & 1.5 & 2.3\\ 0.6 & -2.7 & 1.2\\ 0.5 & 0.8 & -2.1\end{pmatrix}, \qquad D = \begin{pmatrix} 0.2 & 0.1 & 0.5\\ 0.3 & 0.4 & 0.2\\ 0.5 & 0.1 & 0.2\end{pmatrix}, \qquad B = \begin{pmatrix} 2.3 & 1.3 & 1.5\\ 0.5 & 1.8 & 2.4\\ 2.4 & 5.0 & 2.1\end{pmatrix},$$

$$\underline{C}^{--} = \begin{pmatrix} -3.0 & 1.0\\ 1.0 & -2.0\end{pmatrix}, \qquad \underline{C}^{-+} = \begin{pmatrix} 2.0\\ 1.0\end{pmatrix},$$

$$\overline{C}^{+-} = \begin{pmatrix} 1.5 & 1.3\end{pmatrix}, \qquad \overline{C}^{++} = (-4.1), \qquad \overline{D}^{+-} = \begin{pmatrix} 0.4 & 0.5\end{pmatrix}, \qquad \overline{D}^{++} = (0.3),$$

$$\overline{B}^{+-} = \begin{pmatrix} 1.1 & 5.0\end{pmatrix}, \qquad \overline{B}^{++} = (2.7).$$

In this case, we have

$$\text{mean drift} = -0.046225 < 0, \qquad \alpha = 0.012454, \qquad c = 2.213758.$$

We further consider the following two cases.

(case 1) Change the jump size due to D-transition: $[B]_{21} = 5.0 \to [B]_{21} = 15.0$. Then we have

$$\text{mean drift} = -0.546513 < 0, \qquad \alpha = 0.077897, \qquad c = 3.469491.$$

because the jump size is increased when the additive component is below level b, the decay rate α considerably gets larger.

(case 2) Change the jump size due to $\overline{D}$-transition: $[\overline{B}]_1 = 5.0 \to [\overline{B}]_1 = 15.0$. Then we have

$$\text{mean drift} = -0.046225, \qquad \alpha = 0.012454, \qquad c = 1.394617.$$

because the jump size is increased when the additive component stays in level b, only the prefactor c decreases.

11.7 Conclusions

In this chapter, we studied the tail behavior of the loss rate for the feedback fluid queue with a finite buffer. By using the relations between the fluid queue and the Markov additive process with downward jumps, we showed that the loss rate asymptotically decays at an exponential rate with a constant prefactor as the buffer size gets large. This decay rate was obtained by the additive process; that is, it is irrelevant to the boundary condition for the fluid queue.

Appendix

Proof of Lemma 11.1. Consider the $\mathscr{S}^+ \times \mathscr{S}^+$ matrix ${}_bP_{00}^{++}$ whose (i,j)th element is given by

$$[{}_bP_{00}^{++}]_{ij} = \mathrm{P}(M(\zeta_0^{(b)+}) = j | X(0) = 0, M(0) = i),$$

where $\zeta_0^{(b)+} = \inf\{t > 0; X(t-) < 0 < X(t+),\ X(u) < b, u \in (0,t)\}$ is the first time when the MAP $(X(t), M(t))$ crosses level 0 from below, avoiding level b (see Fig. 11.5). Note that $\lim_{b\to\infty} {}_bP_{00}^{++} = P_{00}^{++}$, because the effect of level b disappears as b gets large. By conditioning on the event that the MAP $(X(t), M(t))$ crosses the initial level from below for the first time, we have $R^{++}(b) = {}_0A_{0b}^{++} + {}_bP_{00}^{++} R^{++}(b)$.

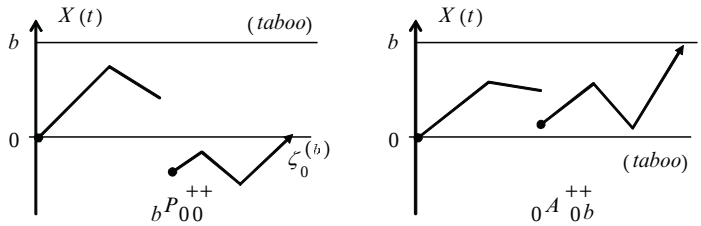


Fig. 11.5 Hitting probabilities for MAP with downward jumps.

By Proposition 11.1, we have

$$ {}_0A_{0b}^{++} = (I - {}_bP_{00}^{++}) \exp(bU). \tag{11.10} $$

Because U is defective, there exists the P-F eigenvalue $-\alpha(<0)$ and the corresponding positive right eigenvector $\mathbf{q}^+$. Note that $\Delta_{\mathbf{q}^+}^{-1}(\alpha I + U)\Delta_{\mathbf{q}^+}$ is a non defective transition rate matrix. So it has the stationary distribution $\mathbf{u}^+$; that is, $\mathbf{u}^+\Delta_{\mathbf{q}^+}^{-1}(\alpha I + U)\Delta_{\mathbf{q}^+} = \mathbf{0}$ and $\mathbf{u}^+\mathbf{1} = 1$. By the standard Markov chain theory, we have

$$ \lim_{b\to\infty} \exp(b\Delta_{\mathbf{q}^+}^{-1}(\alpha I + U)\Delta_{\mathbf{q}^+}) = \mathbf{1}\mathbf{u}^+, $$

which is equivalent to

$$ \lim_{b\to\infty} e^{\alpha b} \exp(bU) = \mathbf{q}^+\mathbf{u}^+\Delta_{\mathbf{q}^+}^{-1}. \tag{11.11} $$

Combining (11.10) with (11.11) yields

$$ \lim_{b\to\infty} e^{\alpha b}\, {}_0A_{0b}^{++} = (I - P_{00}^{++})\mathbf{q}^+\mathbf{u}^+\Delta_{\mathbf{q}^+}^{-1}. $$

By postmultiplying $\mathbf{q}^+$ to (11.2), we have

$$ \left(-\alpha I - \Delta_{\mathbf{r}}^{-1}\left(C + \int_0^\infty \exp(-\alpha u) D(du)\right)\right)\begin{pmatrix} I \\ R_0 \end{pmatrix}\mathbf{q}^+ = \mathbf{0}, $$

which implies that $-\alpha$ is obtained as a solution of $\chi(z) = 0$.

Proof of Theorem 11.1. Consider the hitting probability that the MAP $(X(t), M(t))$ jumps below level 0 with a background state in $\mathscr{S}^+$, starting from level 0 with a background state in $\mathscr{S}^+$, avoiding level b. This is equivalent to the probability that the FFFQ $(Y^{(b)}(t), J^{(b)}(t))$ returns to level 0 while increasing, starting from level 0 with a background state in $\mathscr{S}^+$, avoiding level b. Let ${}_bH^{++}$ be the $\mathscr{S}^+ \times \mathscr{S}^+$ matrix whose (i,j)th element is given by

$$ [{}_bH^{++}]_{ij} = \mathrm{P}(M(\tau_0^-) = j, X(\tau_0^-) < 0, \tau_0^- < \tau_b^+ | X(0) = 0, M(0) = i). $$

Note that $\lim_{b\to\infty} {}_bH^{++} = H^{++}$. By conditioning on the event that $(Y^{(b)}(t), J^{(b)}(t))$ returns to level 0, we have ${}_0\Psi_{0b}^{++} = {}_0A_{0b}^{++} + {}_bH^{++}\,{}_0\Psi_{0b}^{++}$. Because ${}_bH^{++}$ is substochastic, we have

$$ {}_0\Psi_{0b}^{++} = (I - {}_bH^{++})^{-1}\, {}_0A_{0b}^{++}. \tag{11.12} $$

From (11.9) and (11.12), we have $\mathbf{p}^{(b)+} = \mathbf{p}^{(b)-}\underline{C}^{-+}(I - {}_bH^{++})^{-1}\, {}_0A_{0b}^{++}(-Q_{bb}^{(b)})^{-1}$, which implies that

$$ \lim_{b\to\infty} e^{\alpha b}\mathbf{p}^{(b)+} = \mathbf{p}^-\underline{C}^{-+}(I - H^{++})^{-1}(I - P_{00}^{++})(\mathbf{q}^+\mathbf{u}^+\Delta_{\mathbf{q}^+}^{-1})(-\hat{Q}_{00})^{-1} \tag{11.13} $$

by Lemma 11.1. In the following, we compute $\hat{Q}_{00}$, $\mathbf{p}^-$, and $P_{00}^{++}\mathbf{q}^+$ in the right side of (11.13). This completes the proof of Theorem 11.1,

Proof of (I). By the definition of $Q_{bb}^{(b)}$ and the dominated convergence theorem, we have

$$\hat{Q}_{00} = \overline{C}^{++} + \overline{C}^{+-}\hat{\Psi}_{00}^{-+} + \int_0^\infty \overline{D}^{++}(dx)\hat{\Psi}_{x0}^{++} + \int_0^\infty \overline{D}^{+-}(dx)\hat{\Psi}_{x0}^{-+},$$

where $\hat{\Psi}_{00}^{-+} = \lim_{b\to\infty} {}_0\Psi_{bb}^{-+}$, $\hat{\Psi}_{x0}^{++} = \lim_{b\to\infty} {}_0\Psi_{(b-x)b}^{+}$, and $\hat{\Psi}_{x0}^{-+} = \lim_{b\to\infty} {}_0\Psi_{(b-x)b}^{-+}$. From Proposition 11.1 and the definition of ${}_0\Psi_{xb}^{\bullet+}$, we have

$$\hat{\Psi}_{00}^{-+} = R_0, \qquad \hat{\Psi}_{x0}^{++} = \exp(xU), \qquad \hat{\Psi}_{x0}^{-+} = R_0\exp(xU).$$

Thus we have (I).

Proof of (II). By conditioning on the event that the MAP $(X(t), M(t))$ crosses level 0 from below, P_{00}^{++} is given by $H_0^{+-}R_0 + \int_0^\infty H^{+\bullet}(du)R^{\bullet+}(u)$; that is,

$$H_0^{+-}R_0 + \Delta_{\mathbf{r}^+}^{-1}\Delta_{\pi^+}^{-1}\left\{\int_0^\infty du\int_u^\infty \begin{pmatrix}\tilde{R}_0\exp((y-u)\tilde{U})\\ \exp((y-u)\tilde{U})\end{pmatrix}^{\mathrm{T}}\Delta_\pi D(dy)\begin{pmatrix}R_0\\ I\end{pmatrix}\exp(uU)\right\}$$

by Proposition 11.2. By postmultiplying $\mathbf{q}^+$, changing the order of integration, and $U\mathbf{q}^+ = \alpha\mathbf{q}^+$, we have

$$P_{00}^{++}\mathbf{q}^+ = \left\{H_0^{+-}R_0 + \Delta_{\mathbf{r}^+}^{-1}\Delta_{\pi^+}^{-1}\int_0^\infty \begin{pmatrix}\tilde{R}_0\exp(y\tilde{U})\Delta_{\tilde{\mathbf{q}}^+}\int_0^y du\ \exp(-u\hat{U})\Delta_{\tilde{\mathbf{q}}^+}^{-1}\\ \exp(y\tilde{U})\Delta_{\tilde{\mathbf{q}}^+}\int_0^y du\ \exp(-u\hat{U})\Delta_{\tilde{\mathbf{q}}^+}^{-1}\end{pmatrix}^{\mathrm{T}}\right.$$
$$\left.\cdot\,\Delta_\pi D(dy)\begin{pmatrix}R_0\\ I\end{pmatrix}\right\}\mathbf{q}^+, \tag{11.14}$$

where $\hat{U} = \Delta_{\tilde{\mathbf{q}}^+}^{-1}(\alpha I + \tilde{U})\Delta_{\tilde{\mathbf{q}}^+}$. Because $\tilde{U}$ also has the P-F eigenvalue $-\alpha$, denote the corresponding positive right eigenvector by $\tilde{\mathbf{q}}^+$; that is, $\tilde{U}\tilde{\mathbf{q}}^+ = -\alpha\tilde{\mathbf{q}}^+$. Because $\hat{U}$ is a nondefective transition rate matrix, denote its stationary distribution by $\boldsymbol{\kappa}$. Then we have

$$\int_0^y du\ \exp(-u\hat{U}) = (\mathbf{1}\boldsymbol{\kappa} - \hat{U})^{-1}(\exp(-y\hat{U}) + y\mathbf{1}\boldsymbol{\kappa} - I). \tag{11.15}$$

From (11.7) and (11.15), we have (II).

References

1. D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic theory of a data handling system with multiple sources," *Bell Systems Technical Journal*, vol. 61, pp. 1871–1894, 1982.
2. T. Bonald, "Comparison of TCP reno and TCP vegas via fluid approximation," *INRIA Research Report*, no. 3563, 1998.
3. N. Foreest, M. Mandjes, and W. Scheinhardt, "Analysis of a feedback fluid model for heterogeneous TCP sources," *Stochastic Models*, vol. 19, pp. 299–324, 2003.
4. D.H. Chiu and R. Jain, "Analysis of the increase and decrease algorithms of congestion avoidance in computer networks," *Computer Networks and ISDN Systems*, vol. 17, pp. 1–14, 1989.
5. A. da Silva Soares and G. Latouche, "A matrix-analytic approach to fluid queues with feedback control," in *Proc. 11th International Conference on Analytical and Stochastic Modelling Techniques and Applications*, pp. 190–198, 2004.
6. S. Asmussen and M. Pihlsgård, "Loss rates for Lévy processes with two reflecting barriers," *Mathematics of Operations Research*, vol. 32, pp. 308–321, 2007.
7. M. Miyazawa, "Hitting probabilities in a Markov additive process with linear movements and upward jumps: applications to risk and queueing processes," *The Annals of Applied Probability*, vol. 14, pp. 1029–1054, 2004.
8. M. Miyazawa and H. Takada, "A matrix exponential form for hitting probabilities and its application to a Markov-modulated fluid queue with downward jumps," *Journal of Applied Probability*, vol. 39, pp. 604–618, 2002.
9. V. Ramaswami, "Matrix analytic methods for stochastic fluid flows," in *Proc.* 16*th International Teletraffic Congress*, pp. 1019–1030, 1999.
10. A. da Silva Soares and G. Latouche, "Matrix-analytic methods for fluid queues with finite buffers," *Performance Evaluation*, vol. 63, pp. 295–314, 2006.
11. M. Miyazawa, Y. Sakuma, and S. Yamaguchi, "Asymptotic behaviors of the loss probability for a finite buffer queue with QBD structure," *Stochastic Models*, vol. 23, pp. 79–95, 2007.
12. H. Takada, "Markov modulated fluid queues with batch fluid arrivals," *Journal of the Operations Research Society of Japan*, vol. 44, pp. 344–365, 2001.
13. R. Bellman, "Introduction to Matrix Analysis," *SIAM*, Philadelphia, 2nd edition, 1997.

Chapter 12
Explicit Probability Density Function for the Length of a Busy Period in an M/M/1/K Queue

Hideaki Takagi and Ahmed M.K. Tarabia

Abstract A new closed-form explicit expression is derived for the probability density function of the length of a busy period starting with i customers in an M/M/1/K queue, where K is the capacity of the system. The density function is given as a weighted sum of K negative exponential distributions with coefficients calculated from K distinct zeros of a polynomial that involves Chebyshev polynomials of the second kind. The mean and second moment of the busy period are also shown explicitly. In addition, the symmetric results for the first passage time from state i to state K are presented. We also consider the regeneration cycle of state i.

12.1 Introduction

Busy period analysis plays a significant role in the understanding of queueing systems and their efficient management. In particular, queueing systems with finite capacity are important in the design and development of telecommunication systems. The reader is referred to Perros and Altiok [1] for further details of such applications. A busy period in a queueing system normally starts with the arrival of a customer who finds the system empty, and ends with the first time at which the system becomes empty again. One may also consider a busy period starting with more than one customer in the system.

The transient behavior of the queue size in an M/M/1/K queue has a nice closed-form explicit expression for the probability distribution as shown in Takács

H. Takagi
Graduate School of Systems and Information Engineering, University of Tsukuba, Ibaraki 305-8573, Japan
e-mail: takagi@sk.tsukuba.ac.jp

A.M.K. Tarabia
Mathematics Department, Damietta Faculty of Science, New Damietta, Egypt
e-mail: a_tarabia@hotmail.com

W. Yue et al. (eds.), *Advances in Queueing Theory and Network Applications,*

[2, pp. 12–21]. However, finding the explicit formula of the busy period distribution for the M/M/1/K queue is an open problem. This is because the difficulty in obtaining the exact values of eigenvalues of its transition matrix does not allow an explicit solution.

There are a few results about the busy period of this model. Ismailov [3] obtained the Laplace transform of the duration of the busy period in an M/M/1/K queue. Srivastava and Kashyap [4, p. 61] also show the Laplace transform. Sharma and Shobha [5] obtained a closed-form expression for the busy period density function through an elegant algebraic method. See also Sharma's book [6, pp. 45–48]. But their solution involves the eigenvalues of a matrix. Stadje [7] determined the joint transform of the duration of a busy period and the number of customers served in it for the simple exponential queue with finite capacity. Kinateder and Lee [8] provided a new approach to the computation of the Laplace transform of the length of the busy period of the M/M/1 queue with constrained workload (finite dam) without the use of complex analysis. Reference to the studies of the busy period in non-Markovian queues with finite capacity is omitted here.

In this chapter, our motivation is not only to drive a new modified formula for the busy period distribution, but also to extend it to allow for any arbitrary number of initial customers $i \geq 1$. Moreover, we illustrate that the formula given in Sharma and Shobha [5] is not valid for some values of the traffic intensity. In addition, we refer to the first passage time from state i to K as a symmetric problem dealt with in Saaty [9, p. 129], and derive similar results. Finally we consider the time interval between two instants at which the system enters state i successively.

The chapter is organized as follows. Section 12.2 describes a busy period considered in this chapter in detail. First passage time to the system capacity and regeneration cycle are presented in Sect. 12.3 and Sect. 12.4. We conclude this chapter in Sect. 12.5.

12.2 Busy Period

We consider an M/M/1/K queue with arrival rate λ and service rate μ, where K denotes the capacity of the system including the one in the server. Let $b_i(t)$ be the probability density function (pdf) of the length $\mathscr{B}_i$ of a busy period starting with i customers ($1 \leq i \leq K$). This is equivalent to the first passage time from state i to state 0 in the birth-and-death process with a reflecting barrier at state K, where the state k means that there are k customers present in the system.

Let $N(t)$ be the number of customers present in the system at time t, and

$$P_{ik}(t) := P\{N(t) = k, 0 < N(u) \leq K \text{ for } 0 \leq u < t \mid N(0) = i\}, \qquad 1 \leq k \leq K.$$

Because $b_i(t)\Delta t + o(\Delta t) = P\{t < \mathscr{B}_i < t + \Delta t\} = P_{i1}(t) \cdot \mu \Delta t + o(\Delta t)$, it follows that

$$b_i(t) = \mu P_{i1}(t), \qquad 1 \leq i \leq K.$$

For the Laplace transforms

$$P_{ik}^*(s) := \int_0^\infty e^{-st} P_{ik}(t)dt, \qquad B_i^*(s) := \int_0^\infty e^{-st} b_i(t)dt,$$

we have

$$B_i^*(s) = \mu P_{i1}^*(s).$$

Equations for $\{P_{ik}^*(s), 1 \le k \le K\}$ are given by

$$\begin{aligned}
(s+\lambda+\mu)P_{i1}^*(s) - \mu P_{i2}^*(s) &= \delta_{i1}, \\
-\lambda P_{i,k-1}^*(s) + (s+\lambda+\mu)P_{ik}^*(s) - \mu P_{i,k+1}^*(s) &= \delta_{ik}, \qquad 2 \le k \le K-1, \\
-\lambda P_{i,K-1}^*(s) + (s+\mu)P_{iK}^*(s) &= \delta_{iK},
\end{aligned}$$

where δ_{ik} is the Kronecker delta. Defining the generating function

$$P_i^*(z;s) := \sum_{k=1}^{K} P_{ik}^*(s) z^k,$$

which is a polynomial of degree K in z, we get

$$P_i^*(z;s) = \frac{z^{i+1} + \lambda P_{iK}^*(s)(1-z) - \mu P_{i1}^*(s) z}{sz - (\mu - \lambda z)(1-z)}. \tag{12.1}$$

Let $\xi(s)$ and $\eta(s)$ be the solutions to the quadratic equation

$$\lambda z^2 - (s+\lambda+\mu)z + \mu = 0,$$

namely

$$\begin{aligned}
\xi(s) &:= \frac{s+\lambda+\mu - \sqrt{(s+\lambda+\mu)^2 - 4\lambda\mu}}{2\lambda}, \\
\eta(s) &:= \frac{s+\lambda+\mu + \sqrt{(s+\lambda+\mu)^2 - 4\lambda\mu}}{2\lambda}.
\end{aligned}$$

It can be shown that $|\xi(s)| < 1 < |\eta(s)|$ if $\Re(s) > 0$.

Then the numerator in (12.1) must be null at $z = \xi(s)$ and $z = \eta(s)$, which determines $P_{i1}^*(s)$. Hence we get [4, (19), p. 61]

$$B_i^*(s) = \rho^{-i} \frac{[\eta(s)]^{K-i}[\eta(s)-1] + [\xi(s)]^{K-i}[1-\xi(s)]}{[\eta(s)]^K[\eta(s)-1] + [\xi(s)]^K[1-\xi(s)]}, \tag{12.2}$$

where $\rho := \lambda/\mu$. The case of $i = 1$ is obtained by Sharma and Shobha [5] in different notation.

We now try to invert (12.2). To do so, let us introduce

$$x := \frac{s+\lambda+\mu}{2\sqrt{\lambda\mu}} = \begin{cases} \cos\theta, & -1 \le x \le 1,\ 0 \le \theta \le \pi \\ \cosh\tau, & x \ge 1,\ \tau \ge 0 \end{cases} \tag{12.3}$$

and

$$g_K(x) := \rho^{K/2} \frac{[\eta(s)]^K[\eta(s)-1] + [\xi(s)]^K[1-\xi(s)]}{\eta(s)-\xi(s)}$$
$$= \begin{cases} \dfrac{\sin(K+1)\theta - \sqrt{\rho}\,\sin K\theta}{\sin\theta}, & -1 \le x \le 1,\ 0 \le \theta \le \pi \\ \dfrac{\sinh(K+1)\tau - \sqrt{\rho}\,\sinh K\tau}{\sinh\tau}, & x \ge 1,\ \tau \ge 0. \end{cases}$$

Note that the Chebyshev polynomial of the second kind is defined by

$$U_K(x) := \begin{cases} \dfrac{\sin(K+1)\theta}{\sin\theta} = \displaystyle\sum_{j=0}^{\lfloor K/2\rfloor} (-1)^j \binom{K-j}{j} (2\cos\theta)^{K-2j}, & 0 \le \theta \le \pi \\ \dfrac{\sinh(K+1)\tau}{\sinh\tau} = \displaystyle\sum_{j=0}^{\lfloor K/2\rfloor} (-1)^j \binom{K-j}{j} (2\cosh\tau)^{K-2j}, & \tau \ge 0 \end{cases}$$
$$= \sum_{j=0}^{\lfloor K/2\rfloor} (-1)^j \binom{K-j}{j} (2x)^{K-2j}, \qquad x \ge -1,$$

as a polynomial of degree K, where $\lfloor x \rfloor$ denotes the largest integer not exceeding x. For example,

$$U_0(x) = 1, \quad U_1(x) = 2, \quad U_2(x) = 4x^2 - 1, \quad U_3(x) = 8x^3 - 4x,$$
$$U_4(x) = 16x^4 - 12x^2 + 1, \quad U_5(x) = 32x^5 - 32x^3 + 6x,$$
$$U_6(x) = 64x^6 - 80x^4 + 24x^2 - 1, \quad U_7(x) = 128x^7 - 192x^5 + 80x^3 - 8x,$$
$$U_8(x) = 256x^8 - 448x^6 + 240x^4 - 40x^2 + 1.$$

We then have

$$g_K(x) = U_K(x) - \sqrt{\rho}\, U_{K-1}(x), \qquad x \ge -1.$$

Hence we get

$$B_i^*(s) = \begin{cases} \rho^{-i/2} \dfrac{\sin(K-i+1)\theta - \sqrt{\rho}\,\sin(K-i)\theta}{\sin(K+1)\theta - \sqrt{\rho}\,\sin K\theta}, & 0 \le \theta \le \pi \\ \rho^{-i/2} \dfrac{\sinh(K-i+1)\tau - \sqrt{\rho}\,\sinh(K-i)\tau}{\sinh(K+1)\tau - \sqrt{\rho}\,\sinh K\tau}, & \tau \ge 0 \end{cases}$$
$$= \rho^{-i/2} \frac{g_{K-i}(x)}{g_K(x)}, \qquad x \ge -1.$$

Note that $B_i^*(0) = 1$, because $g_K(x)|_{s=0} = \rho^{-K/2}$.

Sharma and Shobha [5] argue that $g_K(x)$ has K distinct real zeros. More specifically, by utilizing the factorization

$$U_K(x) = 2^K \prod_{k=1}^{K} \left[x - \cos\left(\frac{k\pi}{K+1} \right) \right],$$

we see that

$$g_K\left(\cos\frac{j\pi}{K}\right) = U_K\left(\cos\frac{j\pi}{K}\right) = 2^K \prod_{k=1}^{K} \left[\cos\frac{j\pi}{K} - \cos\left(\frac{k\pi}{K+1} \right) \right]$$
$$= \begin{cases} < 0 & \text{if } j \text{ is odd} \\ > 0 & \text{if } j \text{ is even} \end{cases}, \qquad j = 1, 2, \ldots, K.$$

Therefore, $g_K(x)$ has $K-1$ distinct real zeros between $x = \cos(\pi/K)$ and $x = \cos\pi = -1$. Because

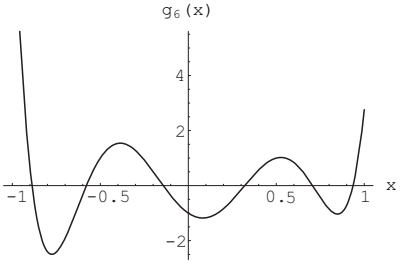
$$g_K(1) = K + 1 - \sqrt{\rho}\, K,$$

it follows that $g_K(x)$ has another real zero between $x = \cos 0 = 1$ and $x = \cos(\pi/K)$ if $\rho < ((K+1)/K)^2$. If $\rho > ((K+1)/K)^2$, then $g_K(1) < 0$ and $g(\infty) = \infty$. Thus there is another real zero at $x = \cosh\tau > 1$, which is uniquely determined by the equation

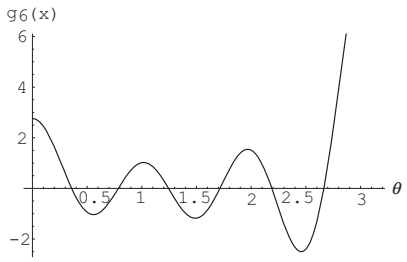
$$\sinh(K+1)\tau - \sqrt{\rho}\, \sinh K\tau = 0, \qquad \tau > 0. \tag{12.4}$$

Figure 12.1 shows $g_6(x)$ with $\rho = 0.5$ as functions in x and in θ. Figure 12.2 shows $g_6(x)$ with $\rho = 2$ as functions in x, in θ, and in τ.

Let the K distinct zeros of $g_K(x)$ be $\{\alpha_j, 1 \le j \le K\}$. From the partial fraction expansion



function in x with zeros at
-0.888418, -0.579196, -0.143028
0.320606, 0.707107, and 0.936483



function in θ with zeros at
0.358333, 0.785398, 1.24443
1.71432, 2.18854, and 2.66468

Fig. 12.1 $g_6(x)$ with $\rho = 0.5$.

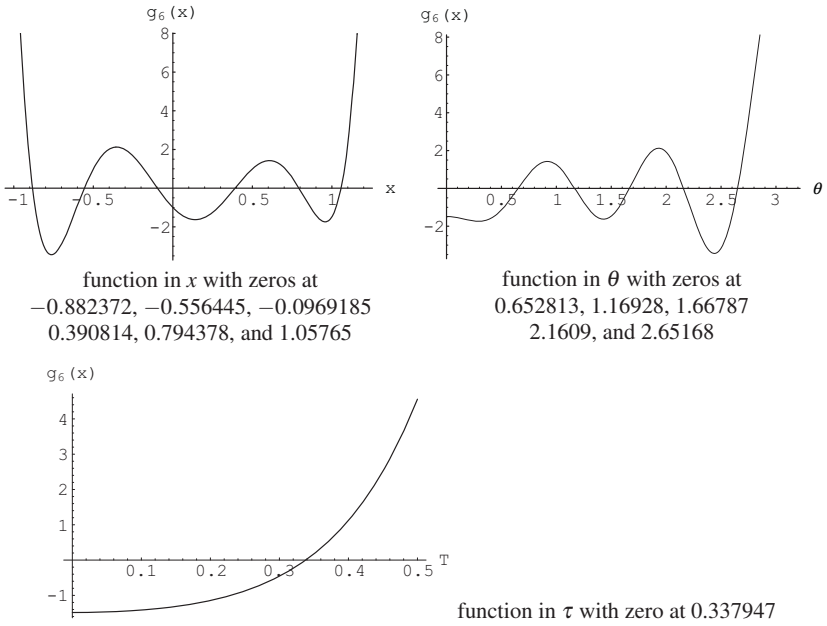


Fig. 12.2 $g_6(x)$ with $\rho = 2$.

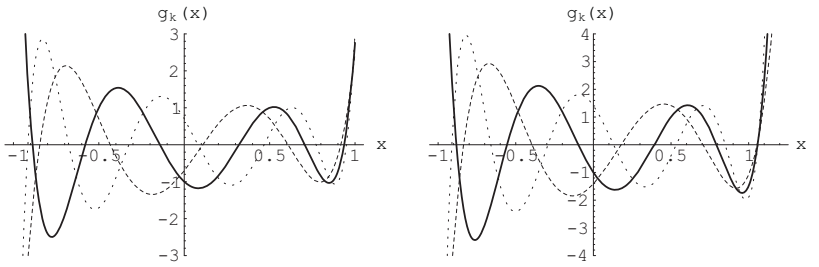
$$B_i^*(s) = \rho^{-i/2}\frac{g_{K-i}(x)}{g_K(x)} = \rho^{-i/2}\sum_{j=1}^{K}\frac{g_{K-i}(\alpha_j)}{g_K'(\alpha_j)(x-\alpha_j)}$$
$$= 2\sqrt{\lambda\mu}\,\rho^{-i/2}\sum_{j=1}^{K}\frac{g_{K-i}(\alpha_j)}{g_K'(\alpha_j)(s+\lambda+\mu-2\sqrt{\lambda\mu}\,\alpha_j)},$$

we obtain the pdf

$$b_i(t) = 2\sqrt{\lambda\mu}\,\rho^{-i/2}e^{-(\lambda+\mu)t}\sum_{j=1}^{K}\frac{g_{K-i}(\alpha_j)}{g_K'(\alpha_j)}e^{2\sqrt{\lambda\mu}\,\alpha_j t}, \qquad t \geq 0.$$

Sharma and Shobha [5] note that the $K-1$ zeros of $g_{K-1}(x)$ interlace the K zeros of $g_K(x)$. In fact, such an interlacing property is common in the transient analysis of finite-state birth-and-death processes [10]. Therefore, $g_{K-1}(x)$ and $g_K'(x)$ have the same sign at the zeros of $g_K(x)$. Figure 12.3 illustrates the situation with $g_5(x)$, $g_6(x)$, and $g_7(x)$. Furthermore, for $-1 < \alpha_j < 1$ we have

$$\lambda+\mu-2\sqrt{\lambda\mu}\,\alpha_j > \lambda+\mu-2\sqrt{\lambda\mu} = \left(\sqrt{\lambda}\,-\sqrt{\mu}\,\right)^2 \geq 0.$$



$\rho = 0.5$ $\rho = 2$

Fig. 12.3 $g_5(x)$ (dashed), $g_6(x)$ (solid), and $g_7(x)$ (dot-dashed).

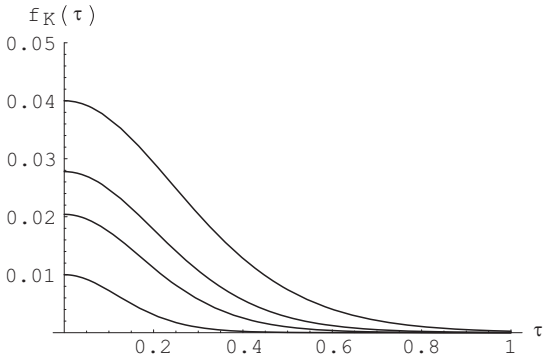


Fig. 12.4 $f_K(\tau) = ((\sinh^2(K+1)\tau)/(\sinh^2 K\tau)) + 1 - ((2\sinh(K+1)\tau\cosh\tau)/(\sinh K\tau))$ for $K = 5, 6, 7, 10$ (from above).

For $\alpha_j = \cosh\tau > 1$, where τ is determined by (12.4), we can show that

$$\lambda + \mu - 2\sqrt{\lambda\mu}\,\cosh\tau = \mu\left(\rho + 1 - 2\sqrt{\rho}\,\cosh\tau\right)$$
$$= \mu\left[\frac{\sinh^2(K+1)\tau}{\sinh^2 K\tau} + 1 - \frac{2\sinh(K+1)\tau\cosh\tau}{\sinh K\tau}\right] > 0, \qquad \tau > 0$$

for every $K \geq 1$; see Fig. 12.4. Hence, the pdf $b_1(t)$ of the busy period is a weighted sum of negative exponential distributions.

For the limit $K \to \infty$, because $|\xi(s)/\eta(s)| < 1$ it follows from (12.2) that

$$\lim_{K\to\infty} B_i^*(s) = [\xi(s)]^i.$$

This is inverted to

$$\lim_{K\to\infty} b_i(t) = i\rho^{-i/2}\frac{e^{-(\lambda+\mu)t}}{t} I_i\left(2\sqrt{\lambda\mu}\,t\right), \qquad t \geq 0,$$

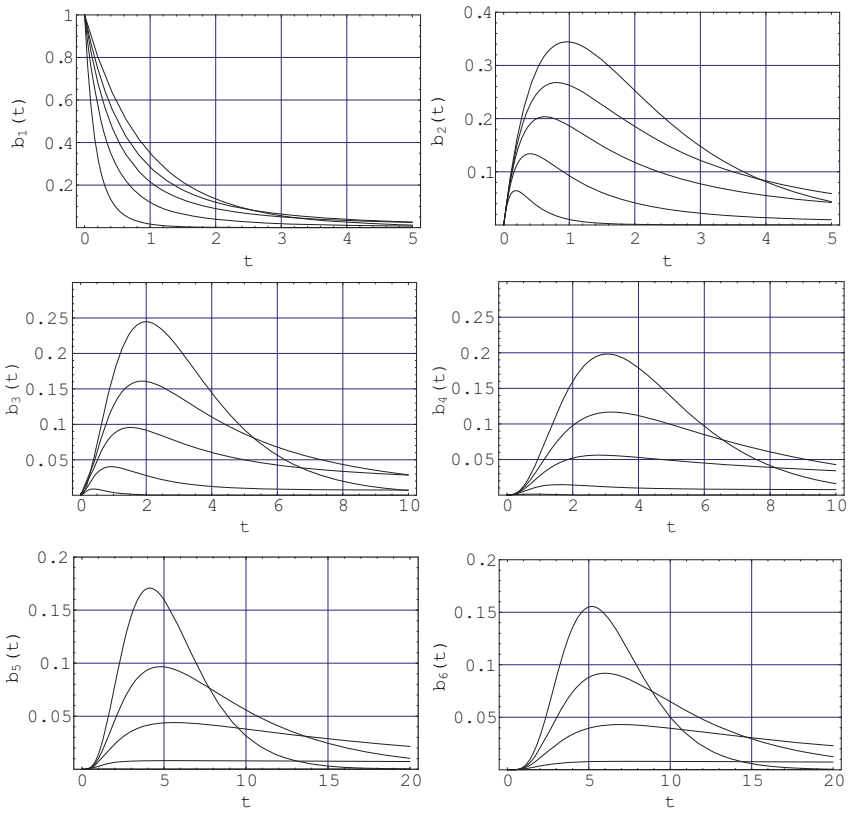
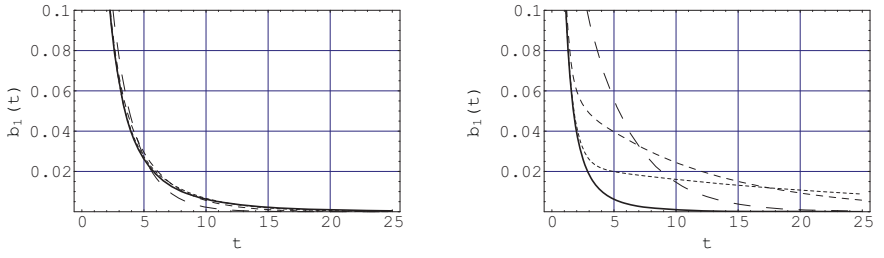


Fig. 12.5 $b_i(t)$ for $K = 6$ with $\lambda = 0.1, 0.5, 1, 2, 5$ (from above), and $\mu = 1$.



$\rho = 0.5$ $\rho = 2$

Fig. 12.6 $b_1(t)$ for $K = 2$ (rough dotted), 3 (medium dotted), 4 (fine dotted), and ∞ (solid) for $\mu = 1$.

where $I_i(x)$ is the modified Bessel function of the first kind of order i. This agrees with the result for an M/M/1 queue.

Figure 12.5 plots $b_i(t)$ for the M/M/1/6 queue with $\lambda = 0.1, 0.5, 1, 2, 5$, and $\mu = 1$ $(1 \le i \le 6)$. Figure 12.6 shows $b_1(t)$ for $K = 2, 3, 4$, and ∞ for $\mu = 1$, where we observe that $b_1(t)$ differs significantly for different values of K when ρ is large.

Moments of the length of a busy period are calculated directly from (12.2). The mean is given by

$$\mathrm{E}[\mathscr{B}_i] = \begin{cases} \dfrac{i}{\mu(1-\rho)} - \dfrac{\rho^{K-i+1}(1-\rho^i)}{\mu(1-\rho)^2}, & \rho \neq 1 \\ \dfrac{i(2K-i+1)}{2\mu}, & \rho = 1 \end{cases}.$$

Note that

$$\mathrm{E}[\mathscr{B}_1] = \frac{1-\rho^K}{\mu(1-\rho)},$$

which can also be derived from the renewal relationship

$$\lim_{t\to\infty} P\{N(t)=0\} = \frac{1-\rho}{1-\rho^{K+1}} = \frac{1/\lambda}{\mathrm{E}[\mathscr{B}_1]+1/\lambda}.$$

The second moment is given by

$$\mathrm{E}[(\mathscr{B}_i)^2] = \frac{i^2}{\mu^2(1-\rho)^2} + \frac{i(1+\rho+2\rho^{K-i+1}+2\rho^{K+1})}{\mu^2(1-\rho)^3} - \frac{2\rho^{K-i+1}(1-\rho^i)[2+2K(1-\rho)+\rho^{K+1}]}{\mu^2(1-\rho)^4}$$

for $\rho \neq 1$. For $\rho = 1$, we have

$$\mathrm{E}[(\mathscr{B}_i)^2] = \frac{i[(i+1)(i-1)(i-2)+4K(2K^2+3K+2-i^2)]}{12\mu^2}.$$

In particular

$$\mathrm{E}[(\mathscr{B}_1)^2] = \begin{cases} \dfrac{2(1-\rho^K)(1+\rho^{K+1})}{\mu^2(1-\rho)^3} - \dfrac{4K\rho^K}{\mu^2(1-\rho)^2}, & \rho \neq 1 \\ \dfrac{K(2K+1)(K+1)}{3\mu^2}, & \rho = 1 \end{cases}$$

as given in [5].

Our solution is based on the roots of the algebraic equation $g_K(x) = 0$, where $g_K(x)$ is a polynomial of order K. Therefore, it can be solved algebraically only for $K = 1, 2, 3$, and 4 in general (Abel's theorem).

For example, we first consider the case $K = 2$. The polynomial

$$g_2(x) = 4x^2 - 2\sqrt{\rho}\, x - 1$$

has two zeros $(\sqrt{\rho} \pm \sqrt{\rho+4}\,)/4$. Thus we get

$$B_1^*(s) = \frac{\rho^{-1/2}}{\sqrt{\rho+4}} \left(\frac{\sqrt{\rho+4} - \sqrt{\rho}}{4x - \sqrt{\rho} - \sqrt{\rho+4}} + \frac{\sqrt{\rho+4} + \sqrt{\rho}}{4x - \sqrt{\rho} + \sqrt{\rho+4}} \right)$$

$$= \frac{\mu\left(1 - \dfrac{\lambda}{\sqrt{\lambda^2+4\lambda\mu}}\right)}{2s+\lambda+2\mu-\sqrt{\lambda^2+4\lambda\mu}} + \frac{\mu\left(1 + \dfrac{\lambda}{\sqrt{\lambda^2+4\lambda\mu}}\right)}{2s+\lambda+2\mu+\sqrt{\lambda^2+4\lambda\mu}}$$

and

$$B_2^*(s) = \frac{2}{\rho\sqrt{\rho+4}} \left(\frac{1}{4x - \sqrt{\rho} - \sqrt{\rho+4}} - \frac{1}{4x - \sqrt{\rho} + \sqrt{\rho+4}} \right)$$

$$= \frac{2\mu^2}{\sqrt{\lambda^2+4\lambda\mu}} \left(\frac{1}{2s+\lambda+2\mu-\sqrt{\lambda^2+4\lambda\mu}} - \frac{1}{2s+\lambda+2\mu+\sqrt{\lambda^2+4\lambda\mu}} \right).$$

The corresponding density functions are given by

$$b_1(t) = \frac{\mu}{2}\left(1 - \frac{\lambda}{\sqrt{\lambda^2+4\lambda\mu}}\right) \exp\left(-\frac{\lambda+2\mu-\sqrt{\lambda^2+4\lambda\mu}}{2}t\right)$$

$$+ \frac{\mu}{2}\left(1 + \frac{\lambda}{\sqrt{\lambda^2+4\lambda\mu}}\right) \exp\left(-\frac{\lambda+2\mu+\sqrt{\lambda^2+4\lambda\mu}}{2}t\right)$$

and

$$b_2(t) = \frac{\mu^2}{\sqrt{\lambda^2+4\lambda\mu}} \left[\exp\left(-\frac{\lambda+2\mu-\sqrt{\lambda^2+4\lambda\mu}}{2}t\right) - \exp\left(-\frac{\lambda+2\mu+\sqrt{\lambda^2+4\lambda\mu}}{2}t\right) \right].$$

12.3 First Passage Time to the System Capacity

In Problem 17 on page 129 of Saaty [9], the time elapsing before the queue grows from size i to the system capacity K at which point the operation stops is studied $(0 \le i \le K-1)$. This is the first passage time from state i to state K in the birth-and-death process with the reflecting barrier at state 0, where the state k means that there are k customers present in the system. This is symmetric to the busy period studied above; just exchange the arrival and service rates and exchange i and $K-i$. The Laplace transform $\hat{B}_i^*(s)$ of the pdf $\hat{b}_i(t)$ for the above-mentioned first passage time $\hat{\mathscr{B}}_i$ is given in [9], but no explicit inversion is shown there.

We only present the results here. The Laplace transform $\hat{B}_i^*(s)$ is given by

$$\hat{B}_i^*(s) = \frac{\rho\{[\eta(s)]^{i+1} - [\xi(s)]^{i+1}\} - \{[\eta(s)]^i - [\xi(s)]^i\}}{\rho\{[\eta(s)]^{K+1} - [\xi(s)]^{K+1}\} - \{[\eta(s)]^K - [\xi(s)]^K\}}.$$

Using x, θ, and τ defined in (12.3), this can be written as

$$\hat{B}_i^*(s) = \begin{cases} \rho^{(K-i)/2} \dfrac{\sqrt{\rho}\,\sin(i+1)\theta - \sin i\theta}{\sqrt{\rho}\,\sin(K+1)\theta - \sin K\theta}, & 0 \le \theta \le \pi \\ \rho^{(K-i)/2} \dfrac{\sqrt{\rho}\,\sinh(i+1)\tau - \sinh i\tau}{\sqrt{\rho}\,\sinh(K+1)\tau - \sinh K\tau}, & \tau \ge 0 \end{cases}$$

$$= \rho^{(K-i)/2} \frac{\hat{g}_i(x)}{\hat{g}_K(x)}, \qquad x \ge -1,$$

where

$$\hat{g}_K(x) := \rho^{(K-1)/2} \frac{\rho\{[\eta(s)]^{K+1} - [\xi(s)]^{K+1}\} - \{[\eta(s)]^K - [\xi(s)]^K\}}{\eta(s) - \xi(s)}$$

$$= \begin{cases} \dfrac{\sqrt{\rho}\,\sin(K+1)\theta - \sin K\theta}{\sin\theta}, & -1 \le x \le 1,\ 0 \le \theta \le \pi \\ \dfrac{\sqrt{\rho}\,\sinh(K+1)\tau - \sinh K\tau}{\sinh\tau}, & x \ge 1,\ \tau \ge 0 \end{cases}$$

$$= \sqrt{\rho}\,U_K(x) - U_{K-1}(x), \qquad x \ge -1.$$

We note that $\hat{g}_K(x)$ has K distinct real zeros in $-1 < x < 1$ if $\rho > (K-(K+1))^2$. If $\rho < (K-(K+1))^2$, $\hat{g}_K(x)$ has $K-1$ distinct real zeros in $-1 < x < 1$ and another real zero in $x > 1$.

If K zeros of $\hat{g}_K(x)$ are denoted by $\{\hat{\alpha}_j, 1 \le j \le K\}$, the pdf for $\hat{\mathscr{B}}_i$ is given by

$$\hat{b}_i(t) = 2\sqrt{\lambda\mu}\,\rho^{(K-i)/2} e^{-(\lambda+\mu)t} \sum_{j=1}^{K} \frac{\hat{g}_i(\hat{\alpha}_j)}{\hat{g}_K'(\hat{\alpha}_j)} e^{2\sqrt{\lambda\mu}\,\hat{\alpha}_j t}, \qquad t \ge 0.$$

Because $\hat{g}_K(x)|_{s=0} = \rho^{(K+1)/2}$ it follows that $\hat{B}_i^*(0) = 1$. The mean first passage time is given by [9, p. 129]

$$\mathrm{E}[\hat{\mathscr{B}}_i] = \begin{cases} \dfrac{K-i}{\mu(\rho-1)} - \dfrac{\rho^{-i} - \rho^{-K}}{\mu(\rho-1)^2}, & \rho \ne 1 \\ \dfrac{(K-i)(K+i+1)}{2\mu}, & \rho = 1 \end{cases}.$$

The second moment is given by

$$\mathrm{E}[(\hat{\mathscr{B}}_i)^2] = \frac{(K-i)^2}{\mu^2(\rho-1)^2} + \frac{(K-i)(1+\rho+2\rho^{-i}+2\rho^{-K})}{\mu^2(\rho-1)^3} - \frac{2\rho^{-i}(1-\rho^{-(K-i)})[2\rho+2K(\rho-1)+\rho^{-K}]}{\mu^2(\rho-1)^4}$$

for $\rho \neq 1$. For $\rho = 1$, we have

$$\mathrm{E}[(\hat{\mathscr{B}}_i)^2] = \frac{(K-i)\{(K-i+1)(K-i-1)(K-i-2)+4K[K^2+(3+2i)K+2-i^2]\}}{12\mu^2}.$$

12.4 Regeneration Cycle

We may also consider the time interval $\bar{\mathscr{B}}_i$ between two successive instants at which the system enters state i. This is the regeneration cycle of state i in the birth-and-death process. Let $\bar{B}_i^*(s)$ be the Laplace transform of the pdf for $\bar{\mathscr{B}}_i$.

In order to find $\bar{B}_i^*(s)$ for $1 \le i \le K-1$, we note that the system started with state i goes to state $i+1$ with probability $\lambda/(\lambda+\mu)$ in an exponentially distributed time with mean $1/(\lambda+\mu)$. The state then behaves as the first passage from state 1 to state 0 in an M/M/1/(K− i) queue. On the other hand, the system started with state i goes to state $i-1$ with probability $\mu/(\lambda+\mu)$ in an exponentially distributed time with mean $1/(\lambda+\mu)$. The state then behaves as the first passage from state $i-1$ to state i in an M/M/1/i queue. Such consideration leads to the following results.

For $1 \le i \le K-1$, we have

$$\begin{aligned}\bar{B}_i^*(s) &= \frac{\mu}{s+\lambda+\mu}\left\{\frac{[\eta(s)]^{K-i-1}[\eta(s)-1]+[\xi(s)]^{K-i-1}[1-\xi(s)]}{[\eta(s)]^{K-i}[\eta(s)-1]+[\xi(s)]^{K-i}[1-\xi(s)]}\right. \\ &\quad \left. + \frac{[\eta(s)]^{i}[1-\xi(s)]+[\xi(s)]^{i}[\eta(s)-1]}{[\eta(s)]^{i+1}[1-\xi(s)]+[\xi(s)]^{i+1}[\eta(s)-1]}\right\} \\ &= \frac{1}{2x}\left[\frac{g_{K-i-1}(x)}{g_{K-i}(x)} + \frac{\hat{g}_{i-1}(x)}{\hat{g}_i(x)}\right],\end{aligned}$$

where x is defined in (12.3), and $g_K(x)$ and $\hat{g}_K(x)$ are defined in the preceding sections. Therefore, we can obtain the pdf for $\bar{\mathscr{B}}_i$ by inverting $\bar{B}_i^*(s)$ similarly.

The formula for $\bar{B}_i^*(s)$ yields the mean

$$\mathrm{E}[\bar{\mathscr{B}}_i] = \begin{cases} \dfrac{\rho^{-i}(1-\rho^{K+1})}{(\lambda+\mu)(1-\rho)}, & \rho \neq 1 \\ \dfrac{K+1}{2\mu}, & \rho = 1 \end{cases}$$

and the second moment

$$\mathrm{E}[(\bar{\mathscr{B}}_i)^2] = \frac{2\rho^{-i}(1-\rho^{K+1})}{(\lambda+\mu)^2(1-\rho)} - \frac{4\rho^{-i}[i+(K-i)\rho^{K+1}]}{(\lambda+\mu)\mu(1-\rho)^2} + \frac{2\rho^{-i+1}(1-\rho^K) - 2\rho^{-i}(1-\rho^{K+2}) + 2\rho^{-2i}(1-\rho^{2(K+1)})}{(\lambda+\mu)\mu(1-\rho)^3}$$

for $\rho \neq 1$. For $\rho = 1$, we have

$$\mathrm{E}[(\bar{\mathscr{B}}_i)^2] = \frac{(K+1)[2K^2+K+3-6i(K-i)]}{6\mu^2}.$$

If the state is started with the boundary ($i = 0$ or $i = K$), we have

$$\bar{B}_0^*(s) = \frac{\lambda}{s+\lambda} B_1^*(s) = \frac{\mu}{s+\lambda} \cdot \frac{[\eta(s)]^{K-1}[\eta(s)-1] + [\xi(s)]^{K-1}[1-\xi(s)]}{[\eta(s)]^K[\eta(s)-1] + [\xi(s)]^K[1-\xi(s)]}$$
$$= \frac{\sqrt{\rho}}{2\sqrt{\rho}\, x - 1} \cdot \frac{g_{K-1}(x)}{g_K(x)}$$

with

$$\mathrm{E}[\bar{\mathscr{B}}_0] = \frac{1-\rho^{K+1}}{\lambda(1-\rho)} \quad \text{for } \rho \neq 1, \qquad \mathrm{E}[\bar{\mathscr{B}}_0] = \frac{K+1}{\lambda} \quad \text{for } \rho = 1,$$

and

$$\bar{B}_K^*(s) = \frac{\mu}{s+\mu} \hat{B}_{K-1}^*(s) = \frac{\mu}{s+\mu} \cdot \frac{[\eta(s)]^K[1-\xi(s)] + [\xi(s)]^K[\eta(s)-1]}{[\eta(s)]^{K+1}[1-\xi(s)] + [\xi(s)]^{K+1}[\eta(s)-1]}$$
$$= \frac{1}{2x - \sqrt{\rho}} \cdot \frac{\hat{g}_{K-1}(x)}{\hat{g}_K(x)}$$

with

$$\mathrm{E}[\bar{\mathscr{B}}_K] = \frac{\rho - \rho^{-K}}{\mu(\rho-1)} \quad \text{for } \rho \neq 1, \qquad \mathrm{E}[\bar{\mathscr{B}}_K] = \frac{K+1}{\mu} \quad \text{for } \rho = 1.$$

12.5 Conclusions

In this chapter, we have derived closed-form explicit expressions for the pdf of the length of a busy period, the first passage time, and the regeneration cycle in an M/M/1/K queue. The pdf is expressed as a weighted sum of K negative exponential distributions with coefficients calculated from K distinct zeros of a polynomial that involves Chebyshev polynomials of the second kind.

In future work, we will study the busy period and first passage time in other queueing systems modeled by the birth-and-death process with finite state space.

References

1. H. G. Perros and T. Altiok, *Queueing Network With Blocking*. Amsterdam: Elsevier Science, 1989.
2. L. Takács, *Introduction to the Theory of Queues*. London: Oxford University Press, 1962.
3. A. I. Ismailov, The distribution of the busy period in a certain queueing model, *Doklady Akademii Nauk UzSSrR*, no. 5, pp. 3–4, 1970 (in Russian).
4. H. M. Srivastava and B. R. K. Kashyap, *Special Functions in Queuing Theory*. New York: Academic Press, 1982.
5. O. P. Sharma and B. Shobha, On the busy period of an M/M/1/N queue, *Journal of Combinatorics, Information and System Sciences*, vol. 11, nos. 2-4, pp. 110–114, 1986.
6. O. P. Sharma, *Markovian Queues*. Chichester, England: Ellis Horwood Limited, 1990.
7. W. Stadje, The busy periods of some queueing systems, *Stochastic Processes and Their Applications*, vol. 55, no. 1, pp. 159–167, 1995.
8. K. K. J. Kinateder and E.-Y. Lee, A new approach to the busy period of the M/M/1 queue, *Queueing Systems: Theory and Applications*, vol. 35, issues 1-4, pp. 105–115, 2000.
9. T. L. Saaty, *Elements of Queueing Theory with Applications*. New York: McGraw-Hill, 1961. Republished by New York: Dover, 1983.
10. W. Ledermann and G. E. H. Reuter, Spectral theory for the differential equations of simple birth and death processes, *Philosophical Transactions of the Royal Society of London*, vol. 246, pp. 321–369, 1954.

Chapter 13
Performance Analysis of ARQ Schemes in Self-Similar Traffic

Shunfu Jin, Wuyi Yue, and Naishuo Tian

Abstract In this chapter, we present a new method to analyze the performance of Automatic Repeat reQuest (ARQ) schemes in self-similar traffic. Taking into account the self-similar nature of a massive-scale wireless multimedia service, we build a batch arrival queueing model and suppose the batch size to be a random variable following a Pareto(c, α) distribution. Considering the delay in the setting up procedure of a data link, we introduce a setup strategy in this queueing model. Thus a batch arrival Geom^X/G/1 queueing system with setup is built in this chapter. By using a discrete-time embedded Markov chain, we analyze the stationary distribution of the queueing system and derive the Probability Generation Functions (P.G.Fs.) of the queueing length and the waiting time of the system. We give the formula for performance measures in terms of the response time of data frames, setup ratios, and offered loads for different ARQ schemes. Numerical results are given to evaluate the performance of the system and to show the influence of the self-similar degree and the delay of the setup procedure on the system performance.

13.1 Introduction

With the rapid development of wireless applications, support for Internet services with excellent reliability is becoming more and more important [1]. In general, error control schemes in communication systems can be classified into two categories: Forward Error Correction (FEC) and Automatic Repeat reQuest (ARQ) schemes [2].

S. Jin
College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China
e-mail: jsf@ysu.edu.cn

W. Yue
Department of Intelligence and Informatics, Konan University, Kobe 658-8501, Japan
e-mail: yue@konan-u.ac.jp

N. Tian
College of Science, Yanshan University, Qinhuangdao 066004, China
e-mail: tiannsh@ysu.edu.cn

W. Yue et al. (eds.), *Advances in Queueing Theory and Network Applications,*

In addition to FEC, ARQ schemes are in most cases used to ensure the transmission of packet data on higher layers, or are used as hybrid ARQ schemes on MAC/PHY layers.

It is a generally accepted view that discrete-time systems may be more complex to analyze than equivalent continuous-time systems. However, [3] has indicated that it would be more accurate and efficient to use discrete-time queueing models than continuous-time queueing models when analyzing and designing digital communication network systems.

The classical discrete-time queueing analyses have been presented in [3] and [4]. Extensive research of advanced ARQ schemes, as well as some performance analyses based on ARQ schemes have been conducted in [5]– [7]. In [5], an analysis of the ARQ feedback types was presented, but no algorithm to select the feedback was given. In [6], the ARQ mechanism were analyzed in the context of real-time flows of small packets. The key features and parameters of the ARQ mechanism were analyzed, and the ARQ block rearrangement, ARQ transmission window, and ARQ block size were researched in [7].

However, some simplifying assumptions considered in the above studies do not hold in practice. For example: self-similar behavior was neglected and the setting up procedure of a data link was omitted. This ignores both the influence of the self-similar degree as well as the delay of the setting up procedure on the system performance in such wireless networks.

In order to satisfy the demands of massive-scale wireless multimedia services and improve the performance of ARQ schemes, more accurate mathematical models that can faithfully capture the self-similar behavior of computer networks and the setting up procedure of a data link need to be constructed.

In this chapter, we avoid this unreal simplification to give a more constructed version, closer in nature to the actual system by considering the self-similar traffic shown in a service-oriented Internet [9]. Taking into account the delay in the setting up procedure of a data link, we build a batch arrival queueing model with a setup strategy. The results obtained in this chapter also include those in [8] for the system having arrivals of data frames. By using a discrete-time embedded Markov chain approach, we analyze the stationary distribution of the system, and present the stochastic decomposition of the queueing length and the waiting time. Based on numerical results, we evaluate the performance of ARQ schemes in terms of the response time of data frames, setup ratio, and the system's offered load. We also show the influence of the delay in the setup procedure and the self-similar degree on the system performance.

The chapter is organized as follows. In Sect. 13.2, the system model is described and some notation definitions are given. In Sect. 13.3, the stationary distribution of the system is derived. Correspondingly, performance measures for ARQ schemes are presented in Sect. 13.4. Numerical results are shown in Sect. 13.5 and conclusions are drawn in Sect. 13.6.

13.2 System Model and Notation

The system under analysis in this chapter consists of a pair of nodes, namely a transmitter and a receiver. When two adjacent nodes need to communicate with each other, a data link must be set up. We assume the time axis to be divided into slots of equal length and batch arrivals to follow a Bernoulli process. There are multiple data frames in a batch.

Self-similarity is the property we associate with one type of fractal, that is, an object whose appearance is unchanged regardless of the scale at which it is viewed [9]. A self-similar process may be constructed by superimposing many simple renewal reward processes, in which the rewards are restricted to the values 0 and 1, and the interrenewal times are heavy-tailed. The simplest heavy-tailed distribution is the Pareto(c,α) distribution [9]. We denote by Λ the number of data frames in a batch called batch size Λ (frames/slot), which is a random variable. The batch size follows a Pareto(c,α) distribution. When the transmission of all the data frames in the output buffer is finished, the data link should be released.

The system works as detailed below.

(1) When a batch arrives in the system, a setup period called "setup period U" is started, where the setup period U corresponds to a time period for setting up a new data link using a three-handshake signaling procedure.
(2) After the setup period U finishes, a busy period called "busy period Θ" begins. Here we define the busy period Θ to be a time period in which data frames are transmitted continuously until the transmitter buffer becomes empty.
(3) When there are no data frames in the output buffer of the transmitter to be transmitted, the data link is released and the system enters an idle period called "idle period I". A batch arriving during the idle period I makes the system enter a new setup period U again.

This process is repeated.

We define a transmission period B called "transmission period B" as being the time period taken to successfully transmit a data frame: that is, the time period from the instant for the first transmission of a data frame to the instant for the departure of the data frame from the transmitter buffer.

The transmission of a data frame only occurs after the correct reception of all data frames with a lower identifier, so we can assume that data frames in batches arriving in the buffer with an infinite capacity are transmitted using a common data link, one by one, in a First-Come First-Served (FCFS) discipline.

The setup period U and the transmission period B are independent and identical discrete-time random variables in slots, and are assumed to be generally distributed with probability distribution u_k and b_k, Probability Generation Functions (P.G.Fs.) $U(z)$ and $B(z)$ are as follows:

$$u_k = P\{U = k\}, \qquad k \geq 1, \qquad U(z) = \sum_{k=1}^{\infty} u_k z^k, \tag{13.1}$$

$$b_k = P\{B = k\}, \qquad k \geq 1, \qquad B(z) = \sum_{k=1}^{\infty} b_k z^k. \tag{13.2}$$

Let $\mathrm{E}[U]$ and $\mathrm{E}[B]$ be the averages of U and B in slots; we have that

$$\mathrm{E}[U] = \sum_{k=1}^{\infty} k u_k, \qquad \mathrm{E}[B] = \sum_{k=1}^{\infty} k b_k.$$

Let $\mathrm{E}[\Lambda]$ be the average of the batch size Λ. We can give the probability λ_k, the P.G.F. $\Lambda(z)$, and average $\mathrm{E}[\Lambda]$ of Λ as

$$\lambda_k = P\{\Lambda = k\}, \qquad k \geq 0, \qquad \Lambda(z) = \sum_{k=0}^{\infty} \lambda_k z^k, \qquad \mathrm{E}[\Lambda] = \sum_{k=0}^{\infty} k\lambda_k, \tag{13.3}$$

where λ_k is the probability that there are k data frames in a batch per slot. Specifically, $\lambda_0 = P\{\Lambda = 0\}$ is the probability that there is no batch ($\Lambda = 0$) arrival in a slot. From (13.1), we also know that the probability of no batch arrival during the transmission period B is $B(\lambda_0) = \lambda_0^B$. The ergodic condition is $\rho = \mathrm{E}[\Lambda]\mathrm{E}[B] < 1$, where ρ is called the offered load.

Let A_U and A_B be random variables representing the numbers of data frames arriving during U and B. We can then give the P.G.Fs. $A_U(z)$ and $A_B(z)$ of A_U and A_B as follows:

$$\begin{aligned} A_U(z) &= \sum_{k=1}^{\infty} u_k(\Lambda(z))^k = U(\Lambda(z)), \\ A_B(z) &= \sum_{k=1}^{\infty} b_k(\Lambda(z))^k = B(\Lambda(z)), \end{aligned} \tag{13.4}$$

where $U(\Lambda(z))$ and $B(\Lambda(z))$ are composed functions of $U(z)$, $B(z)$, and $\Lambda(z)$.

We also define $\Lambda(B(z))$ to be the P.G.F. of the transmission time of a batch in slots. $\Lambda(B(z))$ can be given as

$$\Lambda(B(z)) = \sum_{k=0}^{\infty} \lambda_k (B(z))^k. \tag{13.5}$$

13.3 Performance Analysis

We assume that data frame arrivals and departures occur only at the boundary of a slot. Let $Q_n = Q(\tau_n^+)$ be the number of data frames in the system immediately after the nth data frame departure. Then $\{Q_n, n \geq 1\}$ forms an embedded Markov chain. We define the state of the system by the number Q of data frames in the system at the embedded Markov points as follows:

$$Q_{n+1} = \begin{cases} Q_n - 1 + A_B^{(n+1)}, & Q_n \geq 1 \\ \Lambda^{'} + A_U + A_B^{(n+1)} - 1, & Q_n = 0, \end{cases} \tag{13.6}$$

where $A_B^{(n+1)}$ is the number of data frames arriving during the transmission time of the $(n+1)$th data frame, and $\Lambda^{'}$ denotes the number of data frames that arrive in a slot under the condition that there is at least one data frame arriving in that slot. Obviously, the P.G.F. $\Lambda^{'}(z)$ of $\Lambda^{'}$ can be given as

$$\Lambda^{'}(z) = \frac{\Lambda(z) - \lambda_0}{1 - \lambda_0}. \tag{13.7}$$

From (13.6), we can obtain the P.G.F. $Q(z)$ of Q as

$$\begin{aligned} Q(z) = {} & P\{Q \geq 1\} \mathrm{E}\left[z^{Q + A_B - 1} | Q \geq 1\right] \\ & + P\{Q = 0\} \mathrm{E}\left[z^{\Lambda^{'} + A_U + A_B^{(n+1)} - 1} | Q = 0\right], \end{aligned} \tag{13.8}$$

where $P\{Q = 0\}$ is the probability that there are no data frames to be transmitted in the system at the embedded Markov points, and $P\{Q \geq 1\}$ is the probability that there is at least one data frame to be transmitted in the system at the embedded Markov points.

Substituting (13.7) to (13.8), we can give that

$$Q(z) = P\{Q = 0\} \times \frac{B(\Lambda(z))}{B(\Lambda(z)) - z} \times \left(1 - \frac{\Lambda(z) - \lambda_0}{1 - \lambda_0} U(\Lambda(z))\right). \tag{13.9}$$

Using the normalization condition and the L'Hospital principle in (13.9), we have that

$$P\{Q = 0\} = \frac{(1 - \rho)(1 - \lambda_0)}{\mathrm{E}[\Lambda](1 + \mathrm{E}[U](1 - \lambda_0))}. \tag{13.10}$$

Substituting (13.10) to (13.9), then the P.G.F. $Q(z)$ of Q can be obtained as

$$Q(z) = \frac{(1 - \rho)(1 - \Lambda(z))B(\Lambda(z))}{\mathrm{E}[\Lambda](B(\Lambda(z)) - z)} \times \frac{1 - \lambda_0 - (\Lambda(z) - \lambda_0)U(\Lambda(z))}{1 - \Lambda(z)}. \tag{13.11}$$

Equation (13.11) implies that Q can be decomposed into two parts (i.e., $Q = Q_0 + Q_U$), where Q_0 corresponds to the number of data frames for the classical queue Geom^X/G/1 and Q_U is the number of data frames added by the setup scheme considered in this chapter.

The P.G.F. $Q_0(z)$ of Q_0 can be given as

$$Q_0(z) = \frac{(1 - \rho)(1 - \Lambda(z))B(\Lambda(z))}{\mathrm{E}[\Lambda](B(\Lambda(z)) - z)}$$

and the P.G.F. $Q_U(z)$ of Q_U can be given as

$$\begin{aligned} Q_U(z) &= \frac{1-\lambda_0-(\Lambda(z)-\lambda_0)U(\Lambda(z))}{1-\Lambda(z)} \\ &= \frac{1}{1+(1-\lambda_0)\mathrm{E}[U]} \times U(\Lambda(z)) + \frac{(1-\lambda_0)\mathrm{E}[U]}{1+(1-\lambda_0)\mathrm{E}[U]} \times \frac{1-U(\Lambda(z))}{\mathrm{E}[U](1-\Lambda(z))}. \end{aligned}$$

Obviously, $Q_U(z)$ equals the P.G.F. of the number of data frames arriving during the setup period U with the following probability as

$$\frac{1}{1+(1-\lambda_0)\mathrm{E}[U]}.$$

And $Q_U(z)$ equals the P.G.F. of the number of data frames arriving during the remaining setup period U with the following probability as

$$\frac{(1-\lambda_0)\mathrm{E}[U]}{1+(1-\lambda_0)\mathrm{E}[U]}.$$

Let $\mathrm{E}[X]$ and $X^{(2)}$ be the first and second factorial moments of a discrete-time random variable X by differentiating $X(z)$ with respect to z and evaluating the result at $z=1$ as follows:

$$\mathrm{E}[X] = \left.\frac{dX(z)}{dz}\right|_{z=1}, \qquad X^{(2)} = \left.\frac{d^2X(z)}{dz^2}\right|_{z=1}.$$

Based on the above definition, we can give the average $\mathrm{E}[Q]$ of Q from (13.11) as

$$\mathrm{E}[Q] = \rho + \frac{\Lambda^{(2)} + B^{(2)}\mathrm{E}^3[\Lambda]}{2\mathrm{E}[\Lambda](1-\rho)} + \frac{\mathrm{E}[\Lambda]\left((1-\lambda_0)U^{(2)} + 2\mathrm{E}[U]\right)}{2\left(1+\mathrm{E}[U](1-\lambda_0)\right)}, \tag{13.12}$$

where $U^{(2)}$, $B^{(2)}$, and $\Lambda^{(2)}$ are the second factorial moments of the setup period U, the transmission period B, and batch size Λ.

Now, we begin to analyze the waiting time of a data frame. We focus on an arbitrary data frame in the system called "tagged data frame M". We note that the waiting time W of the tagged data frame M can be divided into two parts as follows. One is the waiting time W_g of the batch to which the tagged data frame M belongs. The other is the total transmission time J of the data frames before the tagged data frame M in the same batch. W_g and J are independent random variables, so we have the P.G.F. $W(z)$ of the waiting time W of the tagged data frame M as follows:

$$W(z) = W_g(z)J(z), \tag{13.13}$$

where $W_g(z)$ and $J(z)$ are P.G.Fs. of W_g and J.

Applying the analysis of the single arrival Geom/G/1 queue model to the setup in [8], we have that

$$W_g(z) = \frac{(1-\rho)(1-z)}{\Lambda(B(z))-z} \times \frac{\mathrm{E}[\Lambda]+(1-z-\mathrm{E}[\Lambda])U(z)}{(1+\lambda \mathrm{E}[U])(1-z)}. \tag{13.14}$$

Referencing [3], with $\Lambda(B(z))$ given in (13.5), we have that

$$J(z) = \frac{1-\Lambda(B(z))}{\mathrm{E}[\Lambda](1-B(z))}. \tag{13.15}$$

Substituting (13.14) and (13.15) to (13.13), then the P.G.F. $W(z)$ and the average $\mathrm{E}[W]$ of W can be obtained as

$$W(z) = \frac{(1-\rho)(1-z)}{\Lambda(B(z))-z} \times \frac{1-\Lambda(B(z))}{\mathrm{E}[\Lambda](1-B(z))} \times \frac{\mathrm{E}[\Lambda]+(1-z-\mathrm{E}[\Lambda])U(z)}{(1+\lambda \mathrm{E}[U])(1-z)},$$

$$\mathrm{E}[W] = \frac{\Lambda^{(2)}\mathrm{E}^2[B]+\mathrm{E}[\Lambda]B^{(2)}}{2(1-\rho)} + \frac{\mathrm{E}[\Lambda]U^{(2)}+2\mathrm{E}[U]}{2(1+\mathrm{E}[\Lambda]\mathrm{E}[U])} + \frac{\Lambda^{(2)}\mathrm{E}[B]}{2\mathrm{E}[\Lambda]}. \tag{13.16}$$

Next, we define the busy cycle called "busy cycle R" as a time period from the instant in which a busy period Θ is completed to the instant in which the next busy period Θ ends. Obviously, a busy cycle R is composed of three parts: a setup period U, a busy period Θ, and an idle period I. Denoted by $\mathrm{E}[R]$, $\mathrm{E}[\Theta]$, and $\mathrm{E}[I]$ the averages of the busy cycle R, the busy period Θ, and the idle period I, respectively, we give that

$$\mathrm{E}[R] = \mathrm{E}[U]+\mathrm{E}[\Theta]+\mathrm{E}[I], \tag{13.17}$$

where $\mathrm{E}[U]$ is defined in (13.1), and $\mathrm{E}[\theta]$ and $\mathrm{E}[I]$ are given below.

Let Q_Θ be the number of data frames at the beginning of a busy period Θ. The P.G.F. $Q_\Theta(z)$ of Q_Θ is then given by

$$Q_\Theta(z) = \frac{\Lambda(z)-\lambda_0}{1-\lambda_0} U(\Lambda(z)). \tag{13.18}$$

Each data frame at the beginning of a busy period Θ will introduce a subbusy period θ. A subbusy period θ of a data frame is composed of the transmission period B of this data frame and the sum of the subbusy period θ incurred by all the data frames arriving during the transmission period B of this data frame. All the subbusy periods brought by the data frames at the beginning of the busy period combine to make a system busy period Θ, so we have that

$$\theta = B + \underbrace{\theta+\theta+\cdots+\theta}_{A_B}, \qquad \Theta = \underbrace{\theta+\theta+\cdots+\theta}_{Q_\Theta},$$

where A_B is the number of data frames arriving during the transmission period B presented in Sect. 13.2.

Considering the Bernoulli arrival process in this system, the P.G.F. $\theta(z)$ of θ can be obtained as follows:

$$\theta(z) = B(z(\Lambda(\theta(z)))),$$

which yields the average $\mathrm{E}[\theta]$ of θ as follows:

$$\mathrm{E}[\theta] = \frac{\mathrm{E}[B]}{1-\rho}. \tag{13.19}$$

From (13.18), we can obtain the P.G.F. $\Theta(z)$ of Θ as

$$\Theta(z) = Q_{\Theta}(z)|_{z=\theta(z)} = \frac{\Lambda(\theta(z)) - \lambda_0}{1-\lambda_0} U(\Lambda(\theta(z))). \tag{13.20}$$

Differentiating (13.20) with respect to z at $z = 1$ and using (13.19), the average $\mathrm{E}[\Theta]$ of Θ is then obtained as

$$\mathrm{E}[\Theta] = \frac{\mathrm{E}[\Lambda](1 + (1-\lambda_0)\mathrm{E}[U])}{(1-\lambda_0)} \times \frac{\mathrm{E}[B]}{(1-\rho)}. \tag{13.21}$$

The idle period I is a residual interarrival; due to the memoryless geometrically distributed interarrival time, we can obtain the average $\mathrm{E}[I]$ of I as

$$\mathrm{E}[I] = \frac{1}{1-\lambda_0}. \tag{13.22}$$

Substituting (13.21) and (13.22) to (13.17), the average $\mathrm{E}[R]$ of the busy cycle R can be given as

$$\begin{aligned} \mathrm{E}[R] &= \mathrm{E}[U] + \frac{\mathrm{E}[\Lambda](1 + (1-\lambda_0)\mathrm{E}[U])}{(1-\lambda_0)} \times \frac{\mathrm{E}[B]}{(1-\rho)} + \frac{1}{1-\lambda_0} \\ &= \frac{1 + (1-\lambda_0)\mathrm{E}[U]}{(1-\lambda_0)(1-\rho)}. \end{aligned} \tag{13.23}$$

13.4 Performance Analysis for Different Kinds of ARQ Schemes

Based on the analysis presented in Sect. 13.3, we can obtain the following performance measurements of the system.

13.4.1 Performance Measures

Response time T is defined as the total delay of a data frame. In our analysis, T is subdivided into two parts. One is the waiting time W of this data frame, which is

the time spent in the buffer before its transmission. The other is the corresponding transmission period B of this data frame. The average $\mathrm{E}[T]$ of T is given as follows:

$$\mathrm{E}[T] = \mathrm{E}[W] + \mathrm{E}[B]. \tag{13.24}$$

Substituting (13.16) to (13.24), we have that

$$\mathrm{E}[T] = \frac{\Lambda^{(2)}\mathrm{E}^2[B] + \mathrm{E}[\Lambda]B^{(2)}}{2(1-\rho)} + \frac{\mathrm{E}[\Lambda]U^{(2)} + 2\mathrm{E}[U]}{2(1+\mathrm{E}[\Lambda]\mathrm{E}[U])} + \frac{\Lambda^{(2)}\mathrm{E}[B]}{2\mathrm{E}[\Lambda]} + \mathrm{E}[B]. \tag{13.25}$$

The setup ratio γ is defined as the number of times that the system goes into the setup period U in a slot. There is a setup period U in the busy cycle R. The setup ratio γ can be given by

$$\gamma = \frac{1}{\mathrm{E}[R]}. \tag{13.26}$$

Substituting (13.23) to (13.26), we have that

$$\gamma = \frac{(1-\lambda_0)(1-\mathrm{E}[\Lambda]\mathrm{E}[B])}{1+(1-\lambda_0)\mathrm{E}[U]}. \tag{13.27}$$

We define the offered load ρ as the average number of data frames actually transmitted during a transmission period B, so the offered load ρ is given by

$$\rho = \mathrm{E}[\Lambda]\mathrm{E}[B]. \tag{13.28}$$

13.4.2 Performance Analysis for ARQ Schemes

In this subsection, we present the performance analysis for ARQ schemes. There are three kinds of basic ARQ schemes: Stop-and-Wait ARQ scheme, Go-Back-N ARQ scheme, and Selective-Repeat ARQ scheme. The principles and the differences among the different ARQ schemes are shown in Figs. 13.1–13.3.

To give the formulas for the performance measures for different kinds of ARQ schemes, the following assumptions and notions are introduced.

(1) The transmissions of the ACK frame and the NACK frame are error-free, and the lengths of the ACK frame and the NACK frame are omitted.

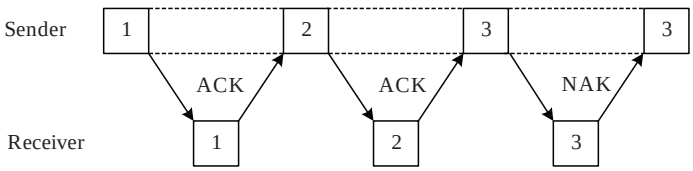


Fig. 13.1 The principle for a Stop-and-Wait ARQ scheme.

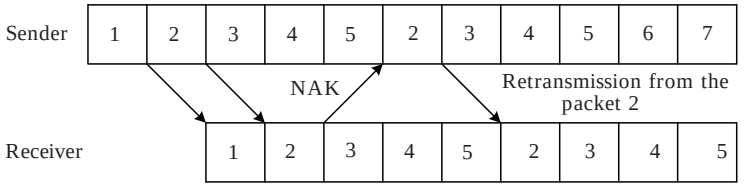


Fig. 13.2 The principle for a Go-Back-N ARQ scheme.

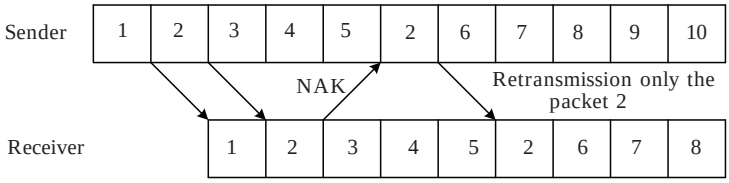


Fig. 13.3 The principle for a Selective-Repeat ARQ scheme.

(2) The rate of the transmission error is e $(0 \le e \le 1)$. Each data frame is correctly transmitted with probability $v = 1 - e$ $(0 \le v \le 1)$, and each data frame will be transmitted or retransmitted until correct reception is achieved.
(3) The round-trip time is assumed to be d slots as a system parameter.

Let N be the number of times of transmission needed for a data frame to be received correctly. Then the probability distribution and the P.G.F. $N(z)$ of N can be given as follows:

$$\begin{aligned} P\{N = n\} &= (1-v)^{n-1}v, \qquad n = 1, 2, \ldots, \\ N(z) &= \sum_{n=1}^{\infty} P\{N = n\}z^n = \frac{vz}{1-(1-v)z}. \end{aligned} \tag{13.29}$$

In the system with a Stop-and-Wait ARQ scheme, we denote by $B_{SW}(z)$, $\mathrm{E}[B_{SW}]$, and $B_{SW}^{(2)}$ the P.G.F. $B(z)$, the average $\mathrm{E}[B]$, and the second factorial moment $B^{(2)}$ of the transmission period B, respectively. From (13.25), we can give the average response time $\mathrm{E}[T]$ denoted by $\mathrm{E}[T_{SW}]$ for a Stop-and-Wait ARQ scheme as follows:

$$\begin{aligned} \mathrm{E}[T_{SW}] = {} & \frac{\Lambda^{(2)}\mathrm{E}^2[B_{SW}] + \mathrm{E}[\Lambda]B_{SW}^{(2)}}{2(1-\rho)} + \frac{\mathrm{E}[\Lambda]U^{(2)} + 2\mathrm{E}[U]}{2(1+\mathrm{E}[\Lambda]\mathrm{E}[U])} \\ & + \frac{\Lambda^{(2)}\mathrm{E}[B_{SW}]}{2\mathrm{E}[\Lambda]} + \mathrm{E}[B_{SW}]. \end{aligned} \tag{13.30}$$

Each transmission in a Stop-and-Wait ARQ scheme will take $1 + d$ slots, no matter whether the transmission is correct or not. So, $B_{SW}(z)$ [3], $\mathrm{E}[B_{SW}]$, and $B_{SW}^{(2)}$ are given as follows:

$$B_{SW}(z) = N(z^{1+d}) = \frac{vz^{1+d}}{1-(1-v)z^{1+d}}, \tag{13.31}$$

$$\mathrm{E}[B_{SW}] = \frac{1+d}{v}, \tag{13.32}$$

$$B_{SW}^{(2)} = \frac{(1+d)(vd+2(1-v)(1+d))}{v^2}. \tag{13.33}$$

Substituting (13.32) to (13.27) and (13.28), we can give the setup ratio γ_{SW} and the offered load ρ_{SW} as follows:

$$\begin{aligned}\gamma_{SW} &= \frac{(1-\lambda_0)(1-\mathrm{E}[\Lambda]\mathrm{E}[B_{SW}])}{1+(1-\lambda_0)\mathrm{E}[U]} \\ &= \frac{(1-\lambda_0)(v-\mathrm{E}[\Lambda](1+d))}{v(1+(1-\lambda_0)\mathrm{E}[U])}, \\ \rho_{SW} &= \mathrm{E}[\Lambda]\mathrm{E}[B_{SW}] = \frac{\mathrm{E}[\Lambda](1+d)}{v}.\end{aligned}$$

In the system with a Go-Back-N ARQ scheme, we denote by $B_{GBN}(z)$, $\mathrm{E}[B_{GBN}]$, and $B_{GBN}^{(2)}$ the P.G.F. $B(z)$, the average $\mathrm{E}[B]$, and the second factorial moment $B^{(2)}$ of the transmission period B, respectively. From (13.25), we can give the average response time $\mathrm{E}[T]$ denoted by $\mathrm{E}[T_{GBN}]$ for a Go-Back-N ARQ scheme as follows:

$$\begin{aligned}\mathrm{E}[T_{GBN}] = {} & \frac{\Lambda^{(2)}\mathrm{E}^2[B_{GBN}] + \mathrm{E}[\Lambda]B_{GBN}^{(2)}}{2(1-\rho)} + \frac{\mathrm{E}[\Lambda]U^{(2)} + 2\mathrm{E}[U]}{2(1+\mathrm{E}[\Lambda]\mathrm{E}[U])} \\ & + \frac{\Lambda^{(2)}\mathrm{E}[B_{GBN}]}{2\mathrm{E}[\Lambda]} + \mathrm{E}[B_{GBN}].\end{aligned} \tag{13.34}$$

In a Go-Back-N ARQ scheme, each error transmission occupies $1+d$ slots, and the last correct transmission takes one slot. So, $B_{GBN}(z)$ [3], $\mathrm{E}[B_{GBN}]$, and $B_{GBN}^{(2)}$ are given as follows:

$$B_{GBN}(z) = \frac{N(z^{1+d})}{z^d} = \frac{vz}{1-(1-v)z^{1+d}}, \tag{13.35}$$

$$\mathrm{E}[B_{GBN}] = \frac{1+(1-v)d}{v}, \tag{13.36}$$

$$B_{GBN}^{(2)} = \frac{(1-v)(1+d)(2+2d-vd)}{v^2}. \tag{13.37}$$

Substituting (13.36) to (13.27) and (13.28), we can give the setup ratio γ_{GBN} and the offered load ρ_{GBN} as

$$\gamma_{GBN} = \frac{(1-\lambda_0)(1-\mathrm{E}[\Lambda]\mathrm{E}[B_{GBN}])}{1+(1-\lambda_0)\mathrm{E}[U]} = \frac{(1-\lambda_0)(v-\mathrm{E}[\Lambda](1+(1-v)d))}{v(1+(1-\lambda_0)\mathrm{E}[U])},$$
$$\rho_{GBN} = \mathrm{E}[\Lambda]\mathrm{E}[B_{GBN}] = \frac{\mathrm{E}[\Lambda](1+(1-v)d)}{v}.$$

In the system with a Selective-Repeat ARQ scheme, we denote by $B_{SR}(z)$, $\mathrm{E}[B_{SR}]$, and $B_{SR}^{(2)}$ the P.G.F. $B(z)$, the average $\mathrm{E}[B]$, and the second factorial moment $B^{(2)}$ of the transmission period B, respectively. From (13.25), we can give the average response time $\mathrm{E}[T]$ denoted by $\mathrm{E}[T_{SW}]$ for a Stop-and-Wait ARQ scheme as follows:

$$\mathrm{E}[T_{SR}] = \frac{\Lambda^{(2)}\mathrm{E}^2[B_{SR}]+\mathrm{E}[\Lambda]B_{SR}^{(2)}}{2(1-\rho)} + \frac{\mathrm{E}[\Lambda]U^{(2)}+2\mathrm{E}[U]}{2(1+\mathrm{E}[\Lambda]\mathrm{E}[U])} + \frac{\Lambda^{(2)}\mathrm{E}[B_{SW}]}{2\mathrm{E}[\Lambda]} + \mathrm{E}[B_{SR}].$$

Each transmission in a Selective-Repeat ARQ scheme, no matter whether it is correct or not, takes, one slot. So, $B_{SR}(z)$, $\mathrm{E}[B_{SR}]$, and $B_{SR}^{(2)}$ are given as follows:

$$B_{SR}(z) = N(z) = \frac{vz}{1-(1-v)z}, \tag{13.38}$$
$$\mathrm{E}[B_{SR}] = \frac{1}{v}, \tag{13.39}$$
$$B_{SR}^{(2)} = \frac{2(1-v)}{v^2}. \tag{13.40}$$

Substituting (13.39) to (13.27) and (13.28), we can also give the setup ratio γ_{SR} and the offered load ρ_{SR} as follows:

$$\gamma_{SR} = \frac{(1-\lambda_0)(1-\mathrm{E}[\Lambda]\mathrm{E}[B_{SR}])}{1+(1-\lambda_0)\mathrm{E}[U]} = \frac{(1-\lambda_0)(v-\mathrm{E}[\Lambda])}{v(1+(1-\lambda_0)\mathrm{E}[U])},$$
$$\rho_{SR} = \mathrm{E}[\Lambda]\mathrm{E}[B_{SR}] = \frac{\mathrm{E}[\Lambda]}{v}.$$

13.5 Numerical Results

In line with prevalent wireless network applications, we let the transmission rate be 50 Mbps. To ensure that the latest conflict signal is sensed by the transmitter before a data frame is completely sent out, we assume the size of a data frame to be 1,250

bytes and the round-triptime to be 0.1 ms. The setup period U follows a geometrical distribution with an average value of 0.2 ms.

At the same time, taking into account the burst data shown in Internet traffic, we suppose the batch size Λ to be a Pareto(c,α) distribution with $\lambda_k = ck^{-(\alpha+1)}$, $k = 0,1,\ldots,$ where c is a normalization factor for $\sum_{k=1}^{\infty}\lambda_k = 1$, and the parameter α is related to the Hurst factor H by $H = (3-\alpha)/2, 0.5 < H < 1, 1 < \alpha < 2$. The smaller the result of α is, the more the burst is shown in Internet traffic. Especially, there is no self-similarity when $\alpha = 2$. Some research shows that the transmission mode of the browser shows self-similarity [9] with $\alpha = 1.16 - 1.5$ and the data of each signal source are self-similar [10] with $\alpha = 1.2$.

With these parameters, we show the setup ratio γ and offered load ρ as functions of the batch arrival rate $\lambda_g = 1 - \lambda_0$ (batches/slot) with the rate of the transmission error $e = 0.1$ under the burst degree of $\alpha = 1.2, 1.6, 2.0$, respectively. For different kinds of ARQ schemes in Figs. 13.4–13.9, where $\alpha = 2.0$ means that there is actually no self-similarity.

In Figs. 13.4–13.6, we show how the setup ratio γ changes with the batch arrival rate λ_g with the rate of the transmission error $e = 0.1$ and with the parameter of burst degree $\alpha = 1.2, 1.6, 2.0$ for different ARQ schemes. It should be noted that for all the burst degree parameters, the setup ratio γ experiences a two-stage trend. In the first stage, the setup ratio γ will increase along with the batch arrival rate λ_g. During this stage, the greater the batch arrival rate λ_g is, the higher the number of data frames arriving in the idle period I will be, and the greater the number of times needed for the setup procedure will be. In the second stage, the setup ratio γ will decrease with the incremental batch arrival rate λ_g. During this period, the greater the batch arrival rate λ_g is, the higher the number of data frames arriving in the busy period Θ will be, and these data frames can be transmitted directly without any setup procedure.

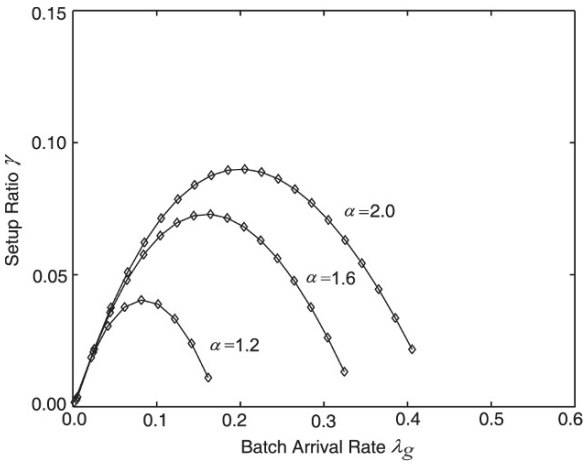


Fig. 13.4 Setup ratio γ for a Stop-and-Wait ARQ scheme.

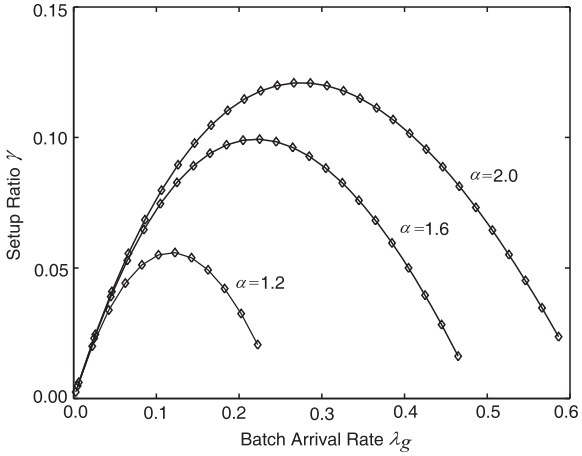


Fig. 13.5 Setup ratio γ for a Go-Back-N ARQ scheme.

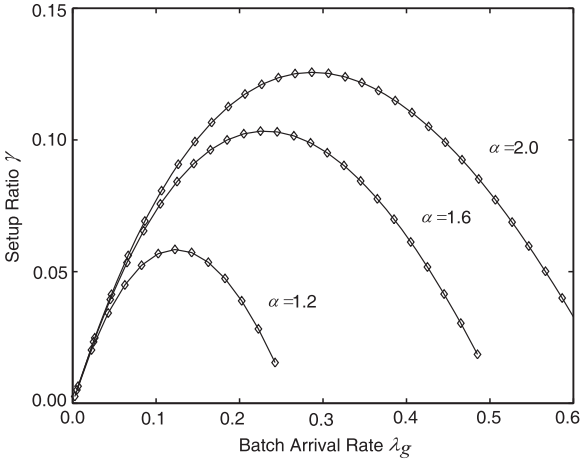


Fig. 13.6 Setup ratio γ for a Selective-Repeat ARQ scheme.

There is a maximal setup ratio γ for all the burst degree parameters, and it can also be observed that the larger the burst degree parameter α is, the greater the maximal setup ratio γ will be, and we can conclude that if we omitted any self-similar Internet traffic, the setup ratio γ would be overevaluated.

In Figs. 13.7–13.9, we compare the offered load ρ with the rate of the transmission error $e = 0.1$ versus batch arrival rate λ_g for the parameters of burst degree $\alpha = 1.2, 1.6, 2.0$ for different ARQ schemes. It can be found that with an increasing batch arrival rate λ_g, the offered load ρ increases also for all the ARQ schemes and all the parameters of burst degree. It should be noted that for the same batch arrival rate λ_g, the lower the parameter of burst degree α is, the larger the offered

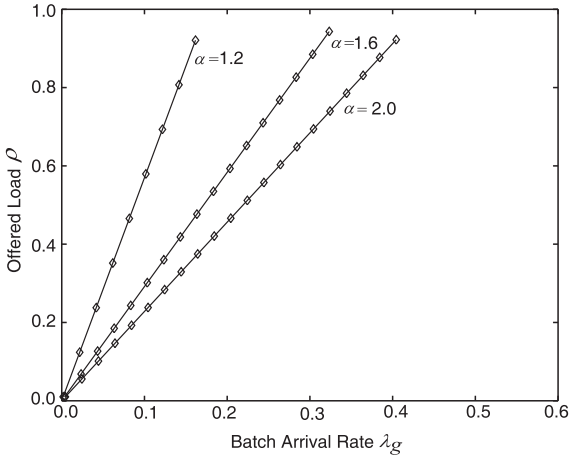


Fig. 13.7 Offered load ρ for a Stop-and-Wait ARQ scheme.

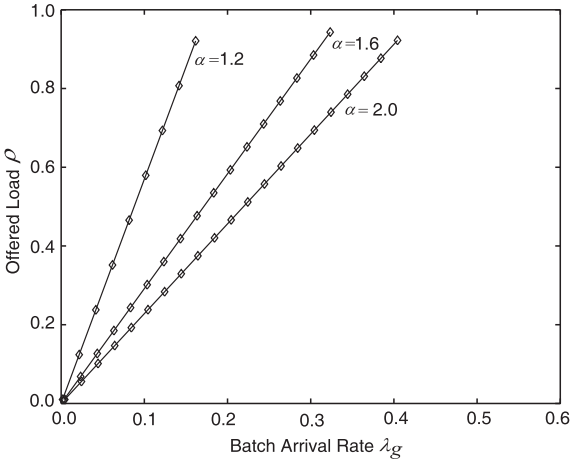


Fig. 13.8 Offered load ρ for a Go-Back-N ARQ scheme.

load ρ will be for all the ARQ schemes. Therefore, we can conclude that if the self-similarity is not considered, the offered load ρ would be underevaluated.

Due to the finite first factorial moment and the infinite second moment of a Pareto distributed stochastic variable, some other performance measures such as the average response time $\mathrm{E}[T]$ in (13.24) are difficult to calculate analytically. So we present the change trend of average response time $\mathrm{E}[T]$ by using the method of simulation.

There is no ready Pareto function in most simulation tools such as Matlab to be used, so we use an inverse function method to generate random number sequences following the Pareto distribution.

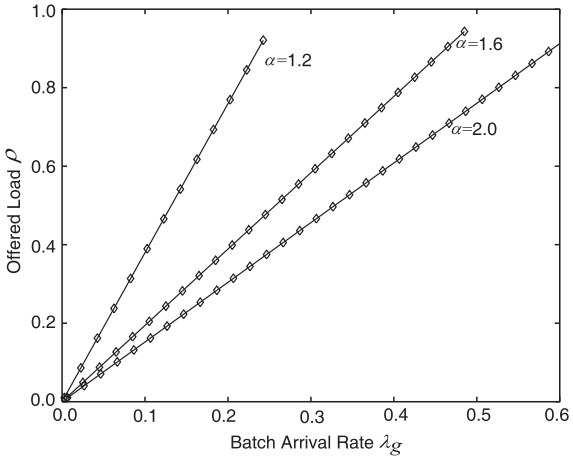


Fig. 13.9 Offered load ρ for a Selective-Repeat ARQ scheme.

Table 13.1 Response time $\mathrm{E}[T]$ of different ARQ schemes for various λ_g with $\alpha = 1.2$ and $e = 0.1$.

Batch Arrival Rate λ_g	0.001	0.02	0.04	0.06	0.08	0.10
Stop-and-Wait ARQ	2.96	39.17	41.30	517.66	1414.30	14087
Go-Back-N ARQ	5.22	77.13	113.14	731.88	4655.90	6130
Selective-Repeat ARQ	5.44	413.99	2671.40	3453.50	6645.60	24328

The general discrete distribution is characterized as follows:

$$\begin{aligned} p_k &= P\{X = k\}, \qquad k \geq 0, \\ F(m) &= \sum_{k=0}^{m} p_k, \qquad m \geq 0. \end{aligned} \tag{13.41}$$

By using a random numbers generation function, we generate random numbers of a $1 \times n$ vector named M whose elements are uniformly distributed in the interval (0,1). On the other hand, following the inverse function method, we introduce another $1 \times n$ vector named N whose elements are set by $N(i) = \min\{m : F(m) > M(i)\}$, where $F(m)$ is given in (13.41) and $m \geq 1$, $i \geq 1$. In this way, the data in the vector N will be Pareto distributed.

The change trend of average response time $\mathrm{E}[T]$ for different ARQ schemes when $\alpha = 1.2$ and $e = 0.1$ with various λ_g is presented in Table 13.1. The measurement of average response time $\mathrm{E}[T]$ behavior for different ARQ schemes when $\alpha = 1.2$ and $\lambda_g = 0.04$ with various error rates e is shown in Table 13.2.

From Tables 13.1 and 13.2, we can observe that with an increasing batch arrival rate λ_g or an increasing error rate e, the average response time $\mathrm{E}[T]$ increases also and tends to be infinite for all the ARQ schemes. This is because of the self-similarity shown in the size of the data frame batch, which is in fact the reason why network performance deteriorates in self-similar traffic.

Table 13.2 Response time E[T] of different ARQ schemes for various e with $\alpha = 1.2$ and $\lambda_g = 0.04$.

Error Rate e	0.02	0.06	0.10	0.14	0.18	0.22
Stop-and-Wait ARQ	17.065	26.081	41.30	64.139	2256.7	7957.9
Go-Back-N ARQ	15.268	63.604	113.14	173.91	686.78	778.46
Selective-Repeat ARQ	70.924	228.49	2671.40	2699.6	7813.5	9924.8

13.6 Conclusions

In this chapter, we presented a new method to analyze the performance of high-reliability Internet systems in self-similar traffic with ARQ schemes. Considering the self-similar nature widely shown in Internet traffic and the setting up procedure of a data link, we built a batch arrival Geom^X/G/1 queue model with a setup strategy. We analyzed the stationary distribution of the system, derived the Probability Generation Functions (P.G.Fs.) of the queueing length and the waiting time of the system. Correspondingly, we gave the formula for performance measures in terms of response time, setup ratio, and offered load for different kinds of ARQ schemes. We presented numerical results to evaluate and compared these performance measures, and to show the influence of the burst degree in self-similar traffic and the delay in the setup procedure on the system performance.

Acknowledgments This work was supported in part by MEXT.ORC (2004-2008) and GRANT-IN-AID FOR SCIENCE RESEARCH (No. 19500070), Japan and was supported in part by NSF (No. 10671170) and NSF (No. 60773100), China.

References

1. W. Yue and Y. Matsumoto, *Performance Analysis of Multi-Channel and Multi-Traffic on Wireless Communication Networks*. Boston: Kluwer Academic, 2002.
2. L. Badia, M. Rossi, and M. Zorzi, SR ARQ packet delay statistics on markov channels in the presence of variable arrival rate, *IEEE Transaction on Wireless Communication*, vol. 5, no. 7, pp. 1639–1644, 2006.
3. H. Takagi, *Queueing Analysis* (*Volume 3*: *Discrete-Time Systems*). Amsterdam: Elsevier Science, 1993.
4. N. Tian and G. Zhang, *Vacation Queueing Models-Theory and Applications*. New York: Springer-Verlag, 2006.
5. M. S. Kang and J. Jang, Performance evaluation of IEEE 802.16d ARQ algorithms with NS-2 simulator, in *Proc. IEEE Asia-Pacific Conference on Communications*, pp. 1–5, 2006.
6. S. Perera and H. Sirisena, Contention based negative feedback ARQ for VoIP services in IEEE 802.16 networks, in *Proc. 4th IEEE International Conference on Networks*, pp. 1–6, 2006.
7. T. Vitaliy, S. Alexander, M. Henrik, A. Olli, and H. Timo, Performance evaluation of the IEEE 802.16 ARQ mechanism, in *Proc. International Conference on Next Generation Teletraffic and Wired/Wireless Advanced Networking, Lecture Notes in Computer Science 4712*, pp. 148–161, 2007.

8. W. Yue and S. Jin, Performance analysis of digital wireless networks with ARQ schemes, in *Proc. International Conference on Computational Science, Part IV, Lecture Notes in Computer Science 4490*, pp. 352–359, 2007.
9. M. Crovella and A. Bestavros, Self-similarity in World Wide Web traffic: Evidence and possible causes, *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, 1997.
10. W. Willinger, M. Taqqu, R. Sherman, and D. Wilson, Self-similarity through high variability: Statistical analysis of ethernet LAN traffic at the source level, *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 71–86, 1997.

Chapter 14
Modeling of P2P File Sharing with a Level-Dependent QBD Process

Sophie Hautphenne, Kenji Leibnitz, and Marie-Ange Remiche

Abstract In this chapter we propose to analyze a peer-to-peer (P2P) file sharing system by means of a so-called level-dependent Quasi Birth-and-Death (QBD) process. We consider the dissemination of a single file consisting of different segments and include a model for the upload queue management mechanism with peers competing for bandwidth. By applying an efficient matrix-analytic algorithm we evaluate the performance of P2P file diffusion in terms of the corresponding extinction probability, that is, the probability that the sharing process ends.

14.1 Introduction

With the introduction of peer-to-peer (P2P) technology in networks for file sharing and content distribution, the volume of transported traffic has recently enormously increased. The nodes participating in the P2P network are called peers and form logical overlay structures on the application layer above the IP topology; see Fig. 14.1. One of the main advantages of using P2P networks for content distribution is their high scalability to a growing number of file requests, especially in the presence of flash crowd arrivals [1]. Unlike conventional client/server architectures, all peers act simultaneously as clients and servers, thus shifting the load from a single server to

S. Hautphenne
Département d'Informatique, Université Libre de Bruxelles, B-1050 Bruxelles, Belgium
e-mail: shautphe@ulb.ac.be

M.-A. Remiche
Faculté des Sciences Appliquées, Université Libre de Bruxelles, B-1050 Bruxelles, Belgium
e-mail: mremiche@ulb.ac.be

K. Leibnitz
Graduate School of Information Science and Technology, Osaka University, Osaka 565-0871, Japan
e-mail: leibnitz@ist.osaka-u.ac.jp

W. Yue et al. (eds.), *Advances in Queueing Theory and Network Applications,*

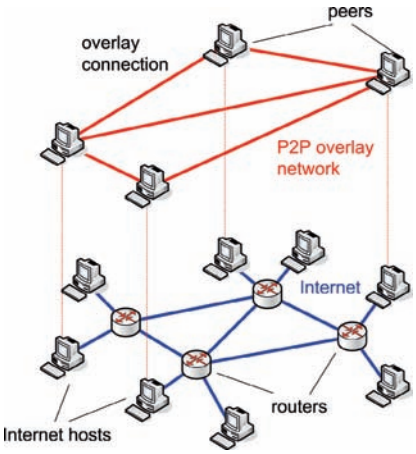


Fig. 14.1 A P2P network consists of peers forming a logical overlay network above the IP topology.

several peers sharing a specific file. Additionally, because the source of a file is no longer stored at a single location, the P2P network is more robust to failures.

However, there are also certain dangers in entirely relying on P2P networks for file distribution. Firstly, as the data are no longer kept at a single trusted source, each peer that hosts the file may modify the data willingly or unwillingly, thus causing the distribution of corrupt information. This is referred to as *poisoning* or *pollution* [2]. Secondly, the existence of a sharing peer in the network cannot be guaranteed due to *churn* (i.e., the process of peers entering and leaving the network). The sharing of files is controlled by the peers' behavior (willingness to share after downloading, patience, etc.) and they may arbitrarily join or leave the network at any instant [3]. If the peer, which has the last part of the file, leaves the network, this information is lost and other peers can no longer retrieve the data. For this reason, specific P2P architectures (e.g., Chord [4]) employ mechanisms to maintain a certain number of replicas of a file in the network.

In this chapter we study the probability that the diffusion of a file will eventually come to a halt in an unstructured P2P file sharing network, which we define as the *extinction* of the file. We extend our previous model in [5], where we used a Markovian Binary Tree (MBT) to model the file sharing network and we formulated an algorithm to compute the extinction probability. However, the previous model only considered the sharing of entire files. In this chapter, we extend the model to include the sharing of individual parts of the file to reflect a more accurate behavior. This is achieved by using a level-dependent Quasi Birth-and-Death (QBD) process. By adapting the logarithmic-reduction algorithm (see Latouche and Ramaswami [6]), we actually compute the probability that file diffusion ends due to the lack of peers sharing a part of the file.

This chapter is organized as follows. First, we briefly summarize some related work on modeling of P2P file sharing mechanisms for content distribution in Sect. 14.2. This is followed by the formulation of our basic assumptions on the

file sharing network in Sect. 14.3. Although we consider a P2P network that roughly resembles the eDonkey protocol, the model is general enough to be easily applied to other file sharing protocols as well. In Sect. 14.4 we formulate two analytical models corresponding to two different systems in which either the sharing process stops when the entire file is lost or when any of the segments is missing. Accordingly, we construct the corresponding level-dependent QBD process and we develop algorithms necessary to obtain the extinction probability in both settings. We provide some numerical results showing the impact of the system parameters on the performance of the system in Sect. 14.5. Finally, conclusions are drawn in Sect. 14.6.

14.2 Related Work

A growing number of studies can be found dealing with the modeling and performance evaluation of P2P file sharing networks. In this section we only highlight a few of them that we consider relevant to this chapter. Most studies on the evaluation of P2P systems as content distribution networks rely on measurements or simulations of existing P2P networks. For example, Saroiu et al. [7] conducted measurement studies of content delivery systems that were accessed by the University of Washington. The authors distinguish between traffic from P2P, WWW, and the Akamai content distribution network, and they found that the majority of volume is transported over P2P. In [8], a measurement-based traffic profile of the eDonkey network is provided and reveals that there is a strong distinction between download flows and nondownload streams. Similar studies exist for the Gnutella network [9] and BitTorrent [10], as well. Hoßfeld et al. [11] provide a simulation study of the eDonkey network and examine the file diffusion properties under constant and flash crowd arrivals.

An analytical model for performance evaluation of a generalized P2P system is given by Ge et al. [12]. On the other hand, other published work mostly considers specific existing applications. For example, Qiu and Srikant [13] used a fluid model for BitTorrent and investigate the performance in steady state. They studied the effectiveness of the incentive mechanism in BitTorrent and proved the existence of a Nash equilibrium. Rubenstein and Sahu [1] mathematically showed that unstructured P2P networks have good scalability and are well suited to cope with flash crowd arrivals. A fluid-diffusive P2P model from statistical physics is presented by Carofiglio et al. [14]. Both the user and the content dynamics are included, but this is only done on the file level and without pollution. All these studies show that by providing incentives to the peers for sharing a file, the diffusion properties are improved. Yang and de Veciana [15] investigated the service capacity of P2P networks by considering two models, one for the transient state with flash crowds and one in steady state.

Christin, Weigend and Chuang [2] measured content availability of popular P2P file sharing networks and used these measurement data for simulating different pollution and poisoning strategies. They show that only a small number of fake peers

can seriously affect the user's perception of content availability. In [16], a diffusion model for modeling eDonkey-like P2P networks is presented based on a model from mathematical biology. This model includes pollution and a patience threshold at which a peer aborts its download attempt and retries again later. It is shown that an evaluation of the diffusion process is not accurate enough when steady state is assumed or the model only considers the transmission of the complete file, especially in the presence of flash crowd arrivals. That model is extended in [17] to analytically compare the performance of P2P file sharing networks to that of client/server systems.

14.3 Peer-to-Peer File Sharing Model

Let us now define the assumptions we make on the P2P file sharing model in this chapter. We assume an unstructured P2P network operating similar to the eDonkey network. However, our model is not restricted to eDonkey, but can in fact be applied to other file sharing networks as well. The sharing of a file with size F is performed in units of chunks, which are further split into smaller units called blocks; see Fig. 14.2. In eDonkey, a chunk has the size of 9.28 MB and a block is 180 kB. After each chunk has been downloaded, it is checked for errors and if the hash value is incorrect, all blocks of the chunk are discarded and downloaded again. After all chunks of a file have been successfully downloaded, the peer may decide to keep the file as a seeder in the network for other peers to download or to remove the file from sharing (leecher or free rider). In this work, we assume that the file consists only of a single chunk, corresponding, for example, to a single mp3 audio file, as this is enough to capture the basic characteristics of the diffusion behavior.

14.3.1 Upload Queue Management and File Segmentation

In order to manage the bandwidth for other peers requesting the file, an upload queue mechanism is maintained. A peer requests individual blocks from other peers

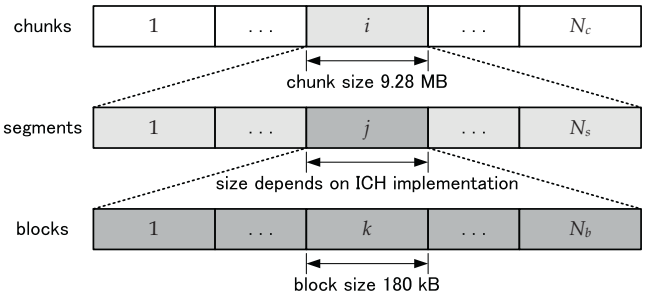


Fig. 14.2 File structure consisting of chunks, segments, and blocks.

sharing the chunk that contains the desired block. All requests are appended to the waiting list of the sharing peer and a weighting mechanism handles the scheduling of the upload queue requests for transmission. The detailed procedure of the queue management takes several features into account that depend on the individual settings of the sharing peer such as upload bandwidth and number of simultaneous uploads.

In our model, an approximative assumption simplifies the upload queue management behavior [11]. If a peer downloads a block from another peer, additional blocks might be of interest, if the providing peer is not already sharing the complete file. The weighting mechanism takes this into account by giving higher priority to peers from which blocks had been previously downloaded. We include this interaction by considering that not individual blocks, but rather a series of blocks is downloaded at a time after moving from the waiting list to the uploading list. The waiting list is modeled as a FIFO (first-in-first-out) queue and the number of consecutively downloaded blocks can be obtained from measurements [8] through the average data volume downloaded per sharing peer.

In the original version of eDonkey, error detection is done after all blocks of a chunk have been received and the complete chunk is discarded in the case of an error. However, this is not very effective and in more recent versions of eDonkey clients (e.g., eMule), the Intelligent Corruption Handling (ICH) mechanism is implemented which performs an error detection on smaller data units than chunks and that we define in the following as segments. Instead of discarding the complete chunk when at least one corrupted block is received, only all blocks of the damaged segment need to be requested again. The actual size of a segment depends on the specific settings of the ICH mechanism.

With the assumptions on the upload queue mechanism and corruption handling, it is sufficient to consider that a chunk only needs to be modeled consisting of few segments instead of several individual blocks. In this study we assume that a chunk consists of two segments (i.e., $N_s = 2$) and the size of a segment is $Z = 4.64$ MB. The size of the whole file F is less than or equal to 9.28 MB.

14.3.2 Download Bandwidth

Let us define the upload and download rates as r_u and r_d, respectively. For the sake of simplicity, we use the same assumption as in [16] of homogeneous users with ADSL connections, resulting in rates of $r_u = 128$ kbps and $r_d = 768$ kbps. Furthermore, let us denote the number of peers sharing a certain segment as S and the peers downloading it as D. Because eDonkey employs a fair share mechanism for the upload rates, there are on average S/D sharing peers serving a single downloading peer and we multiply this value with r_u. This gives us the bandwidth on the uplink.

However, because the download bandwidth could be the limiting factor, the effective downloading rate of a segment consists of the minimum of both

terms, that is, $\min(S/Dr_u, r_d)$. When downloading a segment of size Z, the term $\min(S/Dr_u, r_d)/Z$ represents the proportion of the segment that is downloaded in one unit of time, thus, the rate at which we may observe the arrival of new peers that have completely downloaded the file. We call this rate the effective transition rate. It is worth noting that in general the effective downloading rate depends on the interaction of the peers within the system (namely the number of downloaders and the number of peers sharing the segment) and on the size of segment that is effectively downloaded.

14.4 Analytical P2P File Sharing Model

Let us consider a chunk to be made up of two segments: segment 1 and segment 2 of respective sizes Z and $F-Z$, where F, as defined earlier, is the size of the complete file. We end up with three categories of peers; namely, peers with segment 1 or 2 and peers that have both segments. We say that a peer is in phase i ($i = 1,2$) when it possesses only segment i and in phase 3 in the case where it has both segments. New peers are assumed to appear at random times in the system determined by an exponential random variable whose rate depends both on the effective transition rate we introduced above and on the current state of the system, that is, the number of peers S_i in each phase $i = 1,2$, or 3. For the sake of simplicity, we can assume that the rate at which a peer stops sharing a segment is independent of the segment number, and is equal to d. The ensuing model is now described.

Let us now define the stochastic process $\{(X(t), \varphi(t))\}$, where $X(t)$ counts the total number of peers present in the system at time t, and $\varphi(t) = (\varphi_1(t), \varphi_2(t), \varphi_3(t))$ denotes the number of peers in each phase present in the system at time t, with $\varphi(t)\,\mathbf{1} = X(t)$. Here, $\mathbf{1}$ denotes a vector with ones.

We consider two views to measure the extinction probability of the file sharing process, an optimistic and a pessimistic view. In the optimistic view, we assume that the sharing process ends when no more segments are available in the system. In the pessimistic case, the file sharing process ends as soon as one of the two segments is missing. We call the latter event a catastrophe. Let us explain each resulting model in turn.

14.4.1 Level-Dependent QBD

In this first setting, recall that the sharing process ends when there are no more segments available in the system. The stochastic process $\{(X(t), \varphi(t))\}$ is an absorbing level-dependent quasi birth-and-death process, of which the generator Q can be written as

$$Q = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ A_2^{(1)} & A_1^{(1)} & A_0^{(1)} & 0 & 0 & 0 & \cdots \\ 0 & A_2^{(2)} & A_1^{(2)} & A_0^{(2)} & 0 & 0 & \cdots \\ 0 & 0 & A_2^{(3)} & A_1^{(3)} & A_0^{(3)} & 0 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}. \tag{14.1}$$

This process has been extensively studied in the past (see Latouche and Ramaswami [6] and references therein). In this setting, the time to extinction of the system is clearly equal to the time until absorption. In the remainder of this section, we first elaborate on the content of the $A_i^{(j)}$ matrices (with $i = 0,1,2$ and $j \geq 1$) and then give the algorithmic procedure in order to compute the absorption probability in this level-dependent QBD with generator Q.

14.4.1.1 Level-Dependent QBD Generator Description

When the system is in state (S_1, S_2, S_3), it means that we have S_1 peers in phase 1 (with only segment 1), S_2 peers in phase 2 (with only segment 2), and S_3 peers in phase 3 (with the complete file). We define the state subspace $L(k)$, $k \in \mathbb{N}$, as

$$L(k) = \{(S_1, S_2, S_3) : S_1 \geq 0, S_2 \geq 0, S_3 \geq 0; S_1 + S_2 + S_3 = k\},$$

which gives all states of the system at level k, that is, when k peers are present in the system. Its cardinality is clearly

$$|L(k)| = \frac{1}{2}(k+2)(k+1)$$

and we take the lexicographic order to enumerate the states of each level.

Before proceeding with the description of the transition matrix, we define two functions of crucial interest in the following; these are

$$\mu_i(S,D) = \frac{1}{Z_i} \min\left\{\frac{S}{D} r_u, r_d\right\}, \qquad i = 1,2, \tag{14.2}$$

where $Z_1 = Z$ and $Z_2 = F - Z$ are the sizes of each segment and S and D are the number of all peers currently sharing and downloading the segment, respectively.

When the system contains a single peer (i.e., when its state is in $L(1)$), this peer may stop sharing the one segment it possesses with rate d (the system then moves to $L(0)$) or another peer may start downloading the segment (the system is thus in $L(2)$). The first event occurs at a rate recorded by $A_2^{(1)}$; that is,

$$A_2^{(1)} = \begin{bmatrix} d \\ d \\ d \end{bmatrix}.$$

The latter case occurs at a rate given by the matrix $A_0^{(1)}$ as

$$A_0^{(1)} = \begin{bmatrix} \mu_1(1,1) & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu_2(1,1) & 0 & 0 \\ 0 & 0 & \mu_1(1,1) & 0 & \mu_2(1,1) & 0 \end{bmatrix}.$$

Indeed, if the system is in state $(0,1,0)$, for example, only a new peer with segment 2 may appear; that is, the system moves towards state $(0,2,0)$. This happens at a rate $\mu_2(1,1)$; see (14.2).

Usually, a peer may also perform a change of phase, that is, from 1 to 3 or from 2 to 3. Such a transition keeps the level at 1 because no new peer arrives in the system. However, if a peer in phase 1 (or phase 2) is alone in the system, it will not be able to download the missing segment and to change into phase 3. Thus, the transition rate from phase 1 (or from phase 2) to phase 3 when the system is in level 0, is $\mu_i(0,1) = 0$ for $i = 1,2$ in that particular case. The diagonal elements of $A_1^{(1)}$ (and of all $A_1^{(k)}$, $k \geq 2$) are such that $Q\mathbf{1} = \mathbf{0}$. It finally gives

$$A_1^{(1)} = \begin{bmatrix} -d - \mu_1(1,1) & 0 & 0 \\ 0 & -d - \mu_2(1,1) & 0 \\ 0 & 0 & -d - \mu_1(1,1) - \mu_2(1,1) \end{bmatrix}.$$

The possible transitions from a state $(S_1,S_2,S_3) \in L(k)$ with $k \geq 2$ are described below.

$A_2^{(k)}$: This matrix records the rate at which the system may lose a peer. A peer in phase i disappears with rate d. This latter is multiplied by the number of peers in phase i, that is, S_i with $i = 1,2,3$.

$A_0^{(k)}$: This matrix explains at which average rate a new peer may arrive in the system. There exist two possible transitions, listed in the table below. They both may be interpreted with a similar argumentation, so we limit our explanation to only the first case of possible transitions. The effective downloading rate of a new peer with segment 1 is determined as usual as the minimum between its own physical downloading rate r_d and a rate which depends on the number of peers that are sharing the available total upload bandwidth. Segment 1 is available to peers in phases 1 and 3. However, although there are only S_2 peers interested in downloading segment 1 from peers in phase 1, there are $S_1 + S_2$ peers interested in downloading segment 1 or segment 2 from peers in phase 3. It is important to take into account the S_1 supplementary peers because they also share the available upload bandwidth at peers in phase 3. This leads to an effective transition rate of

$$\mu_3(S_1,S_2,S_3) = \frac{1}{Z} \min\left\{ \left(\frac{S_1}{S_2+1} + \frac{S_3}{S_1+S_2+1} \right) r_u, r_d \right\}$$

Table 14.1 Transitions and rates for matrix $A_0^{(k)}$.

Transitions	Rates
$(S_1,S_2,S_3) \to (S_1+1,S_2,S_3)$	$\mu_3(S_1,S_2,S_3)$
$(S_1,S_2,S_3) \to (S_1,S_2+1,S_3)$	$\mu_4(S_1,S_2,S_3)$

Table 14.2 Transitions and rates for matrix $A_1^{(k)}$.

Transitions	Rates
$(S_1,S_2,S_3) \to (S_1-1,S_2,S_3+1)$	$\mu_2(S_2+S_3,S_1)$
$(S_1,S_2,S_3) \to (S_1,S_2-1,S_3+1)$	$\mu_1(S_1+S_3,S_2)$
Diagonal element	Parameter of the exponential
$(S_1,S_2,S_3) \to (S_1,S_2,S_3)$	$-k\,d-\mu_3(S_1,S_2,S_3)-\mu_4(S_1,S_2,S_3)$ $-\mu_2(S_2+S_3,S_1)-\mu_1(S_1+S_3,S_2)$

and accordingly to

$$\mu_4(S_1,S_2,S_3) = \frac{1}{F-Z}\min\left\{\left(\frac{S_2}{S_1+1}+\frac{S_3}{S_1+S_2+1}\right)r_u, r_d\right\}$$

for the case of a new peer appearing in phase 2. Table 14.1 summarizes the transitions and their corresponding rates.

$A_1^{(k)}$: A peer in phase 1 turns into a peer in phase 3 with the rate $\mu_2(S_2+S_3,S_1)$, because S_1 peers are competing for the $(S_2+S_3)\,r_u$ available bandwidth. The same argument holds for a peer in phase 2 changing into a peer in phase 3. Let us recall that the diagonal elements are such that $Q\mathbf{1}=\mathbf{0}$. The corresponding transitions and rates are shown in Table 14.2.

14.4.1.2 Probability of Extinction

Our interest lies in computing the probability that the sharing process in the particular system setting described in the previous section will terminate at some point. Let $\gamma(0)$ be the first time the system is in level 0; that is no segment is available. Let $\mathbf{e}_i$ be a unit vector with a 1 at the ith entry and 0 elsewhere. In this chapter, an empty product is, by convention, equal to the identity matrix (for $l=0$ in (14.3), for instance). We define $(\mathbf{G}_1)_i$ as the probability that the system starting in level 1 with $\varphi(0)=\mathbf{e}_i$ will eventually reach level 0; that, is,

$$(\mathbf{G}_1)_i = P[\gamma(0)<\infty \,|\, \varphi(0)=\mathbf{e}_i] \qquad i=1,2,3.$$

It was proven in [19] that this vector is explicitly given by

$$\mathbf{G}_1 = \sum_{l=0}^{\infty}\left[\prod_{i=0}^{l-1} U_{2^i}^i\right]\mathbf{D}_{2^l}^l, \tag{14.3}$$

where

$$U_{2^i}^i = P\left[\gamma(2^{i+1}) < \gamma(0) \wedge \varphi(\gamma(2^{i+1}))\,|\,X(0) = 2^i\right],$$
$$\mathbf{D}_{2^l}^l = P\left[\gamma(0) < \gamma(2^{l+1}) \wedge \varphi(\gamma(0))\,|\,X(0) = 2^l\right]$$

and where $\gamma(k)$ is defined as the first passage time to level k; that is,

$$\gamma(k) = \inf\{t \geq 0 : X(t) = k\}$$

with $k \geq 0$. Accordingly, we have

$$\left[\prod_{i=0}^{l-1} U_{2^i}^i\right] \mathbf{D}_{2^l}^l = P[\gamma(2^l) < \gamma(0) < \gamma(2^{l+1}) \wedge \varphi(\gamma(0))\,|\,X(0) = 1], \tag{14.4}$$

that is, the probability that the process starting from level 1, first visits level 2^l, then visits level 0 before visiting level 2^{l+1}. Summing (14.4) over $l = 0$ to infinity clearly gives $\mathbf{G}_1$.

The matrices U_k^l and D_k^l, respectively, of dimensions $|L(k)| \times |L(k+2^l)|$ and $|L(k)| \times |L(k-2^l)|$, are given by the following recursive equations:

$$U_k^0 = \left(-A_1^{(k)}\right)^{-1} A_0^{(k)}, \tag{14.5}$$
$$D_k^0 = \left(-A_1^{(k)}\right)^{-1} A_2^{(k)}, \tag{14.6}$$
$$U_k^l = \left[I - U_k^{l-1} D_{k+2^{l-1}}^{l-1} - D_k^{l-1} U_{k-2^{l-1}}^{l-1}\right]^{-1} U_k^{l-1} U_{k+2^{l-1}}^{l-1}, \qquad l \geq 1, \tag{14.7}$$
$$D_k^l = \left[I - U_k^{l-1} D_{k+2^{l-1}}^{l-1} - D_k^{l-1} U_{k-2^{l-1}}^{l-1}\right]^{-1} D_k^{l-1} D_{k-2^{l-1}}^{l-1}, \qquad l \geq 1. \tag{14.8}$$

Note that for $k = 2^l$ the matrix D_k^l will become a vector. A clear proof is given in [19]. The sum in (14.3) needs to be truncated in order to numerically evaluate $\mathbf{G}_1$. This matter is discussed by Latouche and Ramaswami in [6] and is addressed in our context in Sect. 14.5.

14.4.2 Level-Dependent QBD with Catastrophes

The model in the previous section considered that the file dissemination terminates when no more segments are available for sharing in the system. However, in reality when only an individual segment or an incomplete file remains in the network, no peer is able to retrieve the file completely anymore. Therefore, we now consider that a file is not available for sharing as soon as one of its segments is lost. In this case, the process ends in an absorbing state defined as belonging to $L(0)$ which is defined in this new setting as

$$L(0) = \{(0,0,0),(n,0,0),(0,n,0); n \in \mathbb{N}_0\},$$

where $\mathbb{N}$ (respectively, $\mathbb{N}_0$) is the set of natural numbers (respectively, strictly positive natural numbers). We propose not to differentiate for any $n \in \mathbb{N}_0$ between the states $(n,0,0)$ and $(0,n,0)$, but instead define a kind of metastate labeled $(k,0,0)$ and $(0,k,0)$ that gathers all of these states $(n,0,0)$ and $(0,n,0)$ for $n \in \mathbb{N}_0$, respectively. The subspace $L(0)$ is, thus, composed of three states, that is $\{(0,0,0),(k,0,0),(0,k,0)\}$ and is an absorbing level. Other level state-spaces are for $k \geq 1$:

$$L(k) = \{(i,j,l) \,|\, i,j \in \mathbb{N}, l \in \mathbb{N}_0, i+j+l = k\} \cup \{(i,j,0) \,|\, i,j \in \mathbb{N}_0, i+j = k\}. \tag{14.9}$$

The time to extinction is still equal to the time to absorption and the generator of this new level-dependent QBD is given in (14.10) as follows:

$$Q = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ A_2^{(1)} & A_1^{(1)} & A_0^{(1)} & 0 & 0 & 0 & \cdots \\ A_3^{(2)} & A_2^{(2)} & A_1^{(2)} & A_0^{(2)} & 0 & 0 & \cdots \\ A_3^{(3)} & 0 & A_2^{(3)} & A_1^{(3)} & A_0^{(3)} & 0 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}. \tag{14.10}$$

The rates of catastrophe, determined by matrix $A_3^{(k)}$, are given by the transitions and corresponding rates in Table 14.3. Accordingly, matrix $A_2^{(k)}$ becomes as shown in Table 14.4.

The other transitions in matrices $A_0^{(k)}$ and $A_1^{(k)}$ stay the same as previously described in Sect. 14.4.1.1 for the first model, taking care of the states that now belong to the subspace $L(k)$, as defined in (14.9).

Table 14.3 Transitions and rates for matrix $A_3^{(k)}$.

Transitions		Rates
$S_2 > 0$:	$(0,S_2,1) \to (0,k,0)$	d
$S_2 > 0$:	$(1,S_2,0) \to (0,k,0)$	d
$S_1 > 0$:	$(S_1,0,1) \to (k,0,0)$	d
$S_1 > 0$:	$(S_1,1,0) \to (k,0,0)$	d
	$(0,0,1) \to (0,0,0)$	d

Table 14.4 Transitions and rates for matrix $A_2^{(k)}$ with catastrophes.

Transitions		Rates
$S_1 > 1$ or $S_3 > 0$:	$(S_1,S_2,S_3) \to (S_1-1,S_2,S_3)$	$S_1\, d$
$S_2 > 1$ or $S_3 > 0$:	$(S_1,S_2,S_3) \to (S_1,S_2-1,S_3)$	$S_2\, d$
($S_1 > 0$ and $S_2 > 0$) or $S_3 > 1$:	$(S_1,S_2,S_3) \to (S_1,S_2,S_3-1)$	$S_3\, d$

The extinction probability can now be computed by extending the results by Bean and Latouche [18] to the level-dependent case. The authors in [18] analyze QBD processes with catastrophes as defined in our setting. However, their phase state-space is of infinite size, whereas in our setting this is no longer the case and makes the problem easier to handle from a numerical viewpoint.

We first define $G_0^{(k)}$ as a matrix whose (i,j)th element is the probability that the process reaches level 0 for the first time in phase j, given that the process starts in phase i of level $k \geq 1$ and levels 1 to $k-1$ are taboo. Let G_k be the matrix whose (i,j)th element is the probability that the process reaches level $k-1$ for the first time in phase j, given that the process starts in phase i of level $k \geq 1$. The extinction probability is then given by G_1 which is here also equal to $G_0^{(1)}$ by definition of this quantity. Moreover, we have for $k \geq 2$ that G_k is given by

$$G_k = \left(A_1^{(k)}\right)^{-1} A_2^{(k)} + \left(A_1^{(k)}\right)^{-1} A_0^{(k)} G_{k+1} G_k. \tag{14.11}$$

Indeed, starting from level k, the QBD may directly move to level $k-1$ with probability $\left(A_1^{(k)}\right)^{-1} A_2^{(k)}$, or it may move up to level $k+1$ with probability $\left(A_1^{(k)}\right)^{-1} A_0^{(k)}$. Upon arrival in level $k+1$, it eventually returns to level k with probability G_{k+1} and then to level $k-1$ with probability G_k. However, the equation for G_1 is slightly different and is given by

$$G_1 = \left(A_1^{(1)}\right)^{-1} A_2^{(1)} + \left(A_1^{(1)}\right)^{-1} A_0^{(1)} \left[G_2 G_1 + G_0^{(2)}\right].$$

If the process moves up to level 2 with probability $\left(A_1^{(1)}\right)^{-1} A_0^{(1)}$ (the second term in this sum), then to reach level 0, it may first return to level 1 with probability G_2 and then move to level 0 with probability G_1. It may also be directly absorbed in level 0 this time without returning to level 1 first. This happens with probability $G_0^{(2)}$. Thus, to compute G_1, we need to know G_2 and $G_0^{(2)}$. More generally, $G_0^{(k)}$ satisfies the following recursive equation:

$$G_0^{(k)} = \left(A_1^{(k)}\right)^{-1} A_3^{(k)} + \left(A_1^{(k)}\right)^{-1} A_0^{(k)} \left[G_{k+1} G_0^{(k)} + G_0^{(k+1)}\right]. \tag{14.12}$$

Its interpretation follows directly from the definition of $G_0^{(k)}$ using the same argument as before. Thus, writing $Q_i^{(k)} = \left(-A_1^{(k)}\right)^{-1} A_i^{(k)}$, $0 \leq i \leq 3$, we have explicitly

$$G_0^{(k)} = \left[I - Q_0^{(k)} G_{k+1}\right]^{-1} \left[Q_3^{(k)} + Q_0^{(k)} G_0^{(k+1)}\right]. \tag{14.13}$$

This implies that to obtain $G_0^{(2)}$ we need $G_0^{(3)}$ and so on. So, we have to truncate the QBD after some level M to be able to start the recursion. We start computing G_M using the logarithmic-reduction algorithm as described in [19]; that is,

$$G_M = \sum_{l=0}^{\infty} \left[\prod_{i=0}^{l-1} U^i_{M-1+2^i} \right] D^l_{M-1+2^l}, \tag{14.14}$$

where the matrices U^l_k and D^l_k are given by (14.5)–(14.8). Accordingly, we obtain the matrices $G_{M-1}, G_{M-2}, \ldots, G_2$ with (14.11). Using (14.13), we finally end up with the following system, which provides us the extinction probability G_1:

$$\begin{aligned} G_0^{(M)} &= Q_3^{(M)}, \\ G_0^{(M-1)} &= \left[I - Q_0^{(M-1)} G_M \right]^{-1} \left[Q_3^{(M-1)} + Q_0^{(M-1)} G_0^{(M)} \right], \\ &\vdots \\ G_0^{(1)} &= \left[I - Q_0^{(1)} G_2 \right]^{-1} \left[Q_2^{(1)} + Q_0^{(1)} G_0^{(2)} \right] = G_1. \end{aligned}$$

By truncating the QBD at level M, we actually compute the extinction probability under the taboo of level $M+1$, but a sufficiently large M will provide us a good approximation of this extinction probability.

14.5 Numerical Evaluation

Let us now consider the numerical evaluation of the proposed models, starting with the analysis of the optimistic case. We assume that initially there is a single source sharing both segments in the network, so the system starts at state $(0,0,1)$. The accuracy of our proposed algorithm for computing the extinction probabilities in Sect. 14.4.1 depends on the term l, at which the infinite sum in (14.3) is truncated. Experiments show that in our case the accuracy for $l = 3$ is already sufficient.

The resulting extinction probability as a function over the death rate is illustrated in Fig. 14.3 for file sizes of $F = 9.28$ MB and $F = 6.8$ MB, with $Z = 4.64$ MB as defined earlier being the size of the first segment. The smaller file size has the effect that the second segment is transmitted faster and thus more copies of it exist in the network, which reduces the overall extinction probability slightly. In general, this result can be interpreted as follows. The average death rate d corresponds to the reciprocal of the average sharing time of a peer in the system in seconds. Thus, in order for the content provider to keep a low extinction probability of about 0.01, he should provide incentives that a peer remains in the system for at least 100 s.

We now look at the more pessimistic case that the dissemination stops when at least one segment is no longer available for sharing. In Fig. 14.4, a file size of $F = 9.28$ MB is considered and the death rate d is fixed and equal to 10^{-2}. For the probability that none of both segments are left in the system (i.e., case $(0,0,0)$), we can see that all probabilities are identical and are thus not affected by the truncation level M. However, a slight difference can be seen when we compare the probabilities where only one kind of segment becomes extinct.

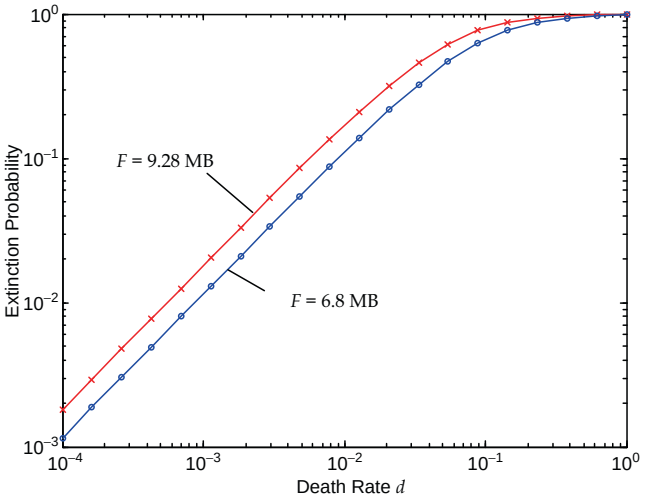


Fig. 14.3 Extinction probability for file sizes $F = 9.28$ MB and $F = 6.8$ MB. When the death rate approaches 1, the extinction probability increases drastically to 1.

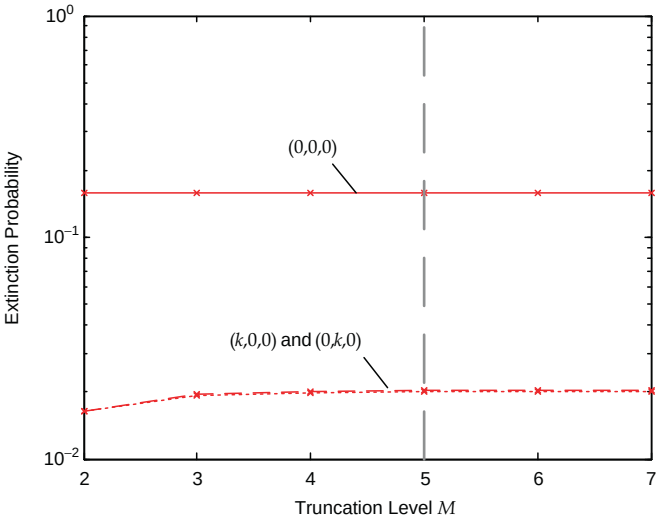


Fig. 14.4 Influence of the truncation level M on the accuracy. A value of about $M = 5$ proves to be accurate enough, so in the following evaluations we use this value as the truncation point.

If we plot the extinction probabilities from the second model with catastrophes over the death rate, we can recognize in Fig. 14.5 that the probabilities to reach $(0,0,0)$ lie above the two curves corresponding to states $(k,0,0)$ and $(0,k,0)$. The reason why they are larger can be interpreted as follows. Initially, the system starts at state $(0,0,1)$, that is, with exactly a single sharing peer. In order to reach the

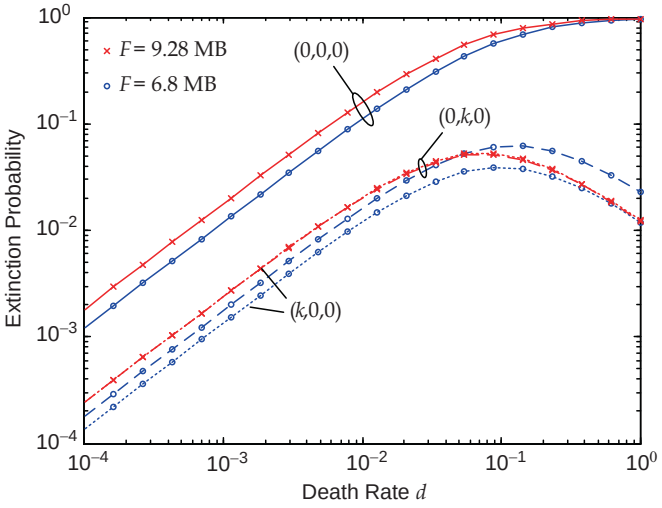


Fig. 14.5 Extinction probabilities with catastrophes for $M = 5$ and file sizes of $F = 9.28$ MB and $F = 6.8$ MB.

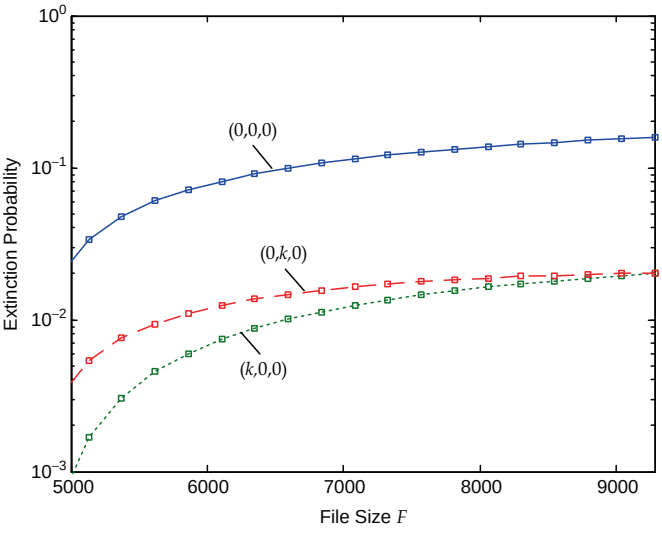


Fig. 14.6 Influence of file size F on the extinction probabilities for $d = 10^{-2}$.

absorbing state $(0,0,0)$, this peer may either make a direct transition by leaving the system or an indirect path by first giving birth to other peers which then all leave after time. On the other hand, in order to reach one of the other absorbing states $(k,0,0)$ or $(0,k,0)$ at least one birth must take place to increment S_1 or S_2, respectively. Thus, a direct transition from $(0,0,1)$ to an absorbing state of that type does not exist in this case, causing a reduction in the weight of the probability.

Additionally, when we look at the shape of the curves, we can recognize that both curves for $(k,0,0)$ and $(0,k,0)$ are identical, when we consider equal segment sizes and the probability for finding and sharing both segments is equal. With $F = 6.8$ MB the second segment is only half in size of the first, which results in a higher extinction probability of the first segment. The curves lie below the corresponding curves for $F = 9.28$ MB when the death rate d is small. However, in both cases we can see that when the death rate exceeds 10^{-1} the extinction probabilities drop again. At this point it is more likely that the sharing process will stop before any segment is actually downloaded at all; that is

$$d \gg \mu_1(1,1) + \mu_2(1,1),$$

where $\mu_1(1,1) + \mu_2(1,1)$ corresponds to the rate of observing a first new peer with any one of the segments.

The influence of the file size F and, thus, the different size of the second segment is illustrated in Fig. 14.6. We can recognize firstly that for a death rate of $d = 10^{-2}$ the extinction probabilities increase with the file size and, secondly, that when the second segment size is small, the difference between the extinction probabilities of states $(k,0,0)$ and $(0,k,0)$ is large. As expected, when both sizes are equal, both curves approach the same value.

14.6 Conclusions

We provided in this chapter an algorithmically tractable analysis of a level-dependent QBD process with and without catastrophe in terms of the absorption probability, which corresponds to the extinction probability of a file, when we apply the model to file diffusion in unstructured P2P file sharing networks. Numerical results have confirmed that there is a need for the content provider to offer incentives to the peers to encourage sharing and a long sojourn time in the system in order to maintain a sufficiently low extinction probability.

In the future we will use this model to analytically derive further performance measures, especially transient ones such as the distribution of the number of peers present in the system. Furthermore, we would like to enhance the model to consider a more sophisticated peer behavior by including, for example, their willingness to share, impatient peers, and pollution.

References

1. D. Rubenstein and S. Sahu, Can unstructured P2P protocols survive flash crowds?, *IEEE/ACM Transactions on Networking*, vol. 13, no. 3, pp. 501–512, 2005.
2. N. Christin, A. S. Weigend, and J. Chuang, Content availability, pollution and poisoning in file sharing peer-to-peer networks, in *Proc. ACM Conference on Electronic Commerce (EC)*, 2005.

3. A. Binzenhöfer and K. Leibnitz, Estimating churn in structured P2P networks, in *Proc. 20th International Teletraffic Congress (ITC)*, 2007.
4. I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan, Chord: A scalable peer-to-peer lookup service for Internet applications, in *Proc. ACM SIGCOMM*, 2001.
5. S. Hautphenne, K. Leibnitz, and M.-A. Remiche, Extinction probability in peer-to-peer file diffusion, *ACM SIGMETRICS Performance Evaluation Review*, vol. 34, no. 2, pp. 3–4, 2006.
6. G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*, Philadelphia: ASA-SIAM Series on Statistics and Applied Probability, 1999.
7. S. Saroiu, K. P. Gummadi, R. J. Dunn, S. D. Gribble, and H. M. Levy, An analysis of internet content delivery systems, in *Proc. 5th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2002.
8. K. Tutschku, A measurement-based traffic profile of the eDonkey filesharing service, in *Proc. 5th Passive & Active Measurement Workshop (PAM)*, 2004.
9. K. Tutschku and H. de Meer, A measurement study on signaling on Gnutella overlay networks, in *Proc. Communication in Distributed Systems (KiVS)*, 2003.
10. M. Izal, G. Urvoy-Keller, E. Biersack, P. Felber, A. A. Hamra, and L. Garces-Erice, Dissecting BitTorrent: Five months in a torrent's lifetime, in *Proc. 5th Passive and Active Measurement Workshop (PAM)*, 2004.
11. T. Hoßfeld, K. Leibnitz, R. Pries, K. Tutschku, P. Tran-Gia, and K. Pawlikowski, Information diffusion in eDonkey-like P2P networks, in *Proc. Australian Telecommunication Networks and Applications Conference (ATNAC)*, 2004.
12. Z. Ge, D. Figueiredo, S. Jaiswal, J. Kurose, and D. Towsley, Modeling peer-peer file sharing systems, in *Proc. 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, pp. 2188–2198, 2003.
13. D. Qiu and R. Srikant, Modeling and performance analysis of BitTorrent-like peer-to-peer networks, in *Proc. ACM SIGCOMM*, 2004.
14. G. Carofiglio, R. Gaeta, M. Garetto, P. Giaccone, E. Leonardi, and M. Sereno, A statistical physics approach for modelling P2P systems, *ACM SIGMETRICS Performance Evaluation Review*, vol. 33, no. 2, pp. 3–5, 2005.
15. X. Yang and G. de Veciana, Service capacity of peer to peer networks, in *Proc. 23rd Conference of the IEEE Communications Society (INFOCOM)*, pp. 2242–2252, 2004.
16. K. Leibnitz, T. Hoßfeld, N. Wakamiya, and M. Murata, On pollution in eDonkey-like peer-to-peer file-sharing networks, in *Proc. 13th GI/ITG Conference on Measurement, Modeling, and Evaluation of Computer and Communication Systems (MMB)*, 2006.
17. K. Leibnitz, T. Hoßfeld, N. Wakamiya, and M. Murata, Peer-to-peer vs. client/server: Reliability and efficiency of a content distribution service, in *Proc. 20th International Teletraffic Congress (ITC)*, 2007.
18. N. Bean and G. Latouche, Numerical approximations for QBDs with infinitely many phases, *Université Libre de Bruxelles, Département d'Informatique, Technical report*, 2007.
19. V. Ramaswami and P. Taylor, Some properties of the rate operators in level dependent quasi-birth-and-death processes with a countable number of phases, *Communications in Statistics: Stochastic Models*, vol. 12, no. 1, pp. 143–164, 1996.

Chapter 15
Performance Analysis of a Decentralized Content Delivery System with FEC Recovery

Kenji Kirihara, Hiroyuki Masuyama, Shoji Kasahara, and Yutaka Takahashi

Abstract This chapter considers the performance of a decentralized content delivery system where video data are simultaneously delivered without duplication by multiple streaming video servers, resulting in a low sending rate per video server. Focusing on a multiple-server video streaming service reinforced by forward error correction (FEC), we model the system as a set of independent GI+M/M/1/K queues, and derive the block-level loss probability. Numerical results show that the decentralized content delivery system with FEC recovery is significantly effective to guarantee video quality even when the background traffic intensity is high.

15.1 Introduction

With the recent advancement of network technologies enabling ultra-high-speed data transmission, video streaming service over the Internet has attracted considerable attention. The Internet, however, is a best-effort network, and thus the quality of service (QoS) for video streaming is not strictly guaranteed due to packet loss and/or delay.

In order to enhance the resilience to packet loss, a number of approaches have been proposed and studied. Among them automatic repeat request (ARQ) and forward error correction (FEC) are commonly deployed for loss recovery. ARQ is an acknowledgment-based error recovery, in which lost data packets are retransmitted reactively by the sender host. ARQ is an efficient resilience mechanism for packet loss if the round-trip time between the sender and receiver hosts is significantly small. However, ARQ is not suitable for a network with a large round-trip time, resulting in a larger end-to-end delay caused by multiplicative transmissions.

K. Kirihara, H. Masuyama, S. Kasahara and Y. Takahashi
Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
e-mail: kirihara@sys.i.kyoto-u.ac.jp; {masuyama, shoji, takahashi}@i.kyoto-u.ac.jp

W. Yue et al. (eds.), *Advances in Queueing Theory and Network Applications,*

On the other hand, FEC is a one-way recovery technique based on open-loop error control. FEC generates redundant data from original data, and both original and redundant data are transmitted to a destination. If the amount of lost data is less than or equal to a prespecified threshold, the lost data can be reconstructed. In this chapter, we consider a packet-level FEC scheme [1]. Because FEC needs no retransmission, it is suitable for real-time applications with stringent delay constraint such as video streaming. However, FEC does not work well against packet burst loss because the amount of redundant data has to be predetermined with the estimate of the packet loss probability.

An alternative approach to guarantee QoS against packet loss is multiple-sender video streaming [2]. This is a decentralized content delivery scheme in which video data are divided into segments to be simultaneously delivered by multiple streaming sender hosts. Each sending rate per server is significantly smaller than that of a single-sender case, achieving a small overall packet loss probability at the destination. In [2], Nguyen and Zakhor proposed a distributed video streaming protocol consisting of a rate allocation algorithm and a packet partition algorithm. Its performance was investigated by simulation and experiments with a real network. FEC recovery performance has also been studied in the literature [3]–[5], however, little work has been devoted to analyze the compound effect of the decentralized content distribution mechanism and FEC.

With the recent advancement of photonic networking technology such as wavelength division multiplexing (WDM), the bottleneck of data transmission has shifted from backbone networks to access ones (the last-mile bandwidth bottleneck [6]). This implies that edge routers of backbone networks are likely to be the bottleneck of data transmission for real-time applications. In real-time applications such as VoIP and Internet TV, packets are sent to the network at a constant bit rate. Therefore, it is important to consider the case where interarrival times of packets to a bottleneck edge router are almost the same.

In this chapter, we analyze the performance of this decentralized content delivery system by a queueing theoretical approach. We consider a multiple-sender video streaming service, and focus on disjoint parts of multiple routes to the destination. Assuming that there exists a bottleneck router along the disjoint part of each route, we model each bottleneck router as a single-server finite queueing system with both general renewal and Poisson inputs. We derive the block-level loss probability, and investigate the resulting video quality of multiple-sender video streaming with and without FEC. Note that the assumption of the general renewal input for main traffic enables us to describe various arrival processes including constant interarrival times.

The chapter is organized as follows. Section 15.2 describes the analysis model in detail, and derives the block-loss probability. Numerical results are presented in Sect. 15.3, and we conclude this chapter in Sect. 15.4.

15.2 Model and Analysis

We consider a multiple-sender distributed video streaming service. Let S denote the number of video servers. A video dataum is divided into S parts, each of which are simultaneously delivered along with different routes. We assume that a video data frame consists of D packets. N redundant data packets are generated from the D original data packets, and a set of $M(=D+N)$ packets is called a block. If the number of lost packets among the M packets is less than or equal to N, the original data packets can be reproduced completely regardless of the lost packets. On the other hand, if the number of lost packets among the M packets is greater than N, the original data packets cannot be recovered. We call this event a block loss.

Video streaming service is supported by S servers. We divide M packets per frame into S groups: group l $(l = 1,\ldots,S)$ with $M^{(l)}$ packets. Note that $\sum_{l=1}^{S} M^{(l)} = M$. Server l manages the $M^{(l)}$ packets in group l and sends those packets to the destination. Note that we have S streaming routes for a video service. Suppose that there exists a bottleneck router along each route and that packet loss occurs independently at each bottleneck router.

We model each bottleneck router as a single-server queueing system with a finite buffer that is fed by two independent input processes. In the following, the packet flow sent from a streaming server is called the main traffic, and the other packet flow the background traffic. The interarrival times of packets in the main traffic sent from server l are independently identically distributed (i.i.d.) with a general distribution $G^{(l)}(x)$. The packet arrivals in the background traffic form a Poisson process with rate $\lambda^{(l)}$. The capacity of the system with server l is $K^{(l)}$. Note that the bottleneck router forwards not only the packets from the video server but also the packets belonging to the background traffic. Therefore, it is natural to assume that the packet size is not the same. Then, we assume that the service time of a packet is exponentially distributed with rate $\mu^{(l)}$. From the above assumptions, we have a GI+M/M/1/K-type queueing model for each bottleneck router.

We derive the block-loss probability that a block is not eventually retrieved at the destination. Let $p^{(l)}(k \mid M^{(l)})$ $(l = 1,\ldots,S)$ denote the probability that k $(k = 0,1,\ldots,M^{(l)})$ packets out of $M^{(l)}$ packets sent from server l are lost. We can compute $p^{(l)}(k \mid M^{(l)})$ from the analytical result in [5]. (The derivation of $p^{(l)}(k \mid M^{(l)})$ is summarized in the appendix.) Noting that original data packets for a video frame can be recovered if the number of lost packets is less than or equal to N, the block-loss probability P_{Loss} is given by

$$P_{Loss} = 1 - \sum_{n=0}^{N} \sum_{k_1+\cdots+k_S=n} p^{(1)}(k_1 \mid M^{(1)}) p^{(2)}(k_2 \mid M^{(2)}) \cdots p^{(S)}(k_S \mid M^{(S)}). \quad (15.1)$$

15.3 Numerical Results

We assume that the transmission rate of a video streaming service for the single-server case is 10 Mbps, and that the output transmission speed of bottleneck routers is 100 Mbps. It is supposed that the video frame rate is 30 [frame/s], and that a frame has the same number of packets as that of a block. The packet size is 1250 bytes. Thus a block has $D = 34$ original data packets, and the packet service rate of packets at bottleneck routers is $\mu = 1 \times 10^4$ [packet/s].

For the multiple-sender case, we assume that the number of video servers is two and that $M^{(1)} = M^{(2)} = (34+N)/2$. The number of FEC redundant packets N is set to 0, 2, and 4. The packet interarrival time of main traffic from each video server is constant. The system capacities $K^{(l)}$ are assumed to be the same and set to $K^{(l)} = K = 10$ and 100. In what follows, we assume that the flow rates of background traffic are equal at both of the bottleneck routers. Note that when N FEC redundant packets are added to D original data packets, the resulting packet transmission rate becomes $(D+N)/D$ times larger than the original one.

The block-loss probability for the multiple-sender case is calculated by (15.1). We also calculate the block-loss probability for the single-sender case using the result in [5]. We define the FEC redundancy as N/D.

15.3.1 Impact of Background Traffic

In this subsection, we investigate how the bandwidth of background traffic affects the block-loss probability.

Figures 15.1 and 15.2 show the block-loss probability against the bandwidth of background traffic in the cases $K = 10$ and 100, respectively. We observe from

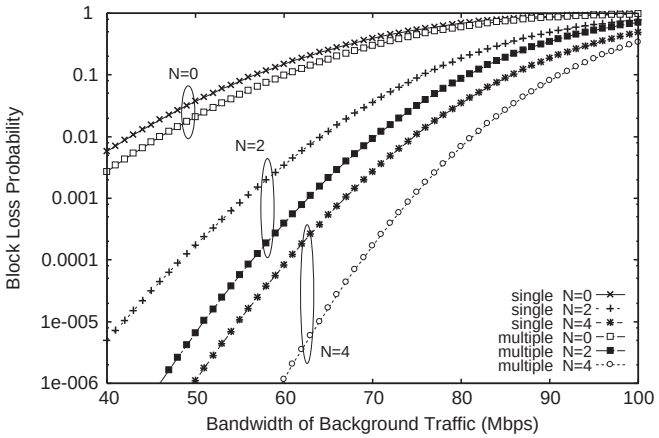


Fig. 15.1 Block-loss probability versus bandwidth of background traffic ($S = 2$, $K = 10$, $D = 34$).

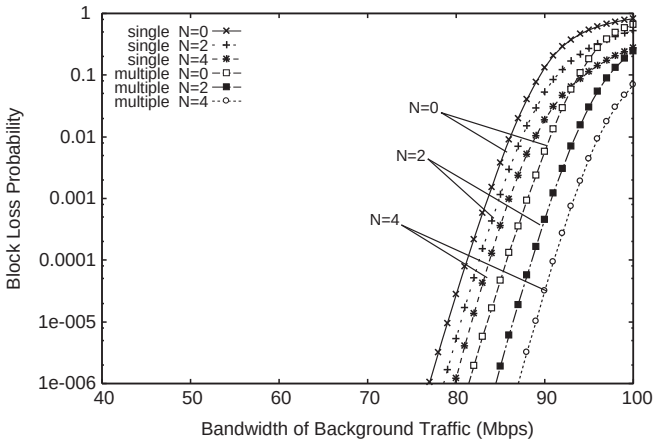


Fig. 15.2 Block-loss probability versus bandwidth of background traffic ($S = 2$, $K = 100$, $D = 34$).

Figs. 15.1 and 15.2 that the block-loss probability increases monotonically when the bandwidth of the background traffic is large, as expected. Owing to the multiple-sender effect, the block-loss probability is further improved for the same number of redundant packets as the single-sender case. We also observe the decrease in the block-loss probability when the number of FEC redundant packets increases, and that the block-loss probability is effectively reduced by FEC redundant packets in the system with a small capacity. When the system capacity is small, a packet-loss event frequently occurs, making the packet-loss process random. Because FEC works well against random packet-loss processes, the block-loss probability is greatly improved by FEC when the system capacity is small.

Next we investigate how the bandwidth of background traffic affects the minimum FEC redundancy. Here, the minimum FEC redundancy is such that the block-loss probability is smaller than a prespecified value $P_{Loss}^{(\alpha)}$. Figures 15.3 and 15.4 illustrate the minimum FEC redundancy against the bandwidth of background traffic in cases of K=10 and K=100, respectively. For each value of K, we calculated the minimum FEC redundancy for the cases $P_{Loss}^{(\alpha)} = 10^{-2}$, 10^{-3} and 10^{-4}.

It is observed from Fig. 15.3 that the FEC redundancy in the multiple-sender case is smaller than that in the single-sender case when $P_{Loss}^{(\alpha)}$ is fixed. This is due to a small packet-loss probability in the multiple-sender case. In Fig. 15.4, the minimum FEC redundancy remains zero at 80 Mbps background traffic in all the cases, because the packet-loss events hardly occur in a system with large capacity. When the bandwidth of background traffic is greater than 80 Mbps, the minimum FEC redundancy increases rapidly. This tendency of the minimum FEC redundancy is the same as in Fig. 15.3. Comparing Fig. 15.3 with Fig. 15.4, FEC is effective in a wide range of background traffic when the system capacity is small.

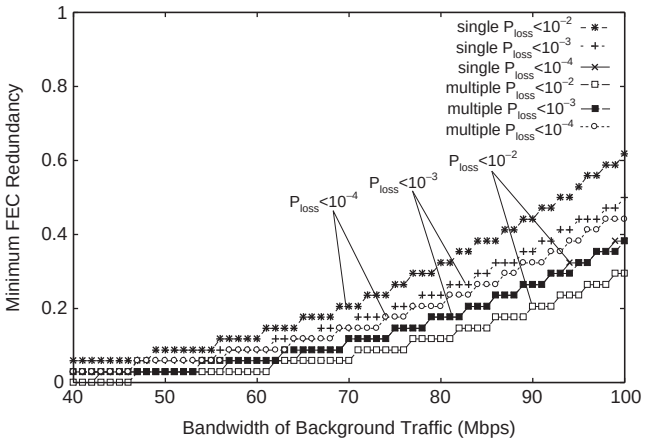


Fig. 15.3 Minimum FEC redundancy versus bandwidth of background traffic ($S = 2$, $K = 10$).

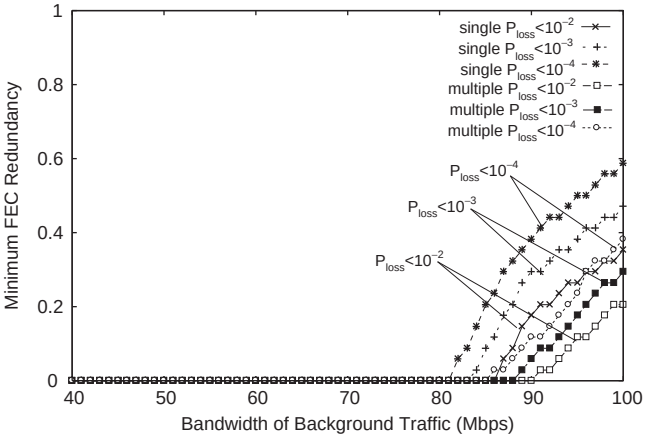


Fig. 15.4 Minimum FEC redundancy versus bandwidth of background traffic ($S = 2$, $K = 100$).

15.3.2 Impact of Service Rate at Bottleneck Router

In this subsection, we investigate how the output transmission speed of the bottleneck router affects the block-loss probability and the minimum FEC redundancy.

Figures 15.5 and 15.6 show the block-loss probability against the transmission speed in the cases $K = 10$ and 100, respectively. Note that when the transmission speed is η bps, the corresponding service rate of a packet at the bottleneck router μ is equal to $\eta \times 10^4$ [packet/s]. The bandwidth of background traffic is set to 50 Mbps.

We observe from Fig. 15.5 that the block-loss probability decreases monotonically and gradually when the transmission speed increases. It is also observed that the decentralized content delivery system is greatly effective for the block-loss

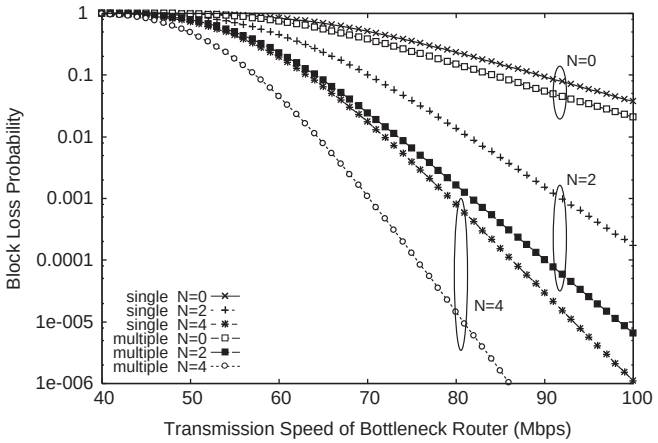


Fig. 15.5 Block-loss probability versus transmission speed ($S = 2$, $K = 10$, $D = 34$).

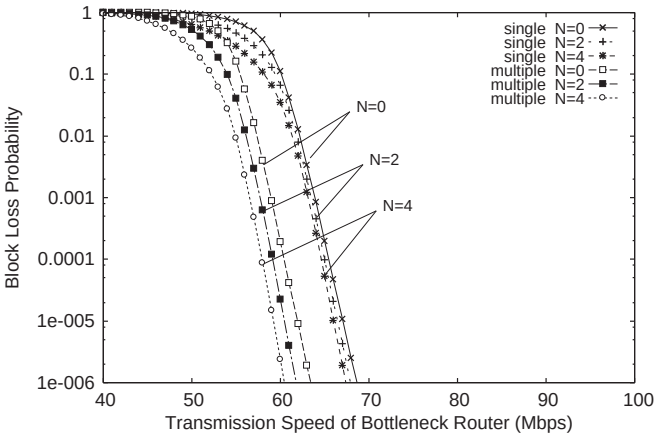


Fig. 15.6 Block-loss probability versus transmission speed ($S = 2$, $K = 100$, $D = 34$).

probability as the FEC redundancy increases. In addition, the block-loss probability for the multiple-sender case is significantly smaller than that for the single-sender case.

It is observed from Fig. 15.6 that the block-loss probability for $K = 100$ exhibits the same tendency as that in Fig. 15.5. Note that the block-loss probability is greatly improved by a high transmission speed, rather than FEC and the decentralized content distribution mechanism.

Figures 15.7 and 15.8 illustrate the minimum FEC redundancy against the transmission speed of the bottleneck router in the cases $K = 10$ and 100, respectively. The minimum FEC redundancy was calculated for $P_{Loss}^{(\alpha)} = 10^{-2}$, 10^{-3}, and 10^{-4}.

In Fig. 15.7, the minimum FEC redundancy in the multiple-sender case is smaller than that in the single-sender case when $P_{Loss}^{(\alpha)}$ is fixed. In addition, the minimum

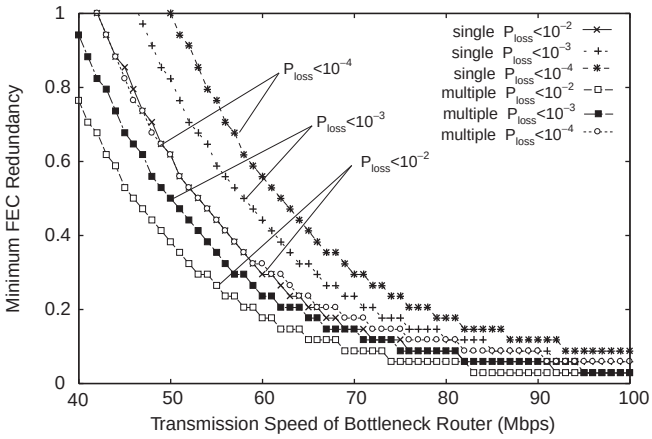


Fig. 15.7 Minimum FEC redundancy versus transmission speed ($S = 2$, $K = 10$).

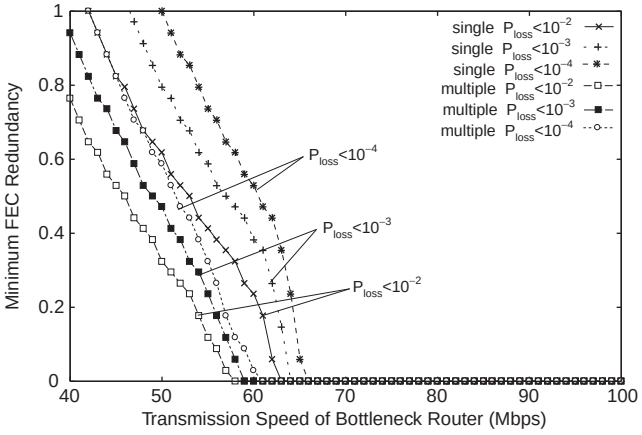


Fig. 15.8 Minimum FEC redundancy versus transmission speed ($S = 2$, $K = 100$).

FEC redundancy gradually decreases when the transmission speed increases. We observe from Fig. 15.8 that the minimum FEC redundancy reaches zero when the transmission speed is larger than around 60 Mbps, and that the other tendency is the same as in Fig. 15.7. These results imply that when the output transmission speed is small, the decentralized content delivery system with FEC recovery is significantly effective for the block-loss probability.

15.3.3 Impact of System Capacity

In this subsection, we investigate the impact of the system capacity on the minimum FEC redundancy. The QoS requirement considered here is $P_{Loss}^{(\alpha)} = 10^{-2}$, 10^{-3}, and 10^{-4}. With each $P_{Loss}^{(\alpha)}$, we calculated the minimum FEC redundancy in cases of

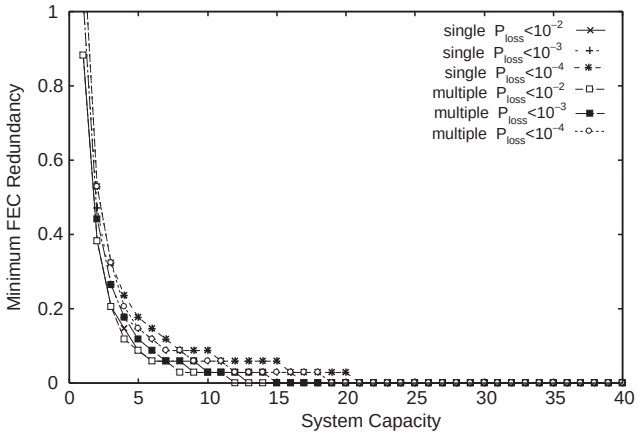


Fig. 15.9 Minimum FEC redundancy versus system capacity ($S = 2$, 50 Mbps background traffic).

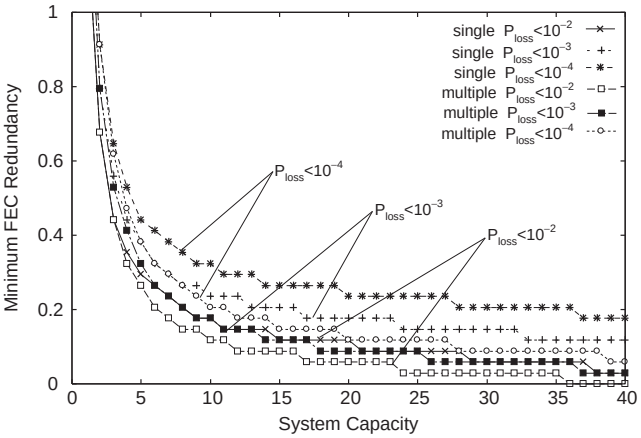


Fig. 15.10 Minimum FEC redundancy versus system capacity ($S = 2$, 80 Mbps background traffic).

50 Mbps and 80 Mbps of the bandwidth of background traffic. It is supposed that the output transmission speed of the bottleneck router is 100 Mbps.

Figure 15.9 shows the minimum FEC redundancy against the system capacity when the bandwidth of background traffic is 50 Mbps. We observe from Fig. 15.9 that the minimum FEC redundancy decreases rapidly when the system capacity is about 5, and that for each P_{Loss} the minimum FEC redundancy reaches zero when the system capacity is greater than 20. Note that the variation of the minimum FEC redundancy is small. This implies that the block-loss probability is greatly improved by the system capacity rather than the decentralized content distribution mechanism with FEC recovery.

Figure 15.10 shows the minimum FEC redundancy when the bandwidth of background traffic is 80 Mbps. In Fig. 15.10, the minimum FEC redundancy decreases

monotonically as the system capacity is large. This tendency is the same as in Fig. 15.9. Comparing Fig. 15.9 with Fig. 15.10, there is a difference between the minimum FEC redundancy in the multiple-sender case and in the single-sender case. A remarkable point of Fig. 15.10 is that the minimum FEC redundancy is likely to remain constant when the system capacity increases. This implies that the decentralized content delivery system with FEC recovery is more effective than enriching system capacity when the system is overloaded.

15.3.4 Impact of Number of Video Servers

Finally, we investigate how the number of video servers improves the video QoS. We set $N = 4$ and hence the number of packets in a block M is 38. We consider four cases of $S = 1, 2, 3$, and 4. Table 15.1 shows the parameter values of $M^{(l)}$s for each S.

Figure 15.11 represents the block-loss probability against the bandwidth of background traffic for $K = 10$ and 100. We assume that all the background traffic intensities at bottleneck routers are the same. This scenario can be regarded as the worst case for multiple-sender transmission.

Table 15.1 Number of packets within each group.

S	Parameters	Values
1	M	38
2	$(M^{(1)}, M^{(2)})$	$(19, 19)$
3	$(M^{(1)}, M^{(2)}, M^{(3)})$	$(13, 13, 12)$
4	$(M^{(1)}, M^{(2)}, M^{(3)}, M^{(4)})$	$(10, 10, 9, 9)$

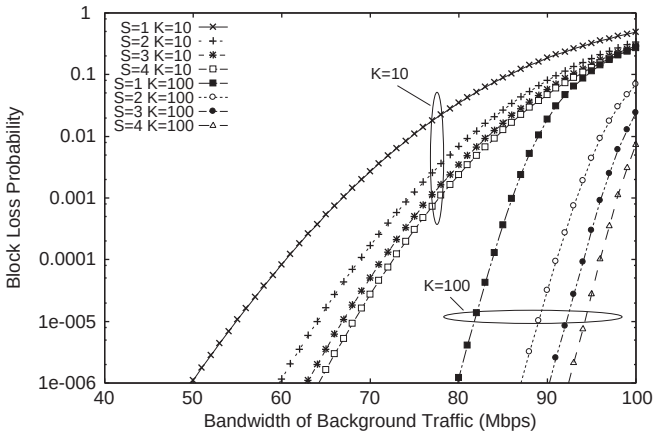


Fig. 15.11 Block-loss probability versus bandwidth of background traffic ($D = 34$, $N = 4$).

It is observed from the figure that for both Ks, the block-loss probability is improved with the increase in the number of video servers, as expected. When $K = 10$, the block-loss probability for $S = 2$ is significantly smaller than that for the single-server case. However, the block-loss probabilities for $S = 3$ and 4 are not greatly improved. Note that $M = 38$ corresponds to 11.4 Mbps of the video sending rate for the single-server case. Roughly speaking, the video sending rate per server is 5.7 Mbps for $S = 2$, 3.8 Mbps for $S = 3$, and 2.9 Mbps for $S = 4$. That is, the resulting video sending rates per server are relatively small in comparison with the background traffic intensity. When $K = 100$, on the other hand, the block-loss probability is significantly small and greatly improved with the increase in the number of video servers. This result suggests that the decentralized content delivery system supported by multiple servers can guarantee video QoS effectively when the network is heavily congested.

Figure 15.12 shows the minimum FEC redundancy against the system capacity per router when the bandwidth of background traffic is 80 Mbps. The QoS requirement considered here is $P_{loss}^{(\alpha)} = 10^{-4}$. In Fig. 15.12, the minimum FEC redundancy for $S = 1$ is the largest, and the minimum FEC redundancy decreases with the increase in S. A remarkable point here is that the minimum FEC redundancies for $S = 2$, 3, and 4 are almost the same, regardless of the system size. Note that the bandwidth of background traffic is 80 Mbps. Because the link capacity is set to 100 Mbps, the overall traffic intensity at each bottleneck router is more than 0.8; that is, the system is heavily utilized. Under such a heavy loaded condition, the increase in the system capacity is more effective for video streaming than increasing the number of video servers. This result also implies that if the buffer size of bottleneck routers is large, video QoS can be guaranteed with two video servers.

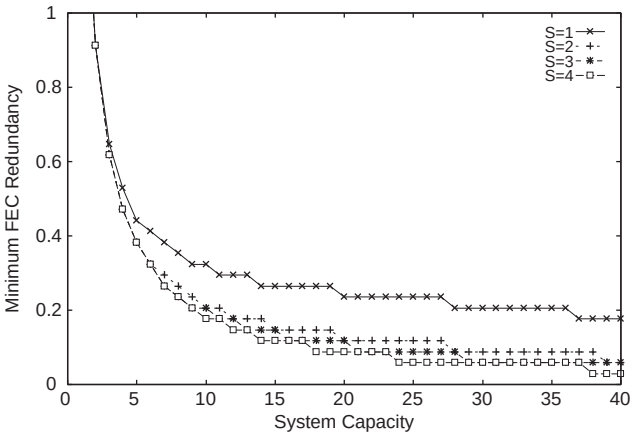


Fig. 15.12 Minimum FEC redundancy versus system capacity (80 Mbps background traffic, $P_{loss} < 10^{-4}$).

15.4 Conclusions

This chapter analyzed the performance of the decentralized content delivery system with FEC recovery. We focused on a multiple-sender video streaming service and modeled it as a set of GI+M/M/1/K queues, deriving the block-level loss probability. Numerical results showed that decentralized content delivery in cooperation with FEC recovery is significantly effective for preserving video quality even when the background traffic intensity is high. In particular, when the system capacity is small and the network is overloaded, multiple-sender video streaming succeeds in guaranteeing video QoS with less FEC redundancy than the single-server case. A remarkable point is that two video servers are enough to guarantee video QoS even when the network is heavily utilized. In this overloaded condition, enriching router buffers is more effective than increasing the number of video servers. In general, increasing the number of video servers causes a large control overhead of video-content management. The fact that a few video servers are enough to guarantee video QoS is significant from the viewpoint of video-content management.

Appendix: Derivation of Probability $\mathbf{p^{(l)}(k \mid M^{(l)})}$

This appendix summarizes the derivation of the probability $p^{(l)}(k \mid M^{(l)})$ ($l = 1,\ldots,S$). For details, see [5], where $\mathbf{p}_M(k)\mathbf{e}$ corresponds to $p^{(l)}(k \mid M^{(l)})$. For simplicity, we omit superscript "$^{(l)}$" in this appendix. Thus, for example, although we write λ, μ, and M for $\lambda^{(l)}$, $\mu^{(l)}$, and $M^{(l)}$, respectively, λs, μs, and Ms herein are different from original λs, μs, and Ms in the preceding sections.

We first consider the stationary queue length distribution immediately before an arrival from main traffic in the GI+M/M/1/K queue, which models a bottleneck router. Let T_m ($m = 0,\pm1,\pm2,\ldots$) denote the arrival epoch of the mth packet from main traffic. We then assume that the system reaches steady state at time T_0. Let L_m^- ($m = 0,\pm1,\pm2,\ldots$) denote the number of packets in the system immediately before time T_m. Note that during each interval between arrivals (T_m, T_{m+1}), the behavior of the queueing process is stochastically the same as that of the M/M/1/K queue with arrival rate λ and service rate μ. Thus $\{L_m^-; m = 0,\pm1,\pm2,\ldots\}$ is a Markov chain whose transition probability matrix Π is given by

$$\Pi = \Lambda \int_0^\infty \exp(\mathbf{Q}x)\,dG(x), \tag{15.2}$$

where $G(x)$ denotes the distribution of interarrival times of packets from main traffic, and where Λ and $\mathbf{Q}$ denote $(K+1)\times(K+1)$ matrices that are given by

$$\Lambda = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ddots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix},$$

$$\mathbf{Q} = \begin{pmatrix} -\lambda & \lambda & 0 & \dots & 0 & 0 \\ \mu & -(\lambda+\mu) & \lambda & \ddots & \vdots & \vdots \\ 0 & \mu & -(\lambda+\mu) & \ddots & 0 & 0 \\ 0 & 0 & \mu & \ddots & \lambda & 0 \\ \vdots & \vdots & \ddots & \ddots & -(\lambda+\mu) & \lambda \\ 0 & 0 & 0 & \ddots & \mu & -\mu \end{pmatrix}.$$

We define π^- as a $1 \times (K+1)$ vector whose jth ($j = 0, 1, \dots, K$) element π_j^- represents $\Pr[L_1^- = j]$. We then have

$$\pi^- \Pi = \pi^-, \qquad \pi^- \mathbf{e} = 1,$$

where $\mathbf{e}$ denotes a column vector of ones with appropriate dimension.

Next in order to derive $p(k \mid M)$, we consider an arbitrary group that consists of M packets sent from a server. We assume that the M packets of the group arrive at the system at times T_1 through T_M. We then call the packet arriving at time T_m ($m = 1, 2, \dots, M$) packet m. Let L_m denote the number of packets in the system at time T_m. Let N_m ($m = 1, 2, \dots, M$) denote the number of lost packets among packets 1 through m at time T_m. We define $\mathbf{p}_m(k)$ ($m = 1, 2, \dots, M$, $k = 0, 1, \dots, M$) as a $1 \times K$ vector whose jth ($j = 1, 2, \dots, K$) element $p_{m,j}(k)$ is given by

$$p_{m,j}(k) = \Pr[N_m = k, L_m = j].$$

Because $N_m \le m$ for all $m = 1, 2, \dots, M$,

$$\mathbf{p}_m(k) = \mathbf{0}, \quad \text{for all } k = m+1, m+2, \dots, M. \tag{15.3}$$

By using $\mathbf{p}_m(k)$, the probability $p(k \mid M)$ is given by

$$p(k \mid M) = \mathbf{p}_M(k)\mathbf{e}. \tag{15.4}$$

In what follows, we discuss the $\mathbf{p}_m(k)$s ($m = 1, 2, \dots, M$, $k = 0, 1, \dots, M$). Note that if $L_1^- < K$, packet 1 can join the queue and hence $L_1 = L_1^- + 1$. Note also that if $L_1^- = K$, packet 1 is lost and $L_1 = K$. Thus $p_{1,j}(0)$ and $p_{1,j}(1)$ ($j = 1, 2, \dots, K$) are given by

$$p_{1,j}(0) = \Pr[L_1^- < K, L_1^- = j-1] = \pi_{j-1}^-, \qquad j = 1,2,\ldots,K, \tag{15.5}$$

$$p_{1,j}(1) = \begin{cases} 0, & j = 1,2,\ldots,K-1 \\ \pi_K^-, & j = K, \end{cases} \tag{15.6}$$

respectively, or equivalently,

$$\mathbf{p}_1(0) = (\pi_0^-, \pi_1^-, \ldots, \pi_{K-1}^-), \qquad \mathbf{p}_1(1) = (0,0,\ldots,0,\pi_K^-).$$

We now define $\mathbf{A}(v)$ $(v = 0,1)$ as a $K \times K$ matrix whose (i,j)th element $A_{i,j}(v)$ $(i,j = 1,2,\ldots,K)$ is given by

$$A_{i,j}(v) = \Pr[\Theta_m = v, L_m = j \mid L_{m-1} = i],$$

where $\Theta_m = 1$ if packet m is lost, and otherwise $\Theta_m = 0$. It is easy to see that for $m = 2,3,\ldots,M$,

$$\mathbf{p}_m(0) = \mathbf{p}_{m-1}(0)\mathbf{A}(0), \tag{15.7}$$

$$\mathbf{p}_m(k) = \mathbf{p}_{m-1}(k-1)\mathbf{A}(1) + \mathbf{p}_{m-1}(k)\mathbf{A}(0), \qquad k = 1,2,\ldots,M. \tag{15.8}$$

Finally, we consider $\mathbf{A}(v)$ $(v = 0,1)$. If $L_m^- < K$, $\Theta_m = 0$ and $L_m = L_m^- + 1$. Thus we have

$$A_{i,j}(0) = \Pr[L_m^- < K, L_m^- = j-1 \mid L_{m-1} = i] = \Gamma_{i,j-1}, \qquad i,j = 1,2,\ldots,K, \tag{15.9}$$

where $\Gamma_{i,j}$ $(i,j = 0,1,\ldots,K)$ denotes the (i,j)th element of $\Gamma = \int_0^\infty \exp(\mathbf{Q}x)dG(x)$. Note here that $\Pi = \Lambda\Gamma$ (see (15.2)). Furthermore, because $\{\Theta_m = 1\}$ is equivalent to $\{L_m = L_m^- = K\}$,

$$A_{i,j}(1) = \begin{cases} 0, & j = 1,2,\ldots,K-1, \\ \Gamma_{i,K}, & j = K, \end{cases} \qquad i = 1,2,\ldots,K. \tag{15.10}$$

In matrix notation, (15.9) and (15.10) are written as follows:

$$\mathbf{A}(0) = \begin{pmatrix} \Gamma_{1,0} & \Gamma_{1,1} & \cdots & \Gamma_{1,K-1} \\ \Gamma_{2,0} & \Gamma_{2,1} & \cdots & \Gamma_{2,K-1} \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{K,0} & \Gamma_{K,1} & \cdots & \Gamma_{K,K-1} \end{pmatrix}, \qquad \mathbf{A}(1) = \begin{pmatrix} 0 & \cdots & 0 & \Gamma_{1,K} \\ 0 & \cdots & 0 & \Gamma_{2,K} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \Gamma_{K,K} \end{pmatrix}.$$

References

1. N. Shacham and P. Pckenney, Packet recovery in high-speed networks using coding and buffer management, in *Proc. IEEE INFOCOM*, vol. 1, pp. 124–131, 1990.
2. T. Nguyen and A. Zakhor, Multiple sender distributed video streaming, *IEEE Transactions on Multimedia*, vol. 6, no. 2, pp. 315–326, 2004.

3. E. Altman and A. Jean-Marie, Loss probabilities for messages with redundant packets feeding a finite buffer, *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 5, pp. 778–787, 1998.
4. I. Cidon, A. Khamisy, and M. Sidi, Analysis of packet loss processes in high-speed networks, *IEEE Transactions on Information Theory*, vol. 39, no. 1, pp. 98–108, 1993.
5. S. Muraoka, H. Masuyama, S. Kasahara, and Y. Takahashi, Performance analysis of FEC recovery using finite-buffer queueing system with general renewal and Poisson inputs, *Managing Traffic Performance in Converged Networks*, LNCS4516, Springer, pp. 707–718, 2007.
6. P.E. Green, Fiber to the home: The next big broadband thing, *IEEE Communications Magazine*, vol. 42, no. 9, pp. 100–106, 2004.

Chapter 16
Blocking Probabilities of Multiple Classes in IP Networks with QoS Routing

Chia-Hung Wang and Hsing Luh

Abstract We propose a mathematical model for calculating blocking probabilities with optimal bandwidth allocation and QoS routing on multiclass communication networks. This scheme is performed by means of a two-phase procedure. The first step determines optimal paths under network constraints. The second step computes the blocking probability with predetermined optimal solutions. The blocking is due to the failure of meeting the demand of end-to-end paths for each class.

16.1 Introduction

Because of the rapid growth of Internet traffic, aggressive deployment of broadband fiber-optic networks, advance of Voice over IP technology, and the global standardization of IP technology, the telecommunications industry is moving toward a converged network, which uses a single global IP-based packet-switching network to carry all types of network traffic, to replace the traditional separated packet-switching and circuit-switching networks. The international telecommunications standard organizations have decided to adopt this new All-IP network as the base transport network for future development.

The quality of the network must be greatly enhanced to support some applications due to inherent problems of packet-switching networks: long delay time, jitter, and packet loss. In this chapter, we investigate the quality issues, especially analyzing the relationship between the blocking probability and the allocated bandwidth allocation under budget-based end-to-end Quality of Service (QoS) management.

C.-H. Wang
Department of Mathematical Sciences, National Chengchi University, Taipei City 11605, Taiwan
e-mail: 93751502@nccu.edu.tw

H. Luh
Department of Mathematical Sciences, National Chengchi University, Taipei City 11605, Taiwan
e-mail: slu@nccu.edu.tw

W. Yue et al. (eds.), *Advances in Queueing Theory and Network Applications,*

Because it is an NP-hard problem, to our knowledge, this is the first novel and analytically possible approach in relating these two critical factors in QoS management.

We deal with the problem of dimensioning bandwidth for elastic data applications in packet-switched communication networks [1], which can be considered as a multiple-objective optimization model. Users' satisfaction is summarized by means of their achievement functions, and each user is allowed to request more than one type of service. The objective is to determine the amount of required bandwidth for each class to maximize the sum of the users' satisfaction.

We study the blocking probability of this end-to-end transmission system with predetermined optimal solutions, which is an important performance measurement of network systems. We focus on obtaining blocking probabilities after optimally allocating resources with proportional fairness [2]. The blocking is due to the failure of meeting the demand of end-to-end paths for each class.

Bandwidth sharing in a network is frequently evaluated in terms of a utility function [3], [4]. The utility of a connection of class i, $f_i(\theta_i)$, is assumed to be an increasing concave function of its bandwidth θ_i, as introduced by Kelly, Maullco, and Tan [2]. Let K_i be the number of class i connections in progress and denote by $\mu_i(\mathbf{K})$ the overall bandwidth allocation of class i in state $\mathbf{K} = \{K_1, \ldots, K_m\}$. The quantity $\mu_i(\mathbf{K})$ can be obtained from a complicated optimization process. The objective is to realize the allocation that maximizes overall utility. That is, for a given connection population $\mathbf{K}$, to choose $\mu_i(\mathbf{K})$ to maximize

$$\sum_{i=1}^{m} K_i f_i\left(\frac{\mu_i(\mathbf{K})}{K_i}\right). \tag{16.1}$$

Assume every connection of the same class i has the same allocated bandwidth allocation θ_i; that is, $\mu_i(\mathbf{K}) = K_i\theta_i$. Then, $f_i(\mu_i(\mathbf{K})/K_i) = f_i(\theta_i)$. Examples of possible utility functions are $f_i(\theta_i) = \log\theta_i$, leading to the so-called proportional fairness of Kelly et al. [2], and $f_i(\theta_i) = \theta_i^{1-\alpha}/(1-\alpha)$ for $0 < \alpha < \infty$, leading more generally to α-fairness defined by Mo and Walrand [5]. Max-min fairness arises in the limit $\alpha \to \infty$ and proportional fairness corresponds to $\alpha \to 1$. In the limit $\alpha \to 0$, the objective is to maximize overall throughput to the detriment of fairness. More general notions of weighted fairness can be defined by multiplying the utility function by a class-dependent weight.

Roberts [6] provided a survey of recent results on the performance of a network handling elastic data traffic under the assumption that flows are generated as a random process. There are very few analytical results available for the throughput performance of α-fair allocations under random traffic. This is mainly because the performance of these networks is not insensitive and depends significantly on detailed traffic characteristics [7].

QoS routing concerns the selection of a path satisfying the QoS requirements of a connection [8]–[10]. The path selection process involves the knowledge of the connection's QoS requirements and information on the availability of bandwidth [11], [12]. Apostolopoulos and Tripathi [13] characterized the processing cost of

QoS routing algorithms that use the constrained widest–shortest path heuristic to compute QoS paths in a link state-based routing environment. Hernández-Orallo and Vila-Carbo [14] presented an efficient routing scheme for Expedited Forwarding (EF) flow path computation. Kumar and Saraph [15] presented a novel approach to achieve end-to-end QoS support by proposing a new Alliance Network model.

Bandwidth sharing efficiency in overload would be improved if it were possible to perform proactive admission control rather than relying on user impatience to stabilize the system [6]. Admission control consists in rejecting a new connection on its arrival in order to preserve the performance of connections already in progress.

Thus, we focus on the precomputation perspective of QoS routing [16], [17]. This scheme is performed by means of a two-phase procedure [18]. The first step determines optimal paths under network constraints, and the second phase selects an adequate path from predetermined optimal paths when connections arrive. We propose a mathematical model to calculate the available bandwidth of the possible QoS routes, so that the destination host can choose the route that is most likely to satisfy the QoS requirements. After the QoS route has been listed in the routing database, the second phase follows, where we propose a novel algorithm to obtain blocking probabilities in terms of optimal bandwidth.

Computing or estimating blocking probabilities is a fundamental ingredient in network design and engineering [19]. To compute it, the classical approach of Erlang provided a very well-tried solution, and it was perfectly adequate for telephone networks. However, computing blocking probabilities becomes much harder in today's complex networks that carry very heterogeneous traffic [20].

The chapter is organized as follows. In Sect. 16.2, we introduce bandwidth allocation schemes. We derive blocking probabilities with predetermined optimal solutions obtained from the mathematical model in Sect. 16.3. We also analyze the relationship between the blocking probability and the allocated bandwidth allocation in Sect. 16.3. Numerical results are shown in Sect. 16.4 and conclusions are drawn in Sect. 16.5.

16.2 Bandwidth Allocation Schemes

We propose a scheme offering a suitable solution to the network optimization problem. This scheme is performed by means of a two-phase procedure [18]. When handling connection requests, the first phase precomputes paths for a wide range of possible constraints, and the second phase just needs to select an adequate path through an online selecting procedure. The first phase (off-line optimization) is executed in advance and its purpose is to precompute solutions, summarized in a database for later usage. When a connection arrives, the second phase is activated as an online process, and its purpose is to promptly select an adequate solution from the database. The key idea of this two-phase procedure is to effectively reduce the time needed to handle network optimization problems (bandwidth allocation and QoS routing) by performing a certain amount of computation in advance.

16.2.1 Problem Definition

Consider a directed network topology $G = (V, E)$, where V and E denote the set of nodes and the set of links in the network, respectively. There are m (different) QoS classes in this network. Let $E_o \subseteq E$ and $E_d \subseteq E$ be subsets of links connected with the source o and destination d, respectively. Each connection is delivered between the same source o and destination d in the core network. We denote $E_v^{in} \subseteq E$ a subset of incoming links to the node $v \in V$, and we also denote $E_v^{out} \subseteq E$ a subset of outgoing links from the node $v \in V$. The maximal link capacity is U_e on each link $e \in E$. For each link $e \in E$, we use d_e and κ_e to represent average delay and the purchasing cost of bandwidth, respectively. Let $A_{i,j}(e)$ represent the bandwidth allocated to link $e \in E$ for connection j in class i. We use $\chi_{i,j}(e)$ to denote the binary variable that determines whether the link e is chosen for connection j in class i.

The decision variable θ_i is the bandwidth allocated to each connection in class i. In each class i, every connection is allocated the same bandwidth θ_i and has the same QoS requirement. The specific QoS requirements include minimal bandwidth requirement b_i and maximal end-to-end delay constraint D_i for each class i. Assume every connection in class i has the same aspiration level and reservation level of bandwidth, a_i and r_i, and assume that the average number of connections in class i is K_i.

16.2.2 First Phase: A Precomputation Scheme for Network Optimization

Under a limited available budget B, we want to allocate the bandwidth in order to provide each class with maximal possible QoS. The purpose of this work is to show that a methodology that allows the decision maker to explore a set of solutions could satisfy preferences with fairness, and choose the solution which the decision maker finds best.

Using the achievement function interpreted as a measure of QoS [21], we can formulate the mathematical model of the fair bandwidth allocation. Depending on the specified aspiration and reservation levels, a_i and r_i, respectively, Wang and Luh [16], [21] transformed the different QoS measurements onto a normalized scale by using achievement functions. At the first phase, a precomputation-based scheme for network optimization is executed:

$$\max \sum_{i=1}^{m} w_i \log_{\alpha_i} \frac{\theta_i}{r_i}, \tag{16.2}$$

$$s.t. \sum_{e \in E} \sum_{i=1}^{m} \sum_{j=1}^{K_i} \kappa_e A_{i,j}(e) \leq B, \tag{16.3}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{K_i} A_{i,j}(e) \leq U_e, \ \forall e \in E, \tag{16.4}$$

$$A_{i,j}(e) - M\chi_{i,j}(e) \le 0,\ \forall e \in E,\ j = 1,\ldots,K_i,\ i = 1,\ldots,m, \tag{16.5}$$

$$\theta_i - A_{i,j}(e) \le M(1-\chi_{i,j}(e)),\ \forall e \in E,\ j = 1,\ldots,K_i,\ i = 1,\ldots,m, \tag{16.6}$$

$$A_{i,j}(e) - \theta_i \le M(1-\chi_{i,j}(e)),\ \forall e \in E,\ j = 1,\ldots,K_i,\ i = 1,\ldots,m, \tag{16.7}$$

$$\theta_i \ge b_i,\ \forall i = 1,\ldots,m, \tag{16.8}$$

$$\sum_{e \in E_o} A_{i,j}(e) = \theta_i,\ \forall j = 1,\ldots,\ K_i, i = 1,\ldots,m, \tag{16.9}$$

$$\sum_{e \in E_d} A_{i,j}(e) = \theta_i,\ \forall j = 1,\ldots,\ K_i, i = 1,\ldots,m, \tag{16.10}$$

$$\sum_{e \in E_v^{in}} A_{i,j}(e) = \sum_{e \in E_v^{out}} A_{i,j}(e),\ \forall v \in V',\ j = 1,\ldots,K_i,\ i = 1,\ldots,m, \tag{16.11}$$

$$A_{i,j}(e) \ge 0,\ \forall e \in E,\ j = 1,\ldots,K_i,\ i = 1,\ldots,m, \tag{16.12}$$

$$\theta_i \ge 0,\ \forall i = 1,\ldots,m, \tag{16.13}$$

$$\chi_{i,j}(e) \in \{0,1\},\ \forall e \in E,\ j = 1,\ldots,K_i,\ i = 1,\ldots,m, \tag{16.14}$$

where w_i is a fixed weight, M is a sufficiently large number, $\alpha_i = a_i/r_i$, and $V' = V \setminus \{o,d\}$. It is a strictly increasing function of θ_i, having value 1 if $\theta_i = a_i$, and value 0 if $\theta_i = r_i$. The use of the logarithmic function prevents the possibility of assigning zero flow to any user, and on the other hand makes it unprofitable to assign too much flow to the users. Note that this allocation is equivalent to proportionally fair allocation [2].

In this scheme, there is a clear dependence between bandwidth reservation and path selection. This chapter uses the bandwidth and budget as constraints of requirements for feasible path computations. Due to the limited budget on network planning, there exists the budget constraint (16.3). Because the aggregate bandwidth of all connections at any link does not exceed the capacity, we have constraint (16.4).

Constraints (16.5)–(16.8) show that every connection in the same class uses the same bandwidth and has the same bandwidth requirement. Constraints (16.9)–(16.11) express the node conservation relations indicating that flow in equals flow out for every connection j in class i. Although $A_{i,j}(e)$ are continuous variables, constraints (16.9) and (16.10) are flow conservation constraints. Continuous decision variables and binary variables must be nonnegative, shown in constraints (16.12)–(16.14).

Under a limited budget B, we can determine the optimal solutions $A^*_{i,j}(e)$ and θ^*_i which represent the optimal bandwidth allocation for each link e and for each connection of class i. The optimal solution θ^*_i is unique and it provides the proportional fairness to every class. This allocation can provide the fair satisfaction to each user of all classes. Consequently, the bandwidths $K_i\theta^*_i$ are allocated to each class i. Moreover, we determine the maximal bandwidth offered by link e for class i, that is, $\sum_{j=1}^{K_i} A^*_{i,j}(e)$.

Bandwidths are allocated along less expensive paths that connect the origin o and the destination d. After solving the mathematical model in the first phase, we determine the optimal end-to-end path from the source to the destination and introduce a routing database with end-to-end QoS guarantees.

Proposition 16.1. *If $p_{i,j} = \{e \in E |\ \chi^*_{i,j}(e) = 1\}$ for connection j in class i, then path $p_{i,j}$ is the Pareto optimal path from the source o to the destination d.*

Proposition 16.2. *The Pareto optimal end-to-end path $p_{i,j}$ is unique for each connection j in class i.*

The routing database P represents the set of all optimal end-to-end paths obtained from the execution of the first phase.

Definition 16.1. The set of all Pareto optimal paths is called the *routing database* P. That is, $P = \{p_{i,j} |\ p_{i,j}$ is the Pareto optimal path from o to d, $\forall\ j = 1, \ldots, K_i,\ i = 1, \ldots, m\}$.

The routing database P includes the inexpensive routes from the source to the destination on the network.

Definition 16.2. Link e is a *bottleneck link* if the usage of bandwidth achieves its link capacity; that is,

$$\sum_{i=1}^{m} \sum_{j=1}^{K_i} A^*_{i,j}(e) = U_e.$$

Proposition 16.3. *Let $\theta_{i,p} \geq 0$, for each class i, be the bandwidth allocated to each optimal path $p \in P$. Then we have*

$$\sum_{p \in P} \theta_{i,p} = K_i \theta^*_i \tag{16.15}$$

and

$$0 \leq \sum_{i=1}^{m} \theta_{i,p} \leq \min_{e \in p} U_e. \tag{16.16}$$

Proposition 16.4. *A link $e \in p$ is a bottleneck link if*

$$\sum_{i=1}^{m} \theta_{i,p} = \min_{e \in p} U_e. \tag{16.17}$$

From the optimization of the precomputation-based scheme (the first phase), we determine the Pareto optimal bandwidth allocation and a routing database. Next, using the output of the first phase, we execute an online routing scheme (the second phase).

16.2.3 Second Phase: An Online Routing Scheme with End-to-End QoS Guarantees

After the off-line precomputation in the first phase, we determine a reduced network $G' = (V, E')$, where V is the original set of nodes and E' is the subset of links used for each end-to-end path p in the routing database P. Each link $e \in E'$ is characterized by the following values [22]:

(1) Maximal bandwidth $\sum_{j=1}^{K_i} A_{i,j}^*(e)$ which the link e can offer to each connection of class i. After a connection of class i with bandwidth $\theta_i^* \le \sum_{j=1}^{K_i} A_{i,j}^*(e)$ is established through link e, the value of $\sum_{j=1}^{K_i} A_{i,j}^*(e)$ becomes $\sum_{j=1}^{K_i} A_{i,j}^*(e) - \theta_i^*$.
(2) A constant delay d_e, related to the link's speed, propagation delay, and maximal transfer unit.

A connection of class i in the network is characterized by the following values [9]:

(1) The source node o and the destination node d
(2) A mean packet size σ_i of a connection in each class i
(3) An allocated bandwidth θ_i^*
(4) A maximal end-to-end delay constraint D_i

In each class i, an arriving connection should be routed through some path p between the source and destination nodes. We represent by $n(p)$ the number of links of a path $p \in P$. When a connection of class i is routed over a path p with a bandwidth θ_i^*, the following end-to-end delay $D(p)$ applies ([9], [12], [22], etc.),

$$D(p) = \frac{n(p)\sigma_i}{\theta_i^*} + \sum_{e \in p} d_e, \tag{16.18}$$

where σ_i is the mean packet size and d_e is a mean delay on link e related to the link's speed, propagation delay, and maximal transfer unit. A path p between o and d is feasible, for a connection of class i, if $D(p) \le D_i$.

The ability to identify a feasible path for a connection does not yet yield a satisfactory QoS routing solution. In order to supervise multiple connections across the network, the routing algorithm must consider the efficient use of the consumed bandwidth. There does not seem to be a precise definition for the optimality of a path in this context, yet it is clear that an efficient scheme should aim at balancing the loads across the network [9]. A better measure for balancing the loads over the network may be one that aims at seeking a path for which the residual bandwidth of its bottleneck link is maximal [9]. For a path $p \in P$, we represent the residual bandwidth of its bottleneck link by $\widehat{\theta}_{i,p}$; that is,

$$\widehat{\theta}_{i,p} = \min_{e \in p} \{ \sum_{j=1}^{K_i} A_{i,j}^*(e) - \theta_i^* \}. \tag{16.19}$$

For each class i, we present one scheme that aims at balancing the loads across the network. The following scheme is executed as an online procedure in the second phase.

$$\begin{aligned} \max \quad & \widehat{\theta}_{i,p}, \\ s.t. \quad & \widehat{\theta}_{i,p} \le \sum_{j=1}^{K_i} A_{i,j}^*(e) - \theta_i^*, \qquad \forall e \in p, \\ & D(p) \le D_i, \\ & p \in P. \end{aligned} \tag{16.20}$$

The optimization goal of this scheme is to enhance the performance of IP traffic while economically utilizing the bandwidth on communication networks. This scheme is to make more efficient use of bandwidth on the network. Hence, the second phase is to find an optimal path p from the routing database P, maximizing the residual bandwidth of its bottleneck link, $\widehat{\theta}_{i,p}$. That is, the online routing scheme distributes the arriving connection among the predetermined paths so as to avoid overloaded links.

Algorithm Online Routing Scheme (16.20) for Class i:
Input Predetermined optimal solutions: $A^*_{i,j}(e)$, θ^*_i, routing database P.
1. **for** $j \leftarrow 1$ to K_i
2. **for** each $p \in P$
3. **do** $D(p) \leftarrow \dfrac{n(p)\sigma_i}{\theta^*_i} + \sum_{e \in p} d_e$
4. **if** $D(p) \le D_i$ **then**
5. **do** $\widehat{\theta}_{i,p} \leftarrow \min_{e \in p} \left\{ \sum_{n=1}^{K_i} A^*_{i,n}(e) - \theta^*_i \right\}$
6. **else**
7. **do** $P \leftarrow P \setminus \{p\}$
8. **do** $p \leftarrow \arg\max_{p \in P} \{\widehat{\theta}_{i,p}\}$
9. **return** $p_{i,j} \leftarrow p$ and $P \leftarrow P \setminus \{p_{i,j}\}$

The goal of the objective function in the scheme (16.20) is to minimize the network utilization under delay constraints, that is, to keep link utilization low. Therefore, after selecting a path, we can check if the utilization of all paths in the routing database is less than a given limit value, K_i. With the reduced network G', we find the path with bounded delay. This procedure can be repeated until a path is found; otherwise, the connection is blocked.

Selecting a path satisfying the QoS requirements of a new connection is based on the knowledge of the connection's requirements and information about the availability of resources in the network. This information is listed in the routing database by which the optimal path is chosen for each connection. The routing database must be periodically updated and distributed to the ingress routers in order to make an accurate path selection.

16.3 Blocking Probability with Predetermined Optimal Solutions

In real networks, connections do not last forever but arrive at random times and leave the network once the corresponding digital document has been transferred. This results in a random dynamic set of active connections. Moreover, the bandwidth allocation allocated to each connection determines how long that connection will stay active and thus affects the evolution of the set of active connections. A new connection in class i will be dropped if the number of active connections equals the predetermined number of connections, K_i.

The blocking probability is an important performance measurement of the network system [23]. In our situation here, the blocking is due to the failure of setting up the number of end-to-end paths K_i for each class i. In this section, we study the blocking probability of an end-to-end transmission system with predetermined optimal solutions, including optimal bandwidth allocation θ_i^*, and K_i optimal end-to-end paths $p_{i,j}$. At the source node o, connections arrive at random times to enter the core network. The predetermined number K_i in the first phase is used to denote the limit on the number of connections in class i. A new connection in class i cannot enter the source node o and is lost when all K_i end-to-end paths are busy. That is, for each class i, a connection gets dropped on its arrival when the number of connections occupying the end-to-end paths equals K_i. Otherwise, it will be routed through an end-to-end path $p_{i,j}$ with allocated bandwidth θ_i^* predetermined by the off-line scheme in the first phase.

The principal quantity of interest is the blocking probability of different QoS classes, that is, the steady-state probability that all K_i end-to-end paths in class i are busy. Our objective is to estimate these blocking probabilities.

16.3.1 M/G/K/K Blocking Probability Model and System Performance

In the busy period, connections occur as a stationary Poisson process. This results from an assumption that individual connections are independently generated by a large population of users. Wang, Yue, and Luh [23] assumed that connections in class i arrive at the source node o in accordance with independent Poisson processes at rate λ_i, but the packet sizes have a general distribution G with mean σ_i. For each class i, we define $\mu_i = \theta_i^*/\sigma_i$, where θ_i^* is the optimal bandwidth allocation for each connection of class i. The average service time corresponds to the packet transmission time and is equal to average pack size divided by bandwidth. That is,

$$\frac{1}{\mu_i} = \frac{\sigma_i}{\theta_i^*}. \tag{16.21}$$

Hence, for each class i, the service times of connections occupying the end-to-end paths have a general distribution G with mean $1/\mu_i = \sigma_i/\theta_i^*$. Suppose that connections occupy the end-to-end paths in the order they arrive and that packet sizes, which need to be transmitted from o to d, are identically distributed, mutually independent, and independent of the interarrival times.

Under these assumptions, we analyze this end-to-end transmission system as M/G/K/K loss systems [24], that is, Poisson arrivals, general service, K_i end-to-end paths with identical bandwidth allocation θ_i^*, and no waiting space. We can derive the steady-state occupancy probabilities from the Erlang loss system [25]. For each class i, the blocking probability is

$$P_i(K_i) = \frac{1}{K_i!}\left(\frac{\sigma_i\lambda_i}{\theta_i^*}\right)^{K_i}\left[\sum_{j=0}^{K_i}\frac{1}{j!}(\frac{\sigma_i\lambda_i}{\theta_i^*})^j\right]^{-1} \tag{16.22}$$

under conditions of Poisson arrival, general service time, and only K_i end-to-end paths. Equation (16.22) is referred to as Erlang's loss formula [25]. If we denote

$$\rho_i = \frac{\sigma_i\lambda_i}{K_i\theta_i^*}, \tag{16.23}$$

then (16.22) can be rewritten as

$$\begin{aligned} P_i(K_i) &= \frac{(K_i\rho_i)^{K_i}}{K_i!}\left[\sum_{j=0}^{K_i}\frac{(K_i\rho_i)^j}{j!}\right]^{-1} \\ &= \frac{(K_i\rho_i)^{K_i}}{K_i!}\left[\exp(K_i\rho_i) - \mathscr{R}_i(K_i)\right]^{-1}, \end{aligned} \tag{16.24}$$

where $\mathscr{R}_i(K_i)$ is the $K_i$thdegree Taylor remainder term of $\exp(K_i\rho_i)$ [26]. It is valid for all service distributions and only depends on the traffic load, ρ_i. From Taylor's formula with remainder [26], we have the following results.

Proposition 16.5. *There exists a real number* $\xi_i \in (0, K_i\rho_i)$*, such that* $\exp(K_i\rho_i) = \sum_{j=0}^{K_i}(((K_i\rho_i)^j)/j!) + \mathscr{R}_i(K_i)$ *as*

$$\mathscr{R}_i(K_i) = \frac{\exp(\xi_i)(K_i\rho_i)^{K_i+1}}{(K_i+1)!}.$$

Moreover,

$$\lim_{K_i\to\infty}\mathscr{R}_i(K_i) = 0.$$

Harel [27] proved that the fraction of customers lost in the M/G/K/K system is convex in the arrival rate, if the traffic intensity is below some ρ^* and concave if the traffic intensity is greater than ρ^*. Some convexity properties of the blocking probability (16.22) are listed below. These results are consistent with convexity properties shown by Harel [27].

Proposition 16.6. *For each* K_i*, there exists a* ρ_i^* *such that for all* $\rho_i < (>)\rho_i^*$*, the blocking probability (16.22) is strictly convex (concave) in* ρ_i*.*

Proposition 16.7. *The blocking probability (16.22) is strictly decreasing and strictly convex in* θ_i^*/σ_i*, provided* λ_i *and* K_i *are fixed.*

If the allocated bandwidth allocation with objective function (16.1) is insufficient, the effect of bandwidth allocation on the blocking probability will be unstable.

Proposition 16.8.

(1) The blocking probability (16.22) is a decreasing function of θ_i^* if $\theta_i^* > \sigma_i\lambda_i/K_i$.
(2) The blocking probability (16.22) is an increasing function of θ_i^* if $\theta_i^* < \sigma_i\lambda_i/K_i$.

Proof. By Proposition 16.5, for a sufficiently large K_i, the blocking probability can be described as

$$P_i(K_i) = \frac{\hat{\rho}_i^{K_i}}{K_i!\exp(\hat{\rho}_i)},$$

where $\hat{\rho}_i = \sigma_i\lambda_i/\theta_i^*$. Its derivative is

$$\frac{\partial P_i(K_i)}{\partial \theta_i^*} = (\hat{\rho}_i - K_i)\frac{\hat{\rho}_i^{K_i}}{\theta_i^* K_i!\exp(\hat{\rho}_i)}.$$

When $\theta_i^* > \sigma_i\lambda_i/K_i$, we have $\partial P_i(K_i)/\partial\theta_i^* < 0$. That is, $P_i(K_i)$ is a decreasing function of θ_i^* if $\theta_i^* > \sigma_i\lambda_i/K_i$. When $\theta_i^* < \sigma_i\lambda_i/K_i$, we have $\partial P_i(K_i)/\partial\theta_i^* > 0$. That is, $P_i(K_i)$ is an increasing function of θ_i^* if $\theta_i^* < \sigma_i\lambda_i/K_i$. □

Proposition 16.9.

(1) The blocking probability (16.22) is strictly convex in ρ_i if $0 < \theta_i^* < \sigma_i\lambda_i/(K_i + \sqrt{K_i})$ or $\theta_i^* > \sigma_i\lambda_i/(K_i - \sqrt{K_i})$.
(2) The blocking probability (16.22) is strictly concave in ρ_i if $\sigma_i\lambda_i/(K_i + \sqrt{K_i}) < \theta_i^* < \sigma_i\lambda_i/(K_i - \sqrt{K_i})$.

Proof. Given a sufficiently large K_i, by Proposition 16.5, the blocking probability can be described as

$$P_i(K_i) = \frac{\hat{\rho}_i^{K_i}}{K_i!\exp(\rho_i)},$$

where $\hat{\rho}_i = \sigma_i\lambda_i/\theta_i^*$. Then, we have the first derivative of $P_i(K_i)$ with respect to $\hat{\rho}_i$,

$$\frac{\partial P_i(K_i)}{\partial \hat{\rho}_i} = (K_i - \hat{\rho}_i)\frac{\hat{\rho}_i^{K_i-1}}{K_i!\exp(\hat{\rho}_i)}.$$

And its second derivative is

$$\frac{\partial^2 P_i(K_i)}{\partial \hat{\rho}_i^2} = \left[(K_i - \hat{\rho}_i)^2 - K_i\right]\frac{\hat{\rho}_i^{K_i-2}}{K_i!\exp(\hat{\rho}_i)}.$$

The inequalities $0 < \theta_i^* < \sigma_i\lambda_i/(K_i + \sqrt{K_i})$ and $\theta_i^* > \sigma_i\lambda_i/(K_i - \sqrt{K_i})$ imply that $\hat{\rho}_i > K_i + \sqrt{K_i}$ and $0 < \hat{\rho}_i < K_i - \sqrt{K_i}$. In such cases, $\partial^2 P_i(K_i)/\partial\hat{\rho}_i^2 > 0$. On the other hand, the inequality $\sigma_i\lambda_i/(K_i + \sqrt{K_i}) < \theta_i^* < \sigma_i\lambda_i/(K_i - \sqrt{K_i})$ implies $K_i - \sqrt{K_i} < \hat{\rho}_i < K_i + \sqrt{K_i}$. In this case, $\partial^2 P_i(K_i)/\partial\hat{\rho}_i^2 < 0$. □

16.3.2 GI/M/K/K Blocking Probability Model and System Performance

Assume that connections arrive at the source node o in accordance with independent general distributions from outside this end-to-end transmission system. For connections in class i, we assume that successive interarrival times are independent and identically distributed (i.i.d.) and that the packet sizes to be transmitted are i.i.d. exponential random variables. Therefore, this end-to-end transmission system can be analyzed as a GI/M/K/K loss system.

We analyze the GI/M/K/K queue through a combination of the supplementary variable and the embedded Markov chain techniques. We use the former technique to derive closed-form relations between prearrival and arbitrary epoch probabilities and the latter one to obtain prearrival epoch probabilities. Various performance measures such as the average system length and blocking probabilities are discussed and evaluated.

The interarrival times of connections of class i are i.i.d. random variables with cumulative distribution function $A_i(u)$, probability density function $a_i(u)$ for $u > 0$, Laplace–Stiltjes' transform $A_i^*(z)$, and mean $1/\lambda_i$. The packet sizes of every connection in class i are i.i.d. random variables following exponential distribution with mean σ_i. The K_i end-to-end paths (servers) have independent, exponentially distributed service times with common average service time $1/\mu_i = \sigma_i/\theta_i^*$, where θ_i^* is the optimal bandwidth allocation for each connection of class i. That is, the average service time corresponds to the packet transmission time and is equal to the average pack size divided by bandwidth. The service discipline is First-Come First-Served (FCFS) and the maximum number of connections allowed in the system at any time is K_i. The interarrival times and service times are mutually independent. The traffic intensity of the system is

$$\rho_i = \frac{\lambda_i}{K_i \mu_i} = \frac{\lambda_i \sigma_i}{K_i \theta_i^*}.$$

Let $N_i(t)$ be the number of connections in class i present, and let $U_i(t)$ be the remaining inter arrival time of the next arrival in class i. At time t, the state of the system for class i is given by $N_i(t)$ and $U_i(t)$. We define

$$P_{i,n}(u,t)du = P\{N_i(t) = n,\ u \le U_i(t) < u + du\}, \qquad u \ge 0,\ n = 0,1,2,\ldots,K_i.$$

It follows that

$$P_{i,n}(t) = P(N_i(t) = n) = \int_0^\infty P_{i,n}(u,t)du, \qquad n = 0,1,2,\ldots,K_i.$$

For simplicity, we skip the notation i in the following mathematical derivation. That is, for every class i, the derivation is conducted in general format.

From the result of Laxmi and Gupta [28], we can obtain the steady-state probabilities of n ($0 \le n \le K$) connections in the system at prearrival epochs P_n^- by relating

the states of the system at two consecutive time epochs t and $t+dt$ and by using probabilistic arguments. The steady-state probabilities P_n^- can be easily obtained from $P_n(0)$ and are given by

$$P_n^- = \frac{P_n(0)}{\sum_{k=0}^{K} P_k(0)} = \frac{P_n(0)}{\lambda}, \qquad 0 \le n \le K. \tag{16.25}$$

Our objective is to find the distributions of the number of connections in the system at arbitrary (P_n) and prearrival (P_n^-) epochs. Next, we develop the relation between P_n and P_n^- and obtain the latter using the embedded Markov chain technique. After some similar manipulation in [28], we obtain

$$P_{n+1} = \frac{\lambda}{\mu(n+1)} P_n^-, \qquad n = 0, 1, \ldots, K-1. \tag{16.26}$$

Once the P_n^- $(0 \le n \le K)$ are known, one can get P_n $(1 \le n \le K)$ from (16.26). Finally, P_0 is obtained by using $\sum_{n=0}^{K} P_k = 1$.

The state probabilities at prearrival epochs, P_n^-s $(0 \le n \le K)$, can be determined by solving the system of linear equations:

$$P_n^- = \sum_{m=0}^{K} P_m^- P_{m,n}, \qquad 0 \le n \le K, \tag{16.27}$$

where $P_{m,n}$s are the one-step transition probabilities. The expression for $P_{m,n}$ is

$$P_{m,n} = \begin{cases} \displaystyle\int_0^\infty \binom{m+1}{n} e^{-\mu n t}(1-e^{-\mu t})^{m+1-n} dA(t), & 0 \le n \le m < K \\ \displaystyle\int_0^\infty \binom{K}{n} e^{-\mu n t}(1-e^{-\mu t})^{K-n} dA(t), & 0 \le n \le m = K \\ \displaystyle\int_0^\infty e^{-\mu n t} dA(t), & 1 \le m+1 = n \le K \\ 0, & 1 \le m+1 < n \le K. \end{cases} \tag{16.28}$$

We briefly explain (16.28) as follows. Let connections arrive at the epochs $0 = \tau_0, \tau_1, \ldots, \tau_t, \ldots$. The interarrival times $T_{t+1} = \tau_{t+1} - \tau_t > 0$, $t = 0, 1, 2, \ldots$, are i.i.d. random variables with the common distribution function $A(u)$. Let τ_t^- denote the time epochs just before the arrival instant τ_t. Furthermore, at time epoch τ_t^-, a connection arrives and finds the system in state m $(0 \le m < K)$, so that the total number of connections at the instant τ_t is $m+1$. Therefore, if at time epoch τ_{t+1}^-, n $(0 \le n \le K)$ connections are needed, then $(m+1-n)$ connections must depart during the interarrival period.

In the following examples, we determine some quantities of (16.28) for interarrival time distributions: exponential, Erlang-k, deterministic, and hyperexponential.

Example 16.1. Assume the interarrival time is exponentially distributed with parameter λ. Then

$$P_{m,n}=\begin{cases}\dfrac{\lambda\sigma}{\theta^*}\dfrac{(m+1)!\Gamma(n+\lambda\sigma/\theta^*)}{n!\Gamma(m+2+\lambda\sigma/\theta^*)}, & 0\le n\le m<K\\ \dfrac{\lambda\sigma}{\theta^*}\dfrac{K!\Gamma(n+\lambda\sigma/\theta^*)}{n!\Gamma(K+1+\lambda\sigma/\theta^*)}, & 0\le n\le m=K\\ \dfrac{\lambda\sigma}{\theta^*}\dfrac{\Gamma(n+\lambda\sigma/\theta^*)}{\Gamma(n+1+\lambda\sigma/\theta^*)}, & 1\le m+1=n\le K\\ 0, & 1\le m+1<n\le K.\end{cases}\tag{16.29}$$

Example 16.2. Assume the interarrival time is deterministic with mean $1/\lambda=d$,

$$a(t)=\delta(t-d)=\begin{cases}\infty, & t=d\\ 0, & t\ne d.\end{cases}$$

We can determine

$$P_{m,n}=\begin{cases}\dbinom{m+1}{n}e^{-\theta^*nd/\sigma}(1-e^{-\theta^*d/\sigma})^{m+1-n}, & 0\le n\le m<K\\ \dbinom{K}{n}e^{-\theta^*nd/\sigma}(1-e^{-\theta^*d/\sigma})^{K-n}, & 0\le n\le m=K\\ e^{-\theta^*nd/\sigma}, & 1\le m+1=n\le K\\ 0, & 1\le m+1<n\le K.\end{cases}\tag{16.30}$$

Example 16.3. Assume the interarrival time is Erlang-k with mean $1/\lambda$. Then

$$P_{m,n}=\begin{cases}\dbinom{m+1}{n}\left(\dfrac{k\lambda\sigma}{\theta^*}\right)^k\displaystyle\sum_{l=0}^{m+1-n}\dbinom{m+1-n}{l}\frac{(-1)^{m+2-n}}{(m+k\lambda\sigma/\theta^*)^k}, & 0\le n\le m<K\\ \dbinom{K}{n}\left(\dfrac{k\lambda\sigma}{\theta^*}\right)^k\displaystyle\sum_{l=0}^{K-n}\dbinom{K-n}{l}\frac{(-1)^{K+1-n}}{(K-1+k\lambda\sigma/\theta^*)^k}, & 0\le n\le m=K\\ (\dfrac{k\lambda\sigma}{\theta^*})^k\dfrac{(-1)}{(n-1+k\lambda\sigma/\theta^*)^k}, & 1\le m+1=n\le K\\ 0, & 1\le m+1<n\le K.\end{cases}\tag{16.31}$$

Example 16.4. Assume the interarrival time is hyperexponential with k exponential stages and parameters λ_l, p_l, $l=1,\ldots,k$, where $0\le p_l\le 1$, $\lambda_l\ge 0$, for each $l=1,\ldots,k$, $\sum_{l=1}^k p_l=1$, and $1/\lambda=\sum_{l=1}^k p_l/\lambda_l$. Then

$$P_{m,n}=\begin{cases}\dfrac{(m+1)!\sigma}{\theta^*n!}\displaystyle\sum_{l=1}^k\frac{p_l\lambda_l\Gamma(n+\lambda_l\sigma/\theta^*)}{\Gamma(m+2+\lambda_l\sigma/\theta^*)}, & 0\le n\le m<K\\ \dfrac{K!\sigma}{\theta^*n!}\displaystyle\sum_{l=1}^k\frac{p_l\lambda_l\Gamma(n+\lambda_l\sigma/\theta^*)}{\Gamma(K+1+\lambda_l\sigma/\theta^*)}, & 0\le n\le m=K\\ \dfrac{\sigma}{\theta^*}\displaystyle\sum_{l=1}^k\frac{p_l\lambda_l\Gamma(n+\lambda_l\sigma/\theta^*)}{\Gamma(n+1+\lambda_l\sigma/\theta^*)}, & 1\le m+1=n\le K\\ 0, & 1\le m+1<n\le K.\end{cases}\tag{16.32}$$

After obtaining $P_{m,n}$s for various interarrival time distributions, we can obtain the state probabilities at prearrival epochs P_n^-s by solving the system of linear equations (16.27). Then, we know all P_n, $0 \le n \le K$, from (16.26) and $\sum_{n=0}^{K} P_n = 1$.

Performance measures are the means to analyze the efficiency of the queueing system under consideration. Let L denote the average system length. Then it is given by

$$L = \sum_{n=1}^{K} nP_n.$$

Let $\mathrm{E}[W]$ denote the average waiting time in the system. Then by Little's rule

$$\mathrm{E}[W] = \frac{L}{\lambda'},$$

where $\lambda' = \bar{g}(1 - P_{BA})\lambda$ is the effective arrival rate.

16.4 Numerical Results

Consider a sample network shown in Fig. 16.1, where $V = \{$node o, node $1, \ldots,$ node $d\}$ and $E = \{e_{o,1}, e_{o,2}, \ldots, e_{11,d}\}$ denote the set of nodes and the set of links in the network, respectively. Let node o and node d be the source and destination, respectively. Each connection is delivered from node o to node d. Table 16.1 shows the capacity U_e, constant delay ℓ_e, and the purchasing cost κ_e of bandwidth for each link $e \in E$.

Four different QoS classes are given (characterized and shown in Table 16.2), where class 1 has the highest priority and class 4 has the lowest priority. We assume every connection in class i, for $i = 1, \ldots, 4$, has the same aspiration level a_i kbps (i.e., kilobits/sec), reservation level r_i kbps, mean packet size σ_i kb, maximal end-to-end delay D_i, and bandwidth requirement b_i kbps. We let θ_i be the bandwidth allocated to each connection in class i, $\forall\, i = 1, \ldots, 4$. Let K_i be the number of connections in each class i for $i = 1, \ldots, 4$.

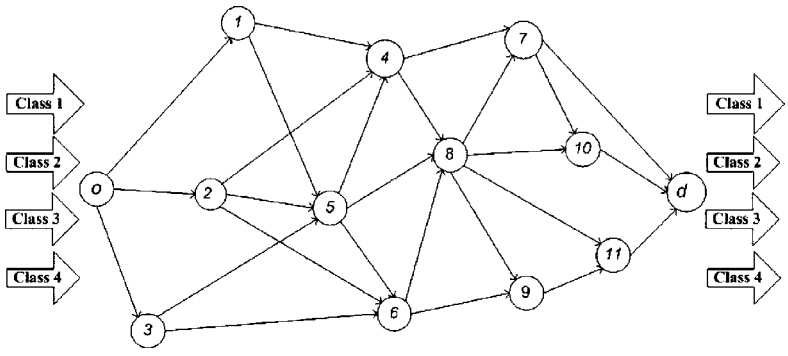


Fig. 16.1 A sample network.

Table 16.1 Characteristics of each link.

Characteristics	$e_{o,1}$	$e_{o,2}$	$e_{o,3}$	$e_{1,4}$	$e_{1,5}$	$e_{2,4}$	$e_{2,5}$	$e_{2,6}$	$e_{3,5}$
Capacity (Mbps)	35	45	55	53	47	36	37	45	40
Cost ($)	7	6	5	14	11	14	7	13	8
Delay (ms)	3	3.2	3.5	1.2	2	1.2	3	1.5	2.7
	$e_{3,6}$	$e_{5,4}$	$e_{5,8}$	$e_{5,6}$	$e_{4,7}$	$e_{4,8}$	$e_{6,8}$	$e_{6,9}$	$e_{8,7}$
Capacity (Mbps)	50	45	46	45	44	46	36	35	54
Cost ($)	14	7	11	5	5	10	5	7	5
Delay (ms)	1.2	3	2	3.5	3.5	2.2	3.5	3	3.5
	$e_{8,10}$	$e_{8,11}$	$e_{8,9}$	$e_{9,11}$	$e_{7,d}$	$e_{7,10}$	$e_{10,d}$	$e_{11,d}$	
Capacity (Mbps)	40	53	41	40	52	44	42	50	
Cost ($)	7	9	6	8	13	6	8	6	
Delay (ms)	3	2.5	3.2	2.7	1.5	3.2	2.7	3.2	

Table 16.2 Characteristics of each QoS class.

Class i	b_i (kbit/s)	r_i (kbit/s)	a_i (kbit/s)	σ_i (kb)	D_i (ms)
1	512	622	1024	2534.2	10.2
2	155	167	512	367.8	19.7
3	45	83	256	128.7	25.4
4	34	38	56	47.1	41.3

16.4.1 Predetermined Optimal Solutions

Under the total available budget $B = 2 \times 10^6$, we plan to allocate the bandwidths in order to provide each class with maximal utility. We provide a routing table as shown in Table 16.3 given parameters $(K_1, K_2, K_3, K_4) = (20, 35, 60, 90)$ and $(w_1, w_2, w_3, w_4) = (0.4, 0.3, 0.2, 0.1)$. The optimal bandwidth allocation is $\theta_1^* = 1024$, $\theta_2^* = 448$, $\theta_3^* = 189$, and $\theta_4^* = 56$. In Table 16.3, it gives, for each path p in the routing table P, the path flow $\theta_{i,p}$ which is computed by (16.15) in Proposition 16.3.

Moreover, it gives the number of connections and number of links $n(p)$ along path p. When connections arrive, these paths are the candidates for the adequate solution with end-to-end QoS guarantees. We can determine the unit path cost $\sum_{e \in p} \kappa_e$ for using one-unit bandwidth along the path $p \in P$. These paths are Pareto optimal solutions with end-to-end QoS guarantees. The path flow $\theta_{i,p}$ in (16.15) and (16.16), for each class i, is the aggregated bandwidth of connections along path p. The number of connections, for each class, along path $p \in P$ is also determined. A path $p_{i,j}$ between o and d is guaranteed if $D(p_{i,j}) \leq D_i$ for a connection j in class i.

We now explore how changes in the total budget affect the optimal allocation. Some numerical results are depicted in Figs. 16.2 to 16.4. Figure 16.3 shows an obvious phenomenon that increasing the total budget will increase the total satisfaction. It is also reflected by bandwidth obtained as shown in Fig. 16.2.

Table 16.3 A routing table as $B = 2 \times 10^6$, $(K_1, K_2, K_3, K_4) = (20, 35, 60, 90)$, and $(w_1, w_2, w_3, w_4) = (0.4, 0.3, 0.2, 0.1)$.

Class i	Opt. Path p	Path Flow $\theta_{i,p}$	No. of Connect.	No. of Links $n(p)$	Unit Path Cost	Delay $D(p)$
1	$e_{o,1} - e_{1,4} - e_{4,7} - e_{7,d}$	6144	6	4	39	9.2
	$e_{o,2} - e_{2,4} - e_{4,7} - e_{7,d}$	14336	14	4	38	9.4
	$e_{o,1} - e_{1,4} - e_{4,7} - e_{7,d}$	448	1	4	39	9.2
	$e_{o,2} - e_{2,4} - e_{4,7} - e_{7,d}$	2688	6	4	38	9.4
2	$e_{o,2} - e_{2,5} - e_{5,4} - e_{4,7} - e_{7,d}$	448	1	5	38	14.2
	$e_{o,2} - e_{2,5} - e_{5,6} - e_{6,8} - e_{8,10} - e_{10,d}$	448	1	6	38	18.9
	$e_{o,2} - e_{2,5} - e_{5,6} - e_{6,8} - e_{8,11} - e_{11,d}$	11200	25	6	38	18.9
	$e_{o,2} - e_{2,5} - e_{5,6} - e_{6,8} - e_{8,7} - e_{7,d}$	448	1	6	41	18.2
	$e_{o,2} - e_{2,4} - e_{4,7} - e_{7,d}$	3213	17	4	38	9.4
3	$e_{o,2} - e_{2,5} - e_{5,6} - e_{6,8} - e_{8,10} - e_{10,d}$	378	2	6	38	18.9
	$e_{o,2} - e_{2,5} - e_{5,6} - e_{6,8} - e_{8,11} - e_{11,d}$	7749	41	6	38	18.9
	$e_{o,2} - e_{2,4} - e_{4,7} - e_{7,d}$	2128	38	4	38	9.4
4	$e_{o,2} - e_{2,5} - e_{5,6} - e_{6,8} - e_{8,10} - e_{10,d}$	112	2	6	38	18.9
	$e_{o,2} - e_{2,5} - e_{5,6} - e_{6,8} - e_{8,11} - e_{11,d}$	2800	50	6	38	18.9

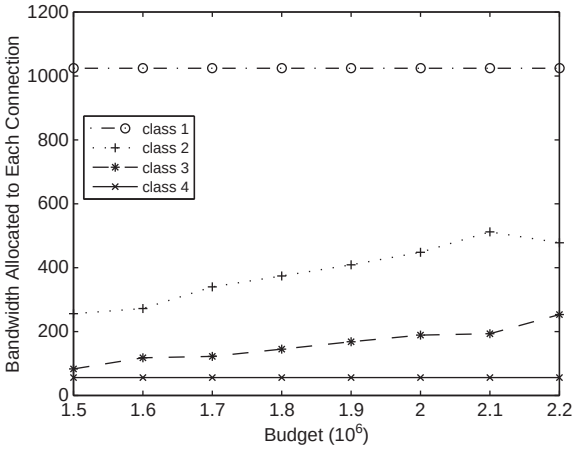


Fig. 16.2 Bandwidth versus budget.

16.4.2 Blocking Probabilities Under M/G/K/K Model

We observe the relationship between blocking probability $P_i(K_i)$ and average arrival rate λ_i. We assume connections of class i arrive at the source node o in accordance with independent Poisson processes at rate λ_i, and the packet sizes to be transmitted have general distributions with mean σ_i. From (16.22), we can determine the blocking probabilities with parameters $\theta_1^* = 1024$, $\theta_2^* = 448$, $\theta_3^* = 189$, $\theta_4^* = 56$, $\sigma_1 = 2534.2$, $\sigma_2 = 367.8$, $\sigma_3 = 128.7$, $\sigma_4 = 47.1$, $K_1 = 20$, $K_2 = 35$, $K_3 = 60$, and $K_4 = 90$. Figure 16.5 shows that class 1 has higher blocking probability than

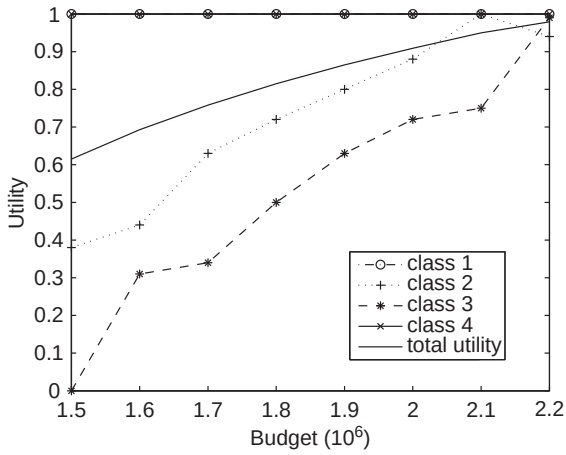


Fig. 16.3 Utility versus budget.

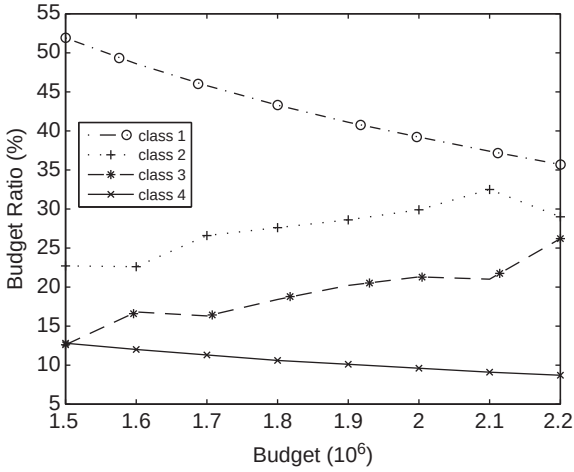


Fig. 16.4 Budget ratio versus budget.

other classes due to the number of their estimated arrivals, K_1 being smaller and the packet transmission, being time (16.21) being longer than that of other classes. From Proposition 16.6, there exists an inflection point for each curve in Fig. 16.5. These curves are convex ahead of inflection points.

Next, we observe the the relationship between blocking probability $P_i(K_i)$ and total budget B by using the formula (16.24) in the M/G/K/K model and optimal solutions in Table 16.4. The numerical results are drawn in Fig. 16.6 given mean arrival rates $\lambda_1 = K_1 = 20$, $\lambda_2 = K_2 = 35$, $\lambda_3 = K_3 = 60$, $\lambda_4 = K_4 = 90$, and mean packet size $\sigma_1 = 2534.2$, $\sigma_2 = 367.8$, $\sigma_3 = 128.7$, and $\sigma_4 = 47.1$. We observe that the effect on blocking probability in class 2 is unstable when $1.5 \times 10^6 < B < 1.8 \times 10^6$, and

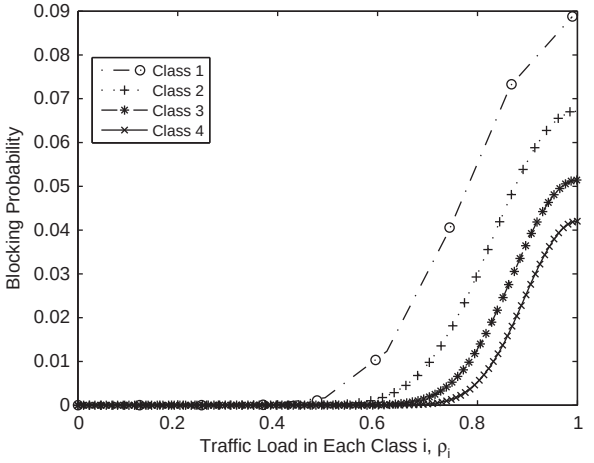


Fig. 16.5 Blocking probability versus traffic load with four classes.

Table 16.4 Change in the budget with $(K_1, K_2, K_3, K_4) = (20, 35, 60, 90)$ and $(w_1, w_2, w_3, w_4) = (0.4, 0.3, 0.2, 0.1)$.

Budget B (10^6)	Bandwidth $(\theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*)$	Utility (f_1, f_2, f_3, f_4)	Total Flow (kbps)	Utility $\sum_{i=1}^{4} w_i f_i$	Budget Ratio %	CPU Time (sec)
1.5	(1024,256,83,56)	(1,0.38,0,1)	34494	0.615	(51.9,22.7,12.6,12.8)	311.63
1.6	(1024,272,118,56)	(1,0.44,0.31,1)	35049	0.693	(48.6,22.6,16.8,12.0)	87.41
1.7	(1024,340,122,56)	(1,0.63,0.34,1)	37403	0.758	(45.8,26.6,16.3,11.3)	290.44
1.8	(1024,374,145,56)	(1,0.72,0.50,1)	38610	0.815	(43.3,27.6,18.4,10.6)	139.91
1.9	(1024,409,168,56)	(1,0.80,0.63,1)	39818	0.865	(41.1,28.6,20.2,10.1)	9981.41
2.0	(1024,448,189,56)	(1,0.88,0.72,1)	41215	0.909	(39.2,29.9,21.3,9.6)	350.91
2.1	(1024,512,193,56)	(1,1,0.75,1)	43440	0.950	(37.4,32.5,21.0,9.1)	1781.72
2.2	(1024,478,253,56)	(1,0.94,0.99,1)	42233	0.979	(35.7,29.0,26.2,8.7)	613.13

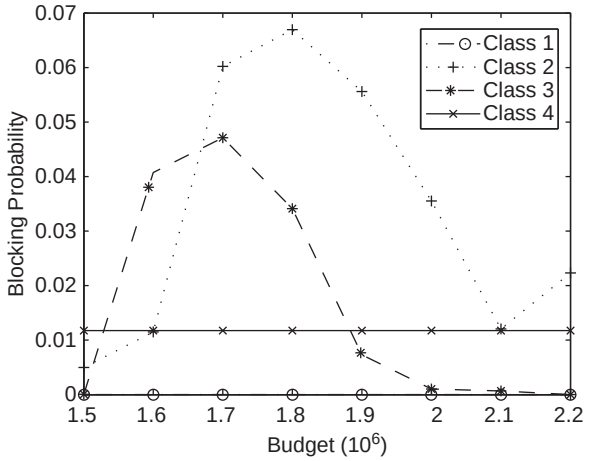


Fig. 16.6 Blocking probability versus total budget under M/G/K/K model.

the effect on blocking probability in class 3 is also unstable when $1.5 \times 10^6 < B < 1.7 \times 10^6$. These unstable situations occur due to insufficient bandwidth allocation to class 2 and class 3 by Proposition 16.8. If the bandwidth allocation is sufficiently large, by Proposition 16.8, the blocking probability will decrease as the allocated bandwidth increases. Computational experiences show different topologies do not change the validity of the model.

16.5 Conclusions

We present an approach for a two-phase modeling of QoS routing in communication networks. The first phase is executed in advance and its purpose is to precompute solutions as a database for later use. The second phase selects one of the solutions precomputed at the first phase by performing a few additional computations. The purpose of the second phase is to promptly provide an adequate solution when connections arrive. Users' utility functions are summarized by means of achievement functions. Using the bandwidth allocation model, we can find a Pareto optimal allocation of bandwidth on the network under a limited available budget, and this allocation can provide the so-called proportional fairness to every class. That is, this allocation can provide similar satisfaction to each user in all classes.

We also derive the blocking probability of an end-to-end transmission system with predetermined optimal solutions, which is an important performance measurement of network systems.

Acknowledgments The authors are grateful to the referees for helpful remarks and suggestions. Special thanks to Professor Yew-Sing Thomas Lee for useful discussions while Chia-Hung Wang visited Department of Information and Decision Sciences, University of Illinois at Chicago. This research was supported in part by the National Science Council, Taiwan, R.O.C., under grant numbers NSC 95-2221-E-004-007 and NSC 96-2917-I-004-015.

References

1. W. Ogryczak, T. Śliwiński, and A. Wierzbicki, Fair resource allocation schemes and network dimensioning problems, *Journal of Telecommunications and Information Technology*, vol. 3, pp. 34–42, 2003.
2. F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, Rate control for communication networks: Shadow prices, proportional fairness and stability, *Journal of the Operational Research Society*, vol. 49, no. 3, pp. 237–252, 1998.
3. A. W. Berger and Y. Kogan, Dimensioning bandwidth for elastic traffic in high-speed data networks, *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 643–654, 2000.
4. S. van Hoesel, Optimization in telecommunication networks, *Statistica Neerlandica*, vol. 59, no. 2, pp. 180–205, 2005.
5. J. Mo and J. Walrand, Fair end-to-end window-based congestion control, *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 556–567, 2000.

6. J. W. Roberts, A survey on statistical bandwidth sharing, *Computer Networks*, vol. 45, no. 3, pp. 319–332, 2004.
7. T. Bonald and A. Proutière, Insensitive bandwidth sharing in data networks, *Queueing Systems*, vol. 44, no. 1, pp. 69–100, 2003.
8. R. A. Guérin and A. Orda, QoS routing in networks with inaccurate information: Theory and algorithms, *IEEE/ACM Transactions on Networking*, vol. 7, no. 3, pp. 350–364, 1999.
9. A. Orda, Routing with end-to-end QoS guarantees in broadband networks, *IEEE/ACM Transactions on Networking*, vol. 7, no. 3, pp. 365–374, 1999.
10. C. Pornavalai, G. Chakraborty, and N. Shiratori, QoS routing algorithms for pre-computed paths, in *Proc. 6th International Conference on Computer Communications and Networks*, pp. 248–251, 1997.
11. J. Gozdecki, A. Jajszczyk, and R. Stankiewicz, Quality of service terminology in IP networks, *IEEE Communications Magazine*, vol. 41, no. 3, pp. 153–159, 2003.
12. I. Atov, H. T. Tran, and R. J. Harris, OPQR-G: Algorithm for efficient QoS partition and routing in multiservice IP networks, *Computer Communications*, vol. 28, no. 18, pp. 1987–1996, 2005.
13. G. Apostolopoulos and S. K. Tripathi, On the effectiveness of path pre-computation in reducing the processing cost of on-demand QoS path computation, in *Proc. 3rd IEEE Symposium on Computers and Communications*, pp. 42–46, 1998.
14. E. Hernández-Orallo and J. Vila-Carbo, Efficient QoS routing for differentiated services EF flows, in *Proc. 10th IEEE Symposium on Computers and Communications*, pp. 91–96, 2005.
15. N. Kumar and G. Saraph, End-to-end QoS in interdomain routing, in *Proc. International conference on Networking and Services*, pp. 82–82, 2006.
16. C.-H. Wang and H. Luh, A precomputation-based scheme for QoS routing and fair bandwidth allocation, in *Proc. 13th Annual IEEE International Conference on High Performance Computing*, 2006, *Lecture Notes in Computer Science*, vol. 4297, pp. 595–606, 2006.
17. C.-H. Wang and H. Luh, A fair QoS scheme for bandwidth allocation by precomputation-based approach, *International Journal of Information and Management Sciences*, vol. 19, no. 3, pp. 391–412, 2008.
18. C.-H. Wang and H. Luh, Two-phase modeling of QoS routing in communication networks, in *Proc. 16th International Conference on Computer Communications and Networks*, pp. 1210–1216, 2007.
19. A. Faragó, Efficient blocking probability computation of complex traffic flows for network dimensioning, *Computers and Operations Research*, vol. 35, no. 12, pp. 3834–3847, 2007.
20. T. Bonald, L. Massoulié, A. Proutière, and J. Virtamo, A queueing analysis of max-min fairness, proportional fairness and balanced fairness, *Queueing Systems*, vol. 53, no. 1-2, pp. 65–84, 2006.
21. C.-H. Wang and H. Luh, Network dimensioning problems of applying achievement functions, in: X.-S. Zhang, D.-G. Liu and L.-Y. Wu (Eds.), *Operations Research and Its Applications, Lecture Notes in Operations Research*, vol. 6, pp. 35–59, 2006.
22. R. Johari and D. K. H. Tan, End-to-end congestion control for the Internet: Delays and stability, *IEEE/ACM Transactions on Networking*, vol. 9, no. 6, pp. 818–832, 2001.
23. C.-H. Wang, W. Yue, and H. Luh, Performance evaluation of predetermined bandwidth allocation for heterogeneous networks, *Technical Report of IEICE*, vol. 107, no. 6, pp. 37–42, 2007.
24. D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Eaglewood cliffs NJ: Prentice Hall, 1992.
25. S. M. Ross, *Stochastic Processes*. New York: Wiley, 1983.
26. T. M. Apostol, *Mathematical Analysis*, 2nd ed. Reading, MA: Addison-Wesley, 1974.
27. A. Harel, Convexity properties of the Erlang loss formula, *Operations Research*, vol. 38, no. 3, pp. 499–505, 1990.
28. P. Vijaya Laxmi and U. C. Gupta, Analysis of finite-buffer multi-server queues with group arrivals: GI^X/M/c/N, *Queueing Systems*, vol. 36, pp. 125–140, 2000.

About the Editors and Authors

Editors

Hideaki Takagi is a professor at the Graduate School of Systems and Information Engineering of the University of Tsukuba, Japan. He received, his B.S. and M.S. in Physics from the University of Tokyo, and Ph.D. in Computer Science from the University of California, Los Angeles. Prior to his current position, he was Consultant Researcher at IBM Research, Tokyo Research Laboratory, Professor and Chair at the Institute of Policy and Planning Sciences of the University of Tsukuba, and Vice President of the University of Tsukuba.

Dr. Hideaki Takagi's research interest includes enumerative combinatorics, probability models, and performance evaluation of computer and communication systems. He is an IEEE Fellow and IFIP Silver Core holder.

Yutaka Takahashi received his B.Eng., M.Eng., and Dr.Eng. degrees from Kyoto University, Kyoto, Japan. He was with the Department of Applied Mathematics and Physics, Faculty of Engineering, Kyoto University from 1980 to 1995 and with the Department of Applied Systems Science at the same faculty from 1995 to 1996. From 1996 to 1999, he was a professor at the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Nara, Japan. Since April 1999, he has been with the Department of Systems Science, the Graduate School of Informatics, Kyoto University, as a professor. From 1983 to 1984 he was with the Institut National de Recherche en Informatique et en Automatique (INRIA), France, as an Invited Professor. He was a cochairman of IFIP TC6 WG6.3 on Performance of Communication Systems from 1992 to 2002 and is an elected full member of IFIP TC6 WG7.3 on Computer Performance Evaluation as well as WG6.3. He is also an associate editor of *Telecommunication Systems*, an area editor of *Mobile Networks & Applications*, an editor of *Wireless Network Journal (WINET)*, and on the editorial board of *Journal of Networks*. He served as an International Advisory Committee member of *NIS* (*Networking and Information Systems*) *Journal* and the project leader for the Kobe Multi-node Integrated

Connection Research Center established by the Telecommunications Advancement Organization of Japan (TAO). He was awarded the Silver Core from IFIP in 2001 and is a fellow of the Operations Research Society of Japan.

His research interests include queueing theory and its application to performance analysis of computer communication systems as well as database systems. Dr. Takahashi is a member of the Institute of Electronics, Information and Communication Engineers, the Information Processing Society of Japan, the Operations Research Society of Japan, the Institute of Systems, Control, and Information Engineers, and the Japan Society for Industrial and Applied Mathematics.

Wuyi Yue received her B.Eng. degree in Electronic Engineering from Tsinghua University, China, and the M.Eng. and Dr.Eng. degrees in Applied Mathematics and Physics from Kyoto University, Japan. She was a researcher and a chief researcher of ASTEM RI, an associate professor of Wakayama University, an associate professor and a professor at the Department of Applied Mathematics, and a professor at the Department of Information Science and Systems Engineering, Konan University, Japan. She is currently a professor at the Department of Intelligence and Informatics, Konan University, Japan. She is also the director of Institute of Intelligent Information and Communications Technology, Konan University.

Dr. Yue is a member of the IEEE, the IEICE of Japan, the System Engineers Society of China, and the Operations Research Society of China, and is a fellow of the Operations Research Society of Japan. She has served at plural international conferences and symposia as chair (co-chair) of the organizing committee, technical program committee, steering committee, and local committee, member of the technical program committee and of the program committee. She is also on the editorial of board of Journal of Industrial and Management Optimization, Journal of International Association of Engineers, Journal of Nonlinear Dynamics and Systems Theory, Journal of Dynamics of Continuous, Discrete and Impulsive Systems, and other six journals.

Dr. Yue's research interests include queueing theory, stochastic processes and optimal methods as applied to system modeling, performance analysis and evaluation, and optimal resource allocation of wired and wireless/mobile communication networks (including mobile cellular, multihop, multitraffic mobile communication networks), multimedia communication networks, traffic systems, stochastic systems, information systems, systems engineering, and operations research.

Authors

Jung Woo Baek received the B.S. and degrees in industrial engineering from Sungkyunkwan University, Seoul Korea. Currently he is a Ph.D student at the same department. His research interests include queueing theory and stochastid modeling.

Kuo-Hwa Chang received his B.S. degree in Applied Mathematics from National Chiao Tung University, Taiwan, and M.S. degree in Operations Research and Ph.D. degree in Industrial and Systems Engineering from the Georgia Institute of Technology, USA. He is now a professor in the Department of Industrial and Systems Engineering in Chung Yuan Christian University, Taiwan.

Dr. Chang's current research interests include queueing system analysis on production systems and financial engineering.

Sophie Hautphenne received her B.A. and M.Sc. degrees in mathematical sciences from the Université Libre de Bruxelles, Brussels, Belgium. Now Sophie Hautphenne is a Ph.D. candidate in applied probability at the Department of Informatics, Université Libre de Bruxelles, Brussels, Belgium.

Sophie Hautphenne's research interests include matrix analytic methods and numerical methods in branching processes.

Tetsuji Hirayama received his B.Econ. degree from Chuo University, Tokyo, Japan, and the M.Sc. and Ph.D. degrees in socioeconomic planning from the University of Tsukuba, Ibaraki, Japan. Now Dr. Hirayama is an assistant professor at the Graduate School of Systems and Information Engineering, University of Tsukuba, Ibaraki, Japan.

Dr. Hirayama's research interests include queueing theory; modeling, analysis and optimization of stochastic systems; and modeling and performance evaluation of computer communication networks.

Dr. Hirayama is a member of the Operation Research Society of Japan.

Shunfu Jin received her B.Eng. degree in computer and applications from North East Heavy Machinery College, Qiqihaer, China, the M.Eng. degree in computer science and Dr.Eng. degree in circuits and systems from Yanshan University, Qinhuangdao, China. Now Dr. Jin is a professor at the College of Information Science and Engineering, Yanshan University, Qinhuangdao, China.

Dr. Jin's research interests include stochastic modeling for telecommunications, performance evaluation for computer systems and networks, and applications for queueing systems.

Shoji Kasahara received his B.Eng., M.Eng., and Dr.Eng. degrees from Kyoto University, Kyoto, Japan, in 1989, 1991, and 1996, respectively. He was with the Educational Center for Information Processing, Kyoto University from 1993 to 1997. In 1996, he was a visiting scholar of University of North Carolina at Chapel Hill, NC, USA. From 1997 to 2005, he was with the Department of Information Systems, Graduate School of Information Science, Nara Institute of Science and Technology. Since 2005, he has been an Associate Professor at the Department of Systems Science, Graduate School of Informatics, Kyoto University.

His research interests include queueing theory and performance analysis of computer and communication systems. Dr. Kasahara is a member of the IEEE, the Institute of Electronics, Information and Communication Engineers (IEICE), the Operations Research Society of Japan, the Information Processing Society of Japan, and the Institute of Systems, Control and Information Engineers.

Kenji Kirihara received his B.Eng. Degree from the Faculty of Engineering, Kyoto University, Japan, in 2007. Since 2007, he has been a master course student at the Department of Systems Science, Graduate School of Informatics, Kyoto University. His research interests include queueing theory and performance analysis of network systems. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE).

Ho Woo Lee received his B.S. degree in industrial engineering from Seoul, National University, Seoul, Korea, and M.S. and Ph.D degrees in industrial and systems engineering from The Ohio State Univeristy, Columbus, Ohio, USA.

Currently Dr. Lee is a professor of systems management engineering and the graduate chairman of Industrial Engineering, Sungkyukwan University, Seoul, Korea. He is also the director of the Research & HRD Group for Management of Complex Systems funded by Brain Korea-21 project of Korean government.

He was the vice-chair of the Korean Institute of Industrial Engineers (KIIE) and the chief editor of the *Journal of KIIE*. Currently he serves as an editor of the *Journal of Applied Mathematics and Stochastic Analysis* (*JAMSA*).

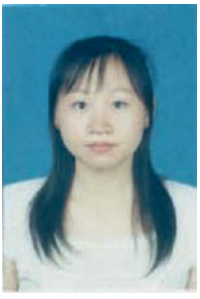
Dr. Lee's research interests include queueing theory, stochastic modeling, and operations research.

Se Won Lee received the B.S., M.S. and Ph.D degrees in industrial engineering from Sungkyunkwan University, Seoul Korea. Currently Dr. Lee is a BK-21 postdoc researcher at the same department.

Dr. Lee's research interests include queueing theory and stochastic modeling.

Kenji Leibnitz received his M.Sc. and Ph.D. degrees in computer science from the University of Würzburg in Germany, where he was also a research assistant at the Institute of Computer Science, Department of Distributed Systems. In 2004, he joined the Graduate School of Information Science and Technology at Osaka University, Japan, as a postdoctoral research fellow and since July 2006 he has been a specially appointed associate professor at the Advanced Network Architecture Laboratory in the Department of Information Networking, working within the framework of the Yuragi Super-COE Project.

His main research interest lies on the performance modeling of communication networks, especially in the application of biologically inspired mechanisms to self-organization and self-adaptation of overlay and sensor networks. He is a member of the IEICE.



Chunyan Li received her B.Sci. degree in information and computing science and the M.Sci. degree in operations research and cybernetics from Yanshan University, Qinhuangdao, China. Now she is an assistant professor at the Department of Science, Zhijiang College, Zhejiang University of Technology, Hangzhou, China.

Chunyan Li's research interests include performance analysis of queueing systems and reliability analysis of repairable systems.

Yang-Shu Lu currently is a Ph.D. student in the Department of Industrial and Systems Engineering, Chung Yuan Christian University, Taiwan. He is working on the queueing system analysis on production systems.

Hsing Luh is a professor at the Department of Mathematical Sciences, National Chengchi University. He received his Ph.D. in Operations Research at North Carolina State University, USA, 1992 and has served as the president of the Operations Research Society of Taiwan, 2005–2007. His research interests include queueing theory and applications, stochastic models and simulations, linear programming, and optimization.

Dr. Hsing Luh's research accomplishments have been recognized by the International Research Award and the Excellent Research Lecture Award of National Chengchi University, and *Marquis' Who's Who in Science and Engineering* since 2003.

Zhanyou Ma received his M.S. degree in operations research and cybernetics and the Ph.D. degree in management science and engineering from Yanshan University, Qinhuangdao, China. Now Dr. Ma is an assistant professor in the College of Sciences, Yanshan University, Qinhuangdao, China.

Dr. Ma's research interests include queueing systems with vacations and performance evaluation models in communication networks.

Hiroyuki Masuyama received his B.Eng. Degree from the Faculty of Engineering, Kyoto University, Japan, in 1999. He received the M.I. and D.I. degrees from the Graduate School of Informatics, Kyoto University, Kyoto, Japan, in 2001, and 2004, respectively. Since 2004, he has been an assistant professor at the Department of Systems Science, Graduate School of Informatics, Kyoto University.

His current research interests include queueing theory, especially, asymptotic analysis of rare events. He is a member of the Operations Research Society of Japan (ORSJ). Dr. Masuyama received the Best Paper Award for Young Researchers from ORSJ in 2007.

Masakiyo Miyazawa received his B.Sci., M.Sci., and Dr.Sci. degrees in applied physics from the Tokyo Institute of Technology. Dr. Miyazawa is a professor at the Department of Information Sciences, Tokyo University of Science.

Dr. Miyazawa's research interests cover asymptotic behaviors of queueing networks, the matrix analytic method for fluid queues, optimal control of queues, and invariance and limiting behaviors in stochastic processes.

Yoshihori Ozaki was born in 1984 in Ehime Prefecture, Japan. He received his bachelor's degree in policy and planning sciences from the University of Tsukuba in March 2006. He is currently with the Master's Program in social systems engineering, Graduate School of Systems and Information Engineering, University of Tsukuba.

His research interests include performance modeling and analysis of mobile communication networks.

No Ik Park received the B.S., M.S. and Ph.D degrees in industrial engineering from Sungkyunkwan University, Seoul Korea. Currently Dr. Park is a researcher in BcN Research Division, ETRI, Daejon, Korea.

Dr. Park's research interests include queueing theory and communication network.

Marie-Ange Remiche received her Ph.D. in 1999 at the Université Libre de Bruxelles, Belgium. After completing a post-doc at RWTH-Aachen, Germany, as a Marie Curie Fellow, she became an associate professor in networking at the Faculty of Engineering at the Université Libre de Bruxelles.

Her main interests are in the field of performance evaluation of telecommunication networks, with a particular interest in queueing theory and matrix analytic methods.

Yutaka Sakuma received his B.Sci., M.Sci., and Dr.Sci. degrees at the Department of Information Sciences from Tokyo University of Science. Dr. Sakuma is an assistant professor at the Department of Information Sciences, Tokyo University of Science.

Dr. Sakuma's research interests include fluid queues, queueing networks, asymptotic behavior for queues, the matrix analytic method, and shortest join queues.

Dr. Hideaki Takagi (see “Editors”.)

Dr. Yutaka Takahashi (see “Editors”.)

A. M. K. Tarabia is currently an associate professor in the Department of Mathematics at Damietta Faculty of Science, Egypt. He was born in Damietta, Egypt, in 1962. He received his B.ESc and B.Sc. and M.Sc. degrees in mathematics from Mansoura University, in 1985, 1989, and 1993, respectively. Further he completed his Ph.D. from Indian Institute of Technology, Delhi, in 1998. Since August 2002, Dr. Tarabia has been traveling to various universities in Egypt and KSA and teaching various courses.

Dr. Tarabia has contributed significantly in the area of modeling and analysis of continuous and discrete-time queueing systems. He has published several research articles in various journals such as *Stochastic Processes and its Applications*, *Mathematical Scientist* (*Journal of Applied Probability*), *Applied Mathematics and Computation*, *An International Journal Computer & Mathematics with Applications*, *Applied Mathematical Modelling*, *Mathematical Methods of Operations Research* (*MMOR*), *Sankhya, Journal of the Operational Research Society of India* (*Opsearch*) and others. Also, he has authored many statistics and mathematics books in the Arabic language.

Naishuo Tian currently is a professor of mathematics in the College of Science, Yanshan University, China. He is a senior commissioner of queueing theory specialty, Operational Research Committee, China. He has supervised many graduate students since 2000.

Prof. Naishuo Tian’s research interests are queueing theory, stochastic models, and their applications. He is the author of a great deal of research studies published by important international and national journals, conference proceedings, as well as books, one of which has been edited by Kluwer Academic Publishers.

Hiroshi Toyoizumi received his B.Science and M.Science degrees in physics, and Dr.Eng. degree in applied physics from Waseda University, Tokyo, Japan. Now Dr. Toyoizumi is a professor at the Graduate School of Accounting and Dept. of Applied Mathematics, Waseda University, Tokyo, Japan.

Dr. Toyoizumi’s research interests include applied probability and stochastic models based on queueing theory.

Chia-Hung Wang is currently a Ph.D. candidate in the Department of Mathematical Sciences at National Chengchi University, Taipei, Taiwan. He received his B.S. degree in mathematics from National Tsing Hua University, Hsinchu, Taiwan in 2002. In 2004, he received his M.S. degree in mathematical sciences from National Chengchi University. From 2007 to 2008, he has a grant from the National Science Council, Taiwan to pursue advanced research at the Department of Information and Decision Sciences, University of Illinois at Chicago, USA. His papers have been published in *Computers & Operations Research*, *International Journal of Information and Management Sciences*, *Applied Mathematical Sciences*, *Lecture Notes in Computer Science*, *International Journal of Computer Science and Network Security*, *Lecture Notes in Operations Research*, *Leading-Edge Applied Mathematical Modeling Research*, and IEEE conference proceedings, among others.

Xiuli Xu received her M. Sc. degree in operations research and cybernetics in 2001 and Dr.Eng. degree in circuits and systems in 2006 both from Yanshan University. She currently is an associate professor of mathematics at the College of Science, Yanshan University, China, and she is a commissioner of queueing theory specialty, Operational Research Committee, China.

Dr. Xiuli Xu's research interest is the queueing system with vacations and its performance evaluation in communication networks. She has published research papers on these topics at important international and national journals as well as conference proceedings.

Dequan Yue received his B.Sci. degree in applied mathematics from Northeast Heavy Machinery Institute, Qiqihaer, China, the M.Sci. degree in applied mathematics from Xi'an Electronic University of Science and Technology, Xi'an, China, and Dr.Sci. degree in operations research and cybernetics from the Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing, China. Now Dr. Yue is a professor at Department of Statistics, Yanshan University, Qinhuangdao, China.

Dr. Yue's research interests include performance analysis of queueing systems, reliability analysis of repair systems, performance evaluation and modeling of communication networks, and stochastic order and its applications.

Dr. Yue has published over 50 papers in journals including *Naval Research Logistics*, *Operations Research Letters*, *Computer Communications*, *Microelectronics and Reliability*, *Nonlinear Dynamics and Systems Theory*, *International Journal of Pure and Applied Mathematics*, *Journal of Systems Science and Complexity*, and other journals.

Dr. Wuyi Yue (see "Editors".)

Index