Advances in Queueing Theory and Network Applications
Wuyi Yue • Yutaka Takahashi • Hideaki Takagi
Editors

Advances in Queueing Theory and Network Applications

Springer
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 .......................................................................................................................... ix

Part I Queueing Processes

1 Two Sided DQBD Process and Solutions to the Tail Decay Rate Problem and Their Applications to the Generalized Join Shortest Queue ................................................................. 3
Masakiyo Miyazawa

2 Analytical Model of On-Demand Streaming Services Based on Renewal Reward Theory ................................................................. 35
Hiroshi Toyoizumi

Part II Single-Server Queues

3 A Pure Decrement Service Geom/G/1 Queue with Multiple Adaptive Vacations ................................................................. 49
Zhanyou Ma, Wuyi Yue, and Naishuo Tian

4 Performance Analysis of an M/M/1 Working Vacation Queue with Setup Times ................................................................. 65
Xiuli Xu and Naishuo Tian

5 Modeling of Production System with Nonrenewal Batch Input, Early Setup, and Extra Jobs ................................................................. 77
Ho Woo Lee, No Ik Park, Se Won Lee, and Jung Woo Baek

6 Performance Analysis of an M/E_k/1 Queue with Balking and Two Service Rates Based on a Single Vacation Policy ................................................................. 103
Chunyan Li, Wuyi Yue, and Dequan Yue

Part III Multiple Queues

7 Markovian Polling Systems: Functional Computation for Mean Waiting Times and its Computational Complexity ................................................................. 119
Tetsuji Hirayama
8 Performance Analysis of a Two-Station MTO/MTS Production System .................................................... 147
Kuo-Hwa Chang and Yang-Shu Lu

Part IV Finite-Buffer Queues

9 Analysis of an M/M/c/N Queueing System with Balking, Reneging, and Synchronous Vacations ................................... 165
Dequan Yue and Wuyi Yue

10 Analysis of Mixed Loss-Delay M/M/m/K Queueing Systems with State-Dependent Arrival Rates ........................ 181
Yoshinori Ozaki and Hideaki Takagi

11 Asymptotic Behavior of Loss Rate for Feedback Finite Fluid Queue with Downward Jumps .................................... 195
Yutaka Sakuma and Masakiyo Miyazawa

12 Explicit Probability Density Function for the Length of a Busy Period in an M/M/1/K Queue ........................................... 213
Hideaki Takagi and Ahmed M.K. Tarabia

Part V Network Applications

13 Performance Analysis of ARQ Schemes in Self-Similar Traffic ........ 229
Shunfu Jin, Wuyi Yue, and Naishuo Tian

14 Modeling of P2P File Sharing with a Level-Dependent QBD Process . 247
Sophie Hautphenne, Kenji Leibnitz, and Marie-Ange Remiche

15 Performance Analysis of a Decentralized Content Delivery System with FEC Recovery .................................................. 265
Kenji Kirihara, Hiroyuki Masuyama, Shoji Kasahara, and Yutaka Takahashi

16 Blocking Probabilities of Multiple Classes in IP Networks with QoS Routing ......................................................... 281
Chia-Hung Wang and Hsing Luh

About the Editors and Authors ........................................ 303

Index ............................................................................. 313
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 balk- ing, 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 \( \text{Geom}^X/G/1 \) 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