

Alexander Dudin
Anatoly Nazarov
Alexander Moiseev (Eds.)

Communications in Computer and Information Science

912

Information Technologies and Mathematical Modelling

Queueing Theory and Applications

17th International Conference, ITMM 2018

Named After A.F. Terpugov

and 12th Workshop on Retrial Queues and Related Topics, WRQ 2018

Tomsk, Russia, September 10–15, 2018, Selected Papers

Communications in Computer and Information Science

912

Commenced Publication in 2007

Founding and Former Series Editors:

Phoebe Chen, Alfredo Cuzzocrea, Xiaoyong Du, Orhun Kara, Ting Liu,
Dominik Ślęzak, and Xiaokang Yang

Editorial Board

Simone Diniz Junqueira Barbosa

*Pontifical Catholic University of Rio de Janeiro (PUC-Rio),
Rio de Janeiro, Brazil*

Joaquim Filipe

Polytechnic Institute of Setúbal, Setúbal, Portugal

Igor Kotenko

*St. Petersburg Institute for Informatics and Automation of the Russian
Academy of Sciences, St. Petersburg, Russia*

Krishna M. Sivalingam

Indian Institute of Technology Madras, Chennai, India

Takashi Washio

Osaka University, Osaka, Japan

Junsong Yuan

University at Buffalo, The State University of New York, Buffalo, USA

Lizhu Zhou

Tsinghua University, Beijing, China

More information about this series at <http://www.springer.com/series/7899>

Alexander Dudin · Anatoly Nazarov
Alexander Moiseev (Eds.)

Information Technologies and Mathematical Modelling

Queueing Theory and Applications

17th International Conference, ITMM 2018

Named After A.F. Terpugov

and 12th Workshop on Retrial Queues and Related Topics, WRQ 2018

Tomsk, Russia, September 10–15, 2018

Selected Papers

Editors

Alexander Dudin
Department of Applied Mathematics
and Computer Science
Belarusian State University
Minsk
Belarus

Alexander Moiseev
Institute of Applied Mathematics
and Computer Science
Tomsk State University
Tomsk
Russia

Anatoly Nazarov
Institute of Applied Mathematics
and Computer Science
Tomsk State University
Tomsk
Russia

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-319-97594-8 ISBN 978-3-319-97595-5 (eBook)
<https://doi.org/10.1007/978-3-319-97595-5>

Library of Congress Control Number: 2018950432

© Springer Nature Switzerland AG 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The series of scientific conferences Information Technologies and Mathematical Modelling (ITMM) was started in 2002. In 2012, the series acquired an international status, and selected revised papers have been published in *Communication in Computer and Information Science* since 2014. The conference series was named by Alexander Terpugov, one of the first organizers of the conference, an outstanding scientist of the Tomsk State University and a leader of the famous Siberian school on applied probability, queueing theory, and applications.

Traditionally, the conferences have about ten sections in various fields of mathematical modelling and information technologies. Throughout the years, the sections on probabilistic methods and models, queueing theory, and communication networks have been the most popular at the conference. These sections gather many scientists from different countries. Many foreign participants come to this Siberia conference every year because of our warm welcome and serious scientific discussions.

This year, the ITMM conference was held in Tomsk together with 12th International Workshop on Retrial Queues and Related Topics (WRQ). This workshop is aimed at a specific area within queueing theory. It has been held since 1998 in different countries and traditionally gathers 20–40 scientists in field of retrial queues.

This volume presents selected papers from the 17th ITMM conference and the 12th WRQ. The papers are devoted to new results in queueing theory and its applications, including retrial queues. Its target audience includes specialists in probabilistic theory, random processes, and mathematical modelling as well as engineers engaged in logical and technical design and operational management of data processing systems, communication, and computer networks.

September 2018

Alexander Dudin
Anatoly Nazarov
Alexander Moiseev

Organization

The conference and workshop were organized by the National Research Tomsk State University, Peoples' Friendship University of Russia (RUDN University), Trapeznikov Institute of Control Sciences of Russian Academy of Sciences.

International Program Committee

A. Dudin (Chair)	Belarusian State University, Belarus
A. Nazarov (Co-chair)	Tomsk State University, Russia
K. Al-Begain	University of South Wales, UK
I. Atencia	University of Malaga, Spain
P. Cabral	Universidade Nova de Lisboa, Portugal
P. F. i Casas	Universitat Politècnica de Catalunya, Spain
S. Chakravarthy	Kettering University, USA
B. D. Choi	Korea University, South Korea
T. Czachórski	Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Poland
R. Dinis	Universidade Nova de Lisboa, Portugal
A. Economou	University of Athens, Greece
D. Efrosinin	Johannes Kepler University Linz, Austria
M. Farhadov	Institute of Control Sciences, Russian Academy of Sciences, Russia
Y. Gaydamaka	Peoples' Friendship University of Russia (RUDN University), Russia
E. Gelenbe	Imperial College, UK
A. Gómez-Corral	Complutense University of Madrid, Spain
A. Gortsev	Tomsk State University, Russia
V. Ivnickii	Railway Research Institute, Russia
B. Kim	Korea University, Korea
C. S. Kim	Sangji University, Korea
A. Kirpichnikov	Kazan National Research Technological University, Russia
U. Krieger	Universität Bamberg, Germany
B. Krishna Kumar	Anna University, Chennai, India
A. Krishnamurthy	Cochin University of Science and Technology, India
Q.-L. Li	Yan Shan University, China
Y. Malinkovsky	Francisk Skorina Gomel State University, Belarus
G. Medvedev	Belarusian State University, Belarus
A. Melikov	National Aviation Academy of Azerbaijan, Azerbaijan
A. Moiseev	Tomsk State University, Russia
S. Moiseeva	Tomsk State University, Russia
P. Montezuma-Carvalho	Universidade Nova de Lisboa, Portugal

R. Nobel	Vrije Universiteit Amsterdam, The Netherlands
M. Pagano	Pisa University, Italy
T. Phung-Duc	University of Tsukuba, Japan
T. B. Preußer	Technische Universität Dresden, Germany
J. Resing	Eindhoven University of Technology, The Netherlands
V. Rykov	Gubkin Russian State University of Oil and Gas, Russia
K. Samuylov	Peoples' Friendship University of Russia (RUDN University), Russia
S. Suschenko	Tomsk State University, Russia
D. Stamate	Goldsmiths College, UK
J. Sztrik	University of Debrecen, Hungary
H. Tijms	Vrije Universiteit Amsterdam, The Netherlands
O. Tikhonenko	Cardinal Stefan Wyszyński University in Warsaw, Poland
G. Tsitsiashvili	Institute of Applied Mathematics, Far Eastern Branch of Russian Academy of Sciences, Russia
V. Vishnevsky	Institute of Control Sciences, Russian Academy of Sciences, Russia

Local Organizing Committee

A. Moiseev (Chair)	M. Matalytski
S. Moiseeva (Co-chair)	S. Paul
A. Voitishak (Co-chair)	S. Rozhkova
A. Zamyatin (Co-chair)	D. Semenova
V. Broner	M. Shklennik
V. Bukreev	A. Shkurkin
E. Danilyuk	I. Shmyrin
E. Fedorova	E. Sopin
R. Gainullin	K. Voytikov
Y. Izmaylova	V. Zadorozhny
I. Lapatin	A. Zorin
E. Lisovskaya	

Contents

A Survey of Recent Results in Finite-Source Retrieal Queues with Collisions	1
<i>Anatoly Nazarov, János Sztrik, and Anna Kvach</i>	
Nonaffine Models of Yield Term Structure.	16
<i>Gennady Medvedev</i>	
On Gaussian Approximation of Queueing Networks with Different Starting Load	27
<i>Eugene Lebedev and Hanna Livinska</i>	
A Retrieal Queueing System with Orbital Search of Customers Lost from an Offer Zone	39
<i>Ambily P. Mathew, Achyutha Krishnamoorthy, and Varghese C. Joshua</i>	
Perishable Queueing Inventory Systems with Delayed Feedback.	55
<i>Agassi Melikov, Achyutha Krishnamoorthy, and Mammad Shahmaliyev</i>	
Methods of Limiting Decomposition and Markovian Summation in Queueing System with Infinite Number of Servers	71
<i>Anatoly Nazarov and Diana Dammer</i>	
Multi-channel Queueing Systems with Markovian Impatience	83
<i>Yury I. Ryzhikov</i>	
Optimal State Estimation of Semi-synchronous Event Flow of the Second Order Under Its Complete Observability	93
<i>Luydmila Nezhelskaya and Diana Tumashkina</i>	
Modelling of Output Flows in Queueing Systems and Networks.	106
<i>Gurami Tsitsiashvili and Marina Osipova</i>	
A Multi-server Queueing System with Backup Servers	117
<i>Valentina Klimenok, Alexander Dudin, and Uladzimir Shumchenia</i>	
Multiclass GI/GI/ ∞ Queueing Systems with Random Resource Requirements	129
<i>Ekaterina Lisovskaya, Svetlana Moiseeva, and Michele Pagano</i>	
Cost and Effect of Replication and Quorum in Desktop Grid Computing	143
<i>Alexander Rumyantsev, Srinivas Chakravarthy, Evsey Morozov, and Stanislav Remnev</i>	

Optimal Estimation of the States of Synchronous Generalized Flow of Events of the Second Order Under Its Complete Observability	157
<i>Luydmila Nezhelskaya and Ekaterina Sidorova</i>	
Asymptotic Sojourn Time Analysis of Finite-Source M/M/1 Retrial Queuing System with Two-Way Communication	172
<i>Anatoly Nazarov, János Sztrik, and Anna Kvach</i>	
An Analysis Method of Queueing Networks with a Degradable Structure and Non-zero Repair Times of Systems	184
<i>Igor E. Tananko and Nadezhda P. Fokina</i>	
An Infinite-Server Queueing $M MAP_k G_k \infty$ Model in Semi-Markov Random Environment Subject to Catastrophes	195
<i>K. Kerobyan, R. Covington, R. Kerobyan, and K. Enakoutsa</i>	
Retrial Queueing Model with Two-Way Communication, Unreliable Server and Resume of Interrupted Call for Cognitive Radio Networks	213
<i>Svetlana Paul and Tuan Phung-Duc</i>	
Mittag-Leffler Function in Applied Problems of Queueing Theory	225
<i>Alexander Kirpichnikov, Anton Titovtsev, and Igor Yakimov</i>	
A Contribution to Modeling Two-Way Communication with Retrial Queueing Systems.	236
<i>Attila Kuki, János Sztrik, Ádám Tóth, and Tamás Bérczes</i>	
Steady State Probabilistic Characteristics of the On/Off Production Rate Control Production-Inventory System with MMPP Demand Arrivals	248
<i>Klimentii Livshits, Anna Kitaeva, and Ekaterina Ulyanova</i>	
System State Distribution of a Finite-Source Retrial Queue with Subscribed Customers	263
<i>Velika Dragieva</i>	
Modeling of a Multi-link Transport Connection by a Network of Queueing Systems	274
<i>Pavel Mikhhev, Anastasiya Pichugina, and Sergey Suschenko</i>	
Estimation of Prioritized Disciplines Efficiency Based on the Metamodel of Multi-flows Queueing Systems	290
<i>V. N. Zadorozhnyi, T. R. Zakharenkova, and D. A. Tulubaev</i>	
Analysis of an Infinite-Server Queue $M AP_k G_k \infty$ in Random Environment with k Markov Arrival Streams and Random Volume of Customers	305
<i>K. Kerobyan, R. Kerobyan, and K. Enakoutsa</i>	

Optimization of Two-Level Discount Values Using Queueing Tandem Model with Feedback 321
Maria Shklennik, Svetlana Moiseeva, and Alexander Moiseev

Performance Analysis of an M/G/1 Retrial Queueing System Under LCFS-PR Discipline with General Retrial and Setup Times 333
B. Krishna Kumar, R. Sankar, and R. Rukmani

Method of Generating Functions for Performance Characteristic Analysis of the Polling Systems with Adaptive Polling and Gated Service. 348
Olga V. Semenova and Duy T. Bui

Retrial Queue with Search of Interrupted Customers from the Finite Orbit . . . 360
Dhanya Babu, Achyutha Krishnamoorthy, and Varghese C. Joshua

Traffic Optimization and Multi-sided Pricing in Congested Networks 372
Haroun H. Salih, Dina A. Urusova, and Sergey A. Vasilyev

Retrial Queueing System of MMPP/M/2 Type with Impatient Calls in the Orbit 387
Olga Vygovskaya, Elena Danilyuk, and Svetlana Moiseeva

Author Index 401



A Survey of Recent Results in Finite-Source Retrial Queues with Collisions

Anatoly Nazarov¹, János Sztrik²(✉), and Anna Kvach¹

¹ National Research Tomsk State University, 36 Lenina ave., Tomsk 634050, Russia
nazarov.tsu@gmail.com, kvach_as@mail.ru

² University of Debrecen, Debrecen, Hungary
sztrik.janos@inf.unideb.hu

Abstract. The aim of the present paper is to give a review of recent results on single server finite-source queuing systems with collision of the customers. There are investigations when the server is reliable and there are models when the server is subject to random breakdowns and repairs depending on whether it is idle or busy. Tool supported, numerical, simulation and asymptotic methods are considered under the condition of unlimited growing number of sources. Several cases and examples are treated and the results of different approaches are compared to each other showing the advantages and disadvantages of the given method. In general we could prove that the steady-state distribution of the number of customers in the service facility can be approximated by a normal distribution with given mean and variance. Using asymptotic methods under certain conditions in steady-state the distribution of the sojourn time in the orbit and in the system can be approximated by a generalized exponential one. Furthermore, it is proved that the distribution of the number of retrials until the successful service in the limit is geometrically distributed. By the help of stochastic simulation several systems are analyzed showing directions for further analytic investigations. Tables and Figures are collected to illustrate some special features of these systems.

Keywords: Finite-source queuing system · Retrial queues
Collisions · Server breakdowns and repairs · Analytic results
Algorithmic approach · Stochastic simulation · Asymptotic analysis

1 Introduction

Finite-source retrial queues are very useful and effective stochastic systems to model several problems arising in telephone switching systems, telecommunication networks, computer networks and computer systems, call centers, wireless communication systems, etc. To see their importance the interested reader is referred to the following works and references cited in them, for example [3, 9, 15, 19]. Searching the scientific databases we have noticed that relatively

just a small number of papers have been devoted to systems when the arriving calls (primary or secondary) causes collisions to the request under service and both go to the orbit, see for example [1, 7, 18, 24, 40].

Nazarov and his research group developed a very effective asymptotic method [39] by the help of which various systems have been investigated. Concerning to finite-source retrial systems with collision we should mention the following papers [25–28, 35].

Sztrik and his research group have been dealing with systems with unreliable server/s as can be seen, for example in [2, 44, 45, 51] and that is why it was understandable that the two research groups started cooperation in 2017.

Our investigations have been based on the analytical, numerical, simulation and asymptotic approached as treated in, for example [3, 5, 6, 10, 16, 20, 23, 29, 30, 34, 39, 42, 43, 50, 52].

The primary aim of the present paper is to give a survey on the results obtained in this field in the near past by means of different methods. Doing so we have tried to unify the notation appeared in different publications and to use the standard notation of Western-style papers which is many times differs from the Russian-style ones.

The rest of the paper is organized as follows. In Sect. 2 description of the model is given, the corresponding multi-dimensional non-Markov process is defined. In Sects. 3 and 4 systems with a reliable and an unreliable server are treated, respectively. In the subsections models with exponentially and generally distributed service times are investigated, and then analyzed by means of tool supported, algorithmic, simulation and asymptotic methods, respectively. The main results of the papers are collected and several Figures illustrate the most interesting features of the given system. Finally, the paper ends with a Conclusion and some future plans are highlighted.

2 Model Description and Notations

In the following we introduce the model in the most general form as it was treated by the help of numerical and asymptotic methods.

Let us consider a retrial queuing system of type $M/GI/1//N$ with collision of the customers and an unreliable server (Fig. 1). The number of sources is N and each of them can generate a primary request during an exponentially distributed time with rate λ/N . A source cannot generate a new call until the end of the successful service of this customer.

If a primary request finds the server idle, he enters into service immediately, in which the required service time has a probability distribution function $B(x)$. Let us denote its service rate function by $\mu(y) = B'(y)(1 - B(y))^{-1}$ and its Laplace-Stieltjes transform by $B^*(y)$, respectively. If the server is busy, an arriving (primary or repeated) customer involves into collision with customer under service and they both move into the orbit. The inter-retrial times of customers are supposed to be exponentially distributed with rate σ/N . We assume that the server is unreliable, that is its lifetime is supposed to be exponentially

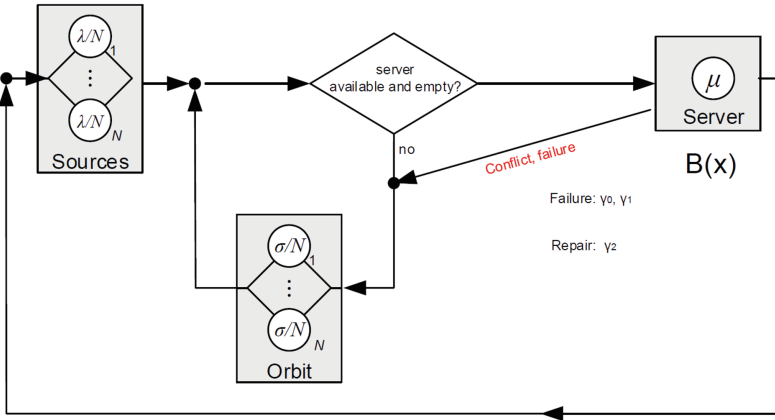


Fig. 1. Retrieal queueing system of type M/GI/1//N with collisions of the customers and an unreliable server

distributed with failure rate γ_0 if the server is idle and with rate γ_1 if it is busy. When the server breaks down, it is immediately sent for repair and the repair time is assumed to be exponentially distributed with rate γ_2 . We deal with the case when the server is down all sources continue generation of customers and send it to the orbit, similarly customers may retry from the orbit to the server but all arriving customers immediately go into the orbit. Furthermore, in this unreliable model we suppose that the interrupted request goes to the orbit immediately and its next service is independent of the interrupted one. Of course in the case of reliable server $\gamma_0 = \gamma_1 = 0$. All random variables involved in the model construction are assumed to be independent of each other.

Let $J(t)$ be the number of customers in the system at time t , that is, the total number of customers in the orbit and in service. Similarly, let $K(t)$ be the server state at time t , that is

$$K(t) = \begin{cases} 0, & \text{if the server is idle,} \\ 1, & \text{if the server is busy,} \\ 2, & \text{if the server is down (under repair).} \end{cases}$$

Thus, we will investigate the process $\{K(t), J(t)\}$, which is not a Markov-process unless the service time is exponentially distributed. To be a Markov one we will use the method of supplementary variables, namely, we will consider two variants: the residual service time method and the elapsed service time method depending on what is the aim of the investigation.

Let us denote by $Y(t)$, and $Z(t)$, the supplementary random process equal to the elapsed service time of the customer till the moment t and by $Z(t)$ the residual service time, that is time interval from the moment t until the end of successful service of the customer, respectively. It is obvious that $\{K(t), J(t), Y(t)\}$ and $\{K(t), J(t), Z(t)\}$ are Markov processes. Let us note, that $Y(t)$ and $Z(t)$ are defined only in those moments when the server is busy, that is, when $K(t) = 1$.

Let us define the stationary probabilities as follows:

$$\begin{aligned} P_0(j) &= P\{K = 0, J = j\}, \\ P_1(j, y) &= P\{K = 1, J = j, Y < y\}, \\ P_1(j, z) &= P\{K = 1, J = j, Z < z\}, \\ P_2(j) &= P\{K = 2, J = j\}. \end{aligned}$$

Of course in the case of exponentially distributed service time the steady-state probabilities are denoted as follows:

$$P_k(j) = P\{K = k, J = j\}, \quad k = 0, 1, 2, \quad j = 0, \dots, N.$$

The steady-state distribution of the server's state is denoted by

$$R_k = P(K = k), k = 0, 1, 2$$

and the distribution of number of customers in the system is designated by

$$P(j) = P(J = j), j = 0, \dots, N.$$

It is clear that in the case of reliable server all the probabilities where $K = 2$ are 0.

The main aim of the investigations is to get these distributions and other performance measures of the systems, such as the distribution of the sojourn time in the system, distribution of the total service time, distribution of the number of retrials. These are very complicated problems and to the best knowledge of the authors there are no exact analytical formulas to the solutions. That is the reason we have tried to obtain the characteristics of different systems by the help of tool supported, algorithmic, stochastic simulation and asymptotic methods.

3 Systems with a Reliable Server

3.1 M/M/1 Systems

Algorithmic Approach. In papers [26,35] the steady-state Kolmogorov equations were derived and the distribution of the system's state were obtained by an algorithmic approach. Then the distribution of the number of customers in the system were calculated and used to validate the asymptotic results.

Asymptotic Approach. The main contribution of paper [35] is that the in steady-state the prelimit distribution of the number of customers in the system can be approximated by a normal distribution with given mean and variance. In paper [35] 2nd and 3rd order approximations of the prelimit distribution were compared to the exact distribution obtained by the algorithmic method.

In different parameter setup and for different N the applicability of the asymptotic method was validated and some conclusions were drawn.

A more complicated problem, namely the distribution of the sojourn time in the service facility was investigated in [25] by the help of asymptotic methods as N tends to infinity. It was proved that the characteristic function of the sojourn time T of a customer spends in the service facility can be approximated by

$$E \exp \{iuT\} \approx q + (1 - q) \frac{\sigma q/N}{\sigma q/N - iu}, \quad q = \frac{\mu R_0}{\delta + \mu}.$$

3.2 M/GI/1 System

This section deals with the results when the required service times are generally distributed but in the examples the gamma distribution is used due to its useful properties. Namely, it is easy to see that its squared coefficient of variation can be less, equal or greater than 1 depending on the values of the shape and scale parameters.

Algorithmic Approach. Paper [27] deals with the algorithmic approach how to get the steady-state distribution of the system. The method of supplementary variable technique with residual service time were applied and several numerical examples were treated with gamma distributed service time. The results helped the validation of asymptotic results for the same model.

Stochastic Simulation. Papers [37,38] are devoted to the asymptotic analysis of the mean total service time, distribution of the sojourn time in the system and the distribution of number of retrials. It must be noted that the results have not been validated by simulation, yet. Meanwhile simulations have been carried out the estimations for the mean and variance of the sojourn time have been obtained, and the distribution of the number of retrials also has been determined. The simulation analysis will be published in the near future.

Asymptotic Approach. In this part the asymptotic results published in [37,38] are summarized. Before doing that we need some notations, namely

$$B^*(\alpha) = \int_0^\infty e^{-\alpha x} dB(x), \quad \delta(\kappa_1) = \lambda + (\sigma - \lambda)\kappa_1.$$

Then κ_1 can be obtained from

$$\kappa_1 = 1 - \frac{\delta(\kappa_1)}{\lambda} \cdot \frac{B^*(\delta(\kappa_1))}{2 - B^*(\delta(\kappa_1))}, \quad (1)$$

and the distribution of the server's state can be determined by

$$R_0 = \frac{1}{2 - B^*(\delta)}, \quad R_1 = \frac{1 - B^*(\delta)}{2 - B^*(\delta)}.$$

Introducing the notations

$$A_1 = \lambda(1 - \kappa_1), \quad R_1^*(\alpha) = -\delta R_0 [B^*(\alpha)],$$

we obtain

$$\kappa_2 = \frac{A_1 \left(R_0 \cdot B^*(\delta) [\delta + A_1] - (\delta + A_1 R_0) \right)}{A_1 (\sigma - \lambda) \left(R_1^*(\delta) - R_1 - R_0 (B^*(\delta) - 1) \right) + \delta \left((\sigma - \lambda) \left(R_1^*(\delta) - R_0 B^*(\delta) \right) - \lambda \right)}.$$

Consequently the steady-state prelimit distribution of the number of customers in the system can be approximated by a normal distribution with mean $N\kappa_1$ and variance $N\kappa_2$.

For the distribution of the number of retrials/transitions of the tagged customer into the orbit we have the following results.

Let ν be the number of transitions of the tagged customer into the orbit, then

$$\lim_{N \rightarrow \infty} \mathbb{E} z^\nu = \frac{q}{1 - (1 - q)z},$$

where value of parameter q has a form

$$q = R_0 B^*(\delta).$$

From the proved theorem it is obviously follows that the probability distribution $P\{\nu = n\}$, $n = \overline{0, \infty}$ of the number of transitions of the tagged customer into the orbit is geometric and

$$P\{\nu = n\} = q(1 - q)^n, \quad n = \overline{0, \infty}.$$

Consequently, by using the law of total probability for the characteristic function of the sojourn/waiting time W of the tagged customer in the orbit we get

$$\mathbb{E} e^{iuW} \approx q + (1 - q) \frac{\sigma q}{\sigma q - iuN}.$$

In the case of $N \rightarrow \infty$ the limiting probability distributions of the sojourn time of the customer in the system T and the sojourn time of the customer in the orbit W coincide, namely

$$\lim_{N \rightarrow \infty} \mathbb{E} \exp \left\{ iu \frac{T}{N} \right\} = \lim_{N \rightarrow \infty} \mathbb{E} \exp \left\{ iu \frac{W}{N} \right\} = q + (1 - q) \frac{\sigma q}{\sigma q - iu}.$$

4 Systems with an Unreliable Server

In many practical situations the server is not reliable and after a random time it can fail and needs repair which also takes a random duration. To deal with these service interruptions several papers have been published, see for example [2, 8, 11, 12, 14, 21, 41, 45, 48, 49, 53]. In the following parts we summarize our results obtained by different methods.

4.1 M/M/1 System

Tool Supported Approach by MOSEL. Because of the fact, that in many practical situations the state space of the describing Markov chain is very large, it is rather difficult to calculate the system measures in the traditional way of writing down and solving the underlying steady-state equations. To simplify this procedure several software packages have been developed and effectively used for performance evaluation of complex systems, see for example [11–14, 17]. In our investigations a similar software tool called MOSEL (Modeling, Specification and Evaluation Language) has been used to formulate the model and to obtain the performance measures. Paper [4] deals with the model formulation, derivation of several performance measures and generation of illustrative examples showing an interesting phenomenon of finite-source retrial queues, that is under specific parameter setup the mean waiting/ sojourn time has a maximum as the arrival intensity is increasing.

Stochastic Simulation. To validate the applicability of the asymptotic approach we need either numerical or simulation results. The correct operation of the simulation software was tested by the numerical sample examples. The investigations carried out by the simulation and asymptotic methods have been submitted for publication, see [31, 32].

Asymptotic Approach. First we deal with the distribution of the number customers in the system as it has been published in [31]. The first order asymptotic results are the following

$$\lim_{N \rightarrow \infty} E \exp \left\{ iw \frac{J}{N} \right\} = \exp \{ iw \kappa_1 \},$$

where κ_1 is the positive solution of the equation

$$(1 - \kappa_1) \lambda - \mu R_1(\kappa_1) = 0,$$

where the stationary distributions of probabilities $R_k(\kappa_1)$ of the server state $k = 0, 1, 2$ are obtained as follows

$$R_0(\kappa_1) = \left\{ \frac{\gamma_0 + \gamma_2}{\gamma_2} + \frac{\gamma_1 + \gamma_2}{\gamma_2} \cdot \frac{a(\kappa_1)}{a(\kappa_1) + \gamma_1 + \mu} \right\}^{-1},$$

$$R_1(\kappa_1) = \frac{a(\kappa_1)}{a(\kappa_1) + \gamma_1 + \mu} \cdot R_0(\kappa_1),$$

$$R_2(\kappa_1) = \frac{1}{\gamma_2} [\gamma_0 R_0(\kappa_1) + \gamma_1 R_1(\kappa_1)],$$

here $a(\kappa_1)$ is

$$a(\kappa_1) = (1 - \kappa_1) \lambda + \sigma \kappa_1.$$

The second order asymptotic results are

$$\lim_{N \rightarrow \infty} E \exp \left\{ iw \frac{J - \kappa_1 N}{\sqrt{N}} \right\} = \exp \left\{ \frac{(iw)^2}{2} \kappa_2 \right\},$$

where κ_2 is

$$\kappa_2 = \frac{\gamma_2 \mu (R_1 - b_1) + (1 - \kappa_1) \lambda \{ (\gamma_1 + \gamma_2) b_1 + (1 - \kappa_1) \lambda R_2 \}}{(\lambda + \mu b_2) \gamma_2 - (1 - \kappa_1) \lambda (\gamma_1 + \gamma_2) b_2},$$

and

$$b_1 = \frac{(1 - \kappa_1) \lambda}{a + \gamma_1 + \mu} R_0, \quad b_2 = \frac{(\sigma - \lambda)(R_0 - R_1)}{a + \gamma_1 + \mu}.$$

Consequently the prelimit distribution of the number of customers in the system can be approximated by a normal distribution with mean $N\kappa_1$ and variance $N\kappa_2$.

One of the main contributions of paper [32] is that for the limit of the characteristic function of the normalized sojourn time we have

$$\lim_{N \rightarrow \infty} E \exp \left\{ iw \frac{T}{N} \right\} = q + (1 - q) \frac{\sigma q}{\sigma q - iw},$$

where q is

$$q = \frac{(1 - \kappa_1) \lambda}{(1 - \kappa_1) \lambda + \sigma \kappa_1}.$$

Consequently the characteristic function of the sojourn time of the customer in the system in the prelimit situation of finite N can be approximated by

$$E e^{iuT} \approx q + (1 - q) \frac{\sigma q}{\sigma q - iuN}. \quad (2)$$

For the distribution of the number of transitions/retrials of the tagged customer into the orbit we got the following results.

Let ν be the number of transitions of the tagged customer into the orbit, then

$$\lim_{N \rightarrow \infty} E z^\nu = \frac{q}{1 - (1 - q)z},$$

resulting that the probability distribution $P \{ \nu = n \}$, $n = \overline{0, \infty}$ of the number of transitions of the tagged customer into the orbit is geometric and has the form

$$P \{ \nu = n \} = q(1 - q)^n, \quad n = \overline{0, \infty}.$$

Consequently the prelimit characteristic function of the sojourn/waiting time W of the tagged customer in an orbit can be approximated as

$$E e^{iuW} \approx q + (1 - q) \frac{\sigma q}{\sigma q - iuN}.$$

In the case of $N \rightarrow \infty$ the limiting probability distributions of the sojourn time of the customer in the system T and the sojourn time of the customer in an orbit W coincide, namely

$$\lim_{N \rightarrow \infty} \mathbb{E} \exp \left\{ iu \frac{T}{N} \right\} = \lim_{N \rightarrow \infty} \mathbb{E} \exp \left\{ iu \frac{W}{N} \right\} = q + (1 - q) \frac{\sigma q}{\sigma q - iu}.$$

4.2 M/GI/1 System

Stochastic Simulation. In paper [47] the required service time is supposed to be gamma distributed and the input parameters of the system are collected in Table 1.

Table 1. Numerical values of model parameters

Case	N	λ/N	γ_0	γ_1	γ_2	σ/N	α	β
1	100	0.01	0.1	0.1	1	0.01	0.5	0.5
2	100	0.01	0.1	0.1	1	0.01	1	1
3	100	0.01	0.1	0.1	1	0.01	2	2

Figure 2 shows the steady-state distribution of the three investigated cases. It is observed the mean number of customers increases as α and β are getting larger. *Case 2* is a special case because when $\alpha = 1$ it represents the exponential distribution. From the shape of the curves it is clearly visible that the

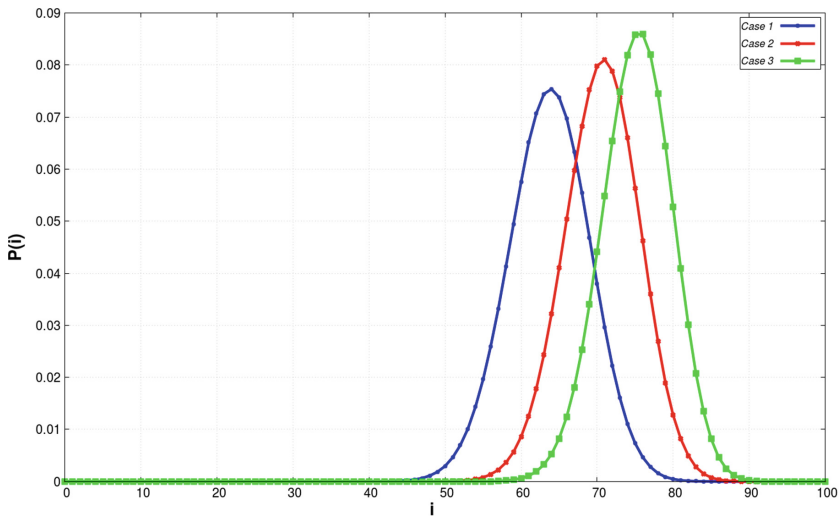


Fig. 2. Comparison of steady-state distributions

steady-state distribution of the cases are normally distributed. The next table presents the considered performance measures in relation with the different cases (see Table 2).

In Table 2 the notations mean the followings: $E(J)$ and $Var(J)$ - mean number and variance of customers in the system, $E(T)$ and $Var(T)$ - mean and variance of response time, $E(W)$ and $Var(W)$ - mean and variance of waiting time, $E(S)$ and $Var(S)$ - mean and variance of successful service time, $E(IS)$ - mean interrupted service time.

Table 2. Simulation results

Case	$E(J)$	$Var(J)$	$E(T)$	$Var(T)$	$E(W)$	$Var(W)$	$E(S)$	$Var(S)$	$E(IS)$
1	63.6842	27.9734	175.3073	65657.3454	174.5884	65434.6696	0.3147	0.1979	0.4041
2	70.5912	24.3012	239.9734	105273.4267	238.9734	104918.6389	0.4784	0.2289	0.5217
3	75.1825	21.2439	302.8106	151781.1411	301.5377	151277.6006	0.6472	0.2095	0.6257

Figure 3 represents the confirmation of mean waiting time. The same parameters are (see Table 2) used as in case of Fig. 2 but here the running parameter is λ/N . As it is expected with the increment of λ/N mean waiting time increases as well but an interesting phenomenon is noticeable namely after λ/N is greater than 0.1 mean waiting time starts to decrease.

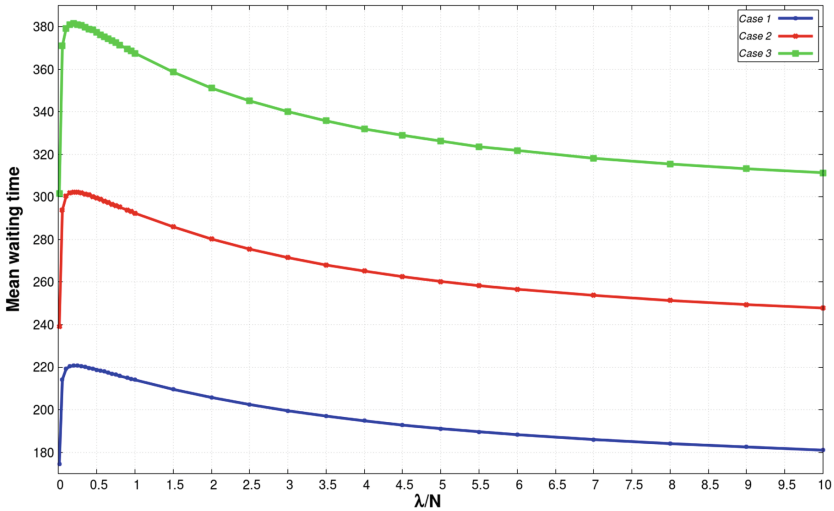


Fig. 3. Mean waiting time vs. intensity of incoming customers

Asymptotic Approach. These results have been published in [36] using supplementary variable technique. The limit of the characteristic function of the

scaled number of customers in the systems can be written in the following form

$$\lim_{N \rightarrow \infty} \mathbf{E} \exp \left\{ iw \frac{J}{N} \right\} = \exp \{ iw \kappa_1 \},$$

where κ_1 is the positive solution of the equation

$$(1 - \kappa_1) \lambda - \delta(\kappa_1) [R_0(\kappa_1) - R_1(\kappa_1)] + \gamma_1 R_1(\kappa_1) = 0,$$

here $\delta(\kappa_1)$ is

$$\delta(\kappa_1) = (1 - \kappa_1) \lambda + \sigma \kappa_1,$$

and the stationary distributions of probabilities $R_k(\kappa_1)$ of the server's state $k = 0, 1, 2$ are determined as follows

$$R_0(\kappa_1) = \left\{ \frac{\gamma_0 + \gamma_2}{\gamma_2} + \frac{\gamma_1 + \gamma_2}{\gamma_2} \cdot \frac{\delta(\kappa_1)}{\delta(\kappa_1) + \gamma_1} [1 - B^*(\delta(\kappa_1) + \gamma_1)] \right\}^{-1},$$

$$R_1(\kappa_1) = R_0(\kappa_1) \frac{\delta(\kappa_1)}{\delta(\kappa_1) + \gamma_1} \cdot [1 - B^*(\delta(\kappa_1) + \gamma_1)],$$

$$R_2(\kappa_1) = \frac{1}{\gamma_2} [\gamma_0 R_0(\kappa_1) + \gamma_1 R_1(\kappa_1)].$$

4.3 Stochastic Simulation of Special Systems

In paper [47] systems with not only gamma distributed service times but also gamma distributed inter-arrival and gamma distributed retrial times have been investigated.

The Effect of Breakdowns Disciplines. In paper [46] the $M/G/1//N$ and $G/M/1//N$ systems were investigated with exponentially distributed operating and repair times. In case of a server failure two operation modes are considered:

- The interrupted request gets into the orbit instantaneously.
- The service of the interrupted request is suspended and it continues after repairing the server.

As it was expected the second operation mode results lower mean sojourn times and higher mean successful service times. The Figures are similar to the cases treated earlier that is why they are omitted.

5 Conclusion

In this paper tool supported, numerical, simulation and asymptotic methods were considered under the condition of unlimited growing number of sources in a finite-source retrial queue with collisions of customers and an unreliable server.

During the survey several cases and examples were treated and the results of different approaches were compared to each other showing the advantages and disadvantages of the given method. Tables and Figures were collected to illustrate some special features of these systems. In the near future the two research groups would like to continue their investigations in this direction including systems with impatient customers, systems embedded in a random environment, systems with two-way communications, just to mention some alternative generalizations.

Acknowledgments. The work/publication of J. Sztrik is supported by the EFOP-3.6.1-16-2016-00022 project. The project is co-financed by the European Union and the European Social Fund.

References

1. Ali, A.A., Wei, S.: Modeling of coupled collision and congestion in finite source wireless access systems. In: *Wireless Communications and Networking Conference (WCNC)*, pp. 1113–1118. IEEE (2015)
2. Almási, B., Roszik, J., Sztrik, J.: Homogeneous finite-source retrial queues with server subject to breakdowns and repairs. *Math. Comput. Model.* **42**(5–6), 673–682 (2005)
3. Artalejo, J., Corral, A.G.: *Retrial Queueing Systems: A Computational Approach*. Springer, Heidelberg (2008)
4. Bérczes, T., Sztrik, J., Tóth, Á., Nazarov, A.: Performance modeling of finite-source retrial queueing systems with collisions and non-reliable server using MOSEL. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) *DCCN 2017. CCIS*, vol. 700, pp. 248–258. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66836-9_21
5. Bhat, U.N.: *An introduction to queueing theory. Modeling and analysis in applications*. 2nd edn. Birkhäuser, Boston (2015)
6. Bossel, H.: *Modeling and simulation*. Springer, Heidelberg (2013)
7. Choi, B.D., Shin, Y.W., Ahn, W.C.: Retrial queues with collision arising from unslotted CSMA/CD protocol. *Queueing Syst.* **11**(4), 335–356 (1992)
8. Dragieva, V.I.: Number of retrials in a finite source retrial queue with unreliable server. *Asia-Pac. J. Oper. Res.* **31**(2), 23 (2014)
9. Falin, G., Artalejo, J.: A finite source retrial queue. *Eur. J. Oper. Res.* **108**, 409–424 (1998)
10. Falin, G., Templeton, J.G.C.: *Retrial Queues*. Chapman and Hall, London (1997)
11. Gharbi, N., Dutheillet, C.: An algorithmic approach for analysis of finite-source retrial systems with unreliable servers. *Comput. Math. Appl.* **62**(6), 2535–2546 (2011)
12. Gharbi, N., Ioualalen, M.: GSPN analysis of retrial systems with servers breakdowns and repairs. *Appl. Math. Comput.* **174**(2), 1151–1168 (2006)
13. Gharbi, N., Mokdad, L., Ben-Othman, J.: A performance study of next generation cellular networks with base stations channels vacations. In: *Global Communications Conference (GLOBECOM)*, pp. 1–6. IEEE (2015)
14. Gharbi, N., Nemmouchi, B., Mokdad, L., Ben-Othman, J.: The impact of breakdowns disciplines and repeated attempts on performances of small cell networks. *J. Comput. Sci.* **5**(4), 633–644 (2014)

15. Gómez-Corral, A., Phung-Duc, T.: Retrial queues and related models. *Ann. Oper. Res.* **247**(1), 1–2 (2016)
16. Harchol-Balter, M.: Performance modeling and design of computer systems. *Queueing Theory in Action*. Cambridge University Press, New York (2013)
17. Ikhlef, L., Lekadir, O., Aïssani, D.: MRSPN analysis of Semi-Markovian finite source retrial queues. *Ann. Oper. Res.* **247**(1), 141–167 (2016)
18. Kim, J.S.: Retrial queueing system with collision and impatience. *Commun. Korean Math. Soc.* **25**(4), 647–653 (2010)
19. Kim, J., Kim, B.: A survey of retrial queueing systems. *Ann. Oper. Res.* **247**(1), 3–36 (2016)
20. Kobayashi, H., Mark, B.L.: System modeling and analysis: Foundations of system performance evaluation. Pearson Education, India (2009)
21. Krishnamoorthy, A., Pramod, P.K., Chakravarthy, S.R.: Queues with interruptions: a survey. *TOP* **22**(1), 290–320 (2014)
22. Kuki, A., T.Bérczes, Sztrik, J., Kvach, A.: Numerical analysis of retrial queueing systems with conflict of customers. *J. Math. Sci.* (2017). (submitted)
23. Kulkarni, V.G.: Modeling and analysis of stochastic systems. CRC Press, Boca Raton (2016)
24. Kumar, B.K., Vijayalakshmi, G., Krishnamoorthy, A., Basha, S.S.: A single server feedback retrial queue with collisions. *Comput. Oper. Res.* **37**(7), 1247–1255 (2010)
25. Kvach, A., Nazarov, A.: Sojourn Time analysis of finite Source Markov retrial queueing system with collision. In: Dudin, A., Nazarov, A., Yakupov, R. (eds.) *ITMM 2015*. CCIS, vol. 564, pp. 64–72. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25861-4_6
26. Kvach, A.: Numerical research of a Markov closed retrial queueing system without collisions and with the collision of the customers. In: *Proceedings of Tomsk State University. A series of physics and mathematics*. Tomsk. Materials of the II All-Russian Scientific Conference, vol. 295, pp. 105–112. TSU Publishing House (2014). (in Russian)
27. Kvach, A., Nazarov, A.: Numerical research of a closed retrial queueing system $M/GI/1//N$ with collision of the customers. In: *Proceedings of Tomsk State University. A series of physics and mathematics*. Tomsk. Materials of the III All-Russian Scientific Conference, vol. 297, pp. 65–70. TSU Publishing House (2015). (in Russian)
28. Kvach, A., Nazarov, A.: The research of a closed RQ-system $M/GI/1//N$ with collision of the customers in the condition of an unlimited increasing number of sources. In: *Probability Theory, Random Processes, Mathematical Statistics and Applications: Materials of the International Scientific Conference Devoted to the 80th Anniversary of Professor Gennady Medvedev, Doctor of Physical and Mathematical Sciences*, pp. 65–70 (2015). (in Russian)
29. Lakatos, L., Szeidl, L., Telek, M.: Introduction to queueing systems with telecommunication applications. Springer, New York (2013)
30. Law, A.M., Kelton, W.D.: Simulation modeling and analysis. McGraw-Hill, New York (1991)
31. Nazarov, A., Sztrik, J., Kvach, A., Bérczes, T.: Asymptotic analysis of finite-source $M/M/1$ retrial queueing system with collisions and server subject to breakdowns and repairs. *Ann. Oper. Res.* (2017). (submitted)
32. Nazarov, A., Sztrik, J., Kvach, A., Tóth, A.: Asymptotic sojourn time analysis of Markov finite-source $M/M/1$ retrial queueing system with collisions and server subject to breakdowns and repairs. *Markov Processes and Related Fields* (2017). (submitted)

33. Nazarov, A., Sudyko, E.: Method of asymptotic semi-invariants for studying a mathematical model of a random access network. *Probl. Inf. Transm.* **46**(1), 86–102 (2010)
34. Nazarov, A., Terpugov, A.: *Theory of Mass Service*. NTL Publishing House, Tomsk (2004). (in Russian)
35. Nazarov, A., Kvach, A., Yampolsky, V.: Asymptotic analysis of closed markov retrial queuing system with collision. In: Dudin, A., Nazarov, A., Yakupov, R., Gortsev, A. (eds.) *ITMM 2014. CCIS*, vol. 487, pp. 334–341. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13671-4_38
36. Nazarov, A., Sztrik, J., Kvach, A.: Comparative analysis of methods of residual and elapsed service time in the study of the closed retrial queuing system $M/GI/1//N$ with collision of the customers and unreliable server. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) *ITMM 2017. CCIS*, vol. 800, pp. 97–110. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_8
37. Nazarov, A., Sztrik, J., Kvach, A.: Some features of a finite-source $M/GI/1$ retrial queuing system with collisions of customers. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) *DCCN 2017. CCIS*, vol. 700, pp. 186–200. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66836-9_16
38. Nazarov, A., Sztrik, J., Kvach, A.: Some features of a finite-source $M/GI/1$ retrial queuing system with collisions of customers. In: *Proceedings of International Conference on Distributed Computer and Communication Networks, DCCN 2017*, pp. 79–86 (2017)
39. Nazarov, A., Moiseeva S.P.: *Methods of asymptotic analysis in queueing theory*. NTL Publishing House of Tomsk University (2006). (in Russian)
40. Peng, Y., Liu, Z., Wu, J.: An $M/G/1$ retrial G-queue with preemptive resume priority and collisions subject to the server breakdowns and delayed repairs. *J. Appl. Math. Comput.* **44**(1–2), 187–213 (2014)
41. Roszik, J.: Homogeneous finite-source retrial queues with server and sources subject to breakdowns and repairs. *Ann. Univ. Sci. Budap. Rolando Eötvös, Sect. Comput.* **23**, 213–227 (2004)
42. Rubinstein, R.Y., Kroese, D.P.: *Simulation and the Monte Carlo method*. Wiley, Hoboken (2016)
43. Stewart, W.J.: *Probability, Markov chains, queues, and simulation. the mathematical basis of performance modeling*. Princeton University Press, Princeton (2009)
44. Sztrik, J.: Tool supported performance modelling of finite-source retrial queues with breakdowns. *Publicationes Mathematicae* **66**, 197–211 (2005)
45. Sztrik, J., Almási, B., Roszik, J.: Heterogeneous finite-source retrial queues with server subject to breakdowns and repairs. *J. Math. Sci.* **132**, 677–685 (2006)
46. Tóth, A., Bérczes, T., Sztrik, J., Kuki, A.: Comparison of two operation modes of finite-source retrial queueing systems with collisions and non-reliable server by using simulation. *J. Math. Sci.* (2017). (submitted)
47. Tóth, Á., Bérczes, T., Sztrik, J., Kvach, A.: Simulation of finite-source retrial queueing systems with collisions and non-reliable server. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) *DCCN 2017. CCIS*, vol. 700, pp. 146–158. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66836-9_13
48. Wang, J., Zhao, L., Zhang, F.: Performance analysis of the finite source retrial queue with server breakdowns and repairs. In: *Proceedings of the 5th International Conference on Queueing Theory and Network Applications*, pp. 169–176. ACM (2010)
49. Wang, J., Zhao, L., Zhang, F.: Analysis of the finite source retrial queues with server breakdowns and repairs. *J. Ind. Manag. Optim.* **7**(3), 655–676 (2011)

50. Wehrle, K., Günes, M., Gross, J.: Modeling and tools for network simulation. Springer, Heidelberg (2010)
51. Wüchner, P., Sztrik, J., de Meer, H.: Finite-source retrial queues with applications. In: Proceedings of 8th International Conference on Applied Informatics, Eger, Hungary. vol. 2, pp. 275–285 (2010)
52. Yao, J.: Asymptotic Analysis of Service Systems with Congestion-Sensitive Customers. Columbia University (2016)
53. Zhang, F., Wang, J.: Performance analysis of the retrial queues with finite number of sources and service interruptions. *J. Korean Stat. Soc.* **42**(1), 117–131 (2013)



Nonaffine Models of Yield Term Structure

Gennady Medvedev^(✉)

Department of Applied Mathematics and Computer Science,
Belarusian State University, Nezavisimosti ave. 4, 220030 Minsk, Belarus
MedvedevGA@bsu.by
<http://www.bsu.by>

Abstract. The equation of term structure for the price of a zero-coupon bond is considered, the solution of which in analytical form is known, basically, for the simplest models and has an affine structure with respect to the short-term rate. The paper constructs solutions of this equation for a family of term structure models that are based on short-term rate processes in which the square of volatility is proportional to the third power of the short-term rate in stochastic differential equations. The solution of the equation is sought in the form of a definite functional series and, as a result, is reduced to a confluent hypergeometric function. Three versions of the underlying stochastic differential equations for short-term rate processes are considered: with zero drift, linear drift, and quadratic drift. Numerical examples are given for the yield curve and the forward rate curve for these versions. Some conditions for the existence of nontrivial solutions of the equation of time structure in the family of processes under consideration are formulated.

Keywords: The equation of the yield term structure
The price of zero-coupon bond · The CIR(1980) model
The Ahn – Gao model · The yield curve · The forward curve

1 Introduction

Suppose that the state of the financial market is described by the interest rate $r(t)$, which follows a Markov process homogeneous in time, generated by the stochastic differential equation

$$dr(t) = \mu(r(t))dt + \sigma(r(t))dw(t)$$

with the drift function $\mu(x)$, the volatility function $\sigma(x)$, and the standard Wiener process $w(t)$. For convenience of reasoning, we denote the drift function $m(r) = \mu(r) - \lambda(r)\sigma(r)$ and the diffusion function $s(r) = 0.5\sigma^2(r)$. Here $\lambda(r)$ is the so-called market risk price. Previously [1], the problem of determining the time structure of the yield of a zero-coupon bond, when the functions $m(r)$ and $s(r)$ are polynomials was considered. It turned out, in this case, whether the yield curves can be polynomials or power series in the variable r . It turned

out that this happens if only $m(r)$ and $s(r)$ are polynomials of not more than first degree. In this case, the models of the yield term structure are affine.

In this paper, we consider a similar problem, but the term structure of the price of a zero-coupon bond is sought in the form of a functional series that differs from the power series. It is found out that for some cases such solutions exist. The resulting term structure turns out to be non-affine and is described by confluent hypergeometric functions. This family includes such known models of interest rates as the model CIR(1980) [2] and the model Ahn – Gao [3].

2 The General Equation for the Price of a Bond and Its Components

Consider the equation of term structure for the price of the zero-coupon bond $P(r, \tau)$ [4]

$$-\frac{\partial P(r, \tau)}{\partial \tau} + m(r) \frac{\partial P(r, \tau)}{\partial r} + s(r) \frac{\partial^2 P(r, \tau)}{\partial r^2} - rP(r, \tau) = 0, \quad P(r, 0) = 1. \quad (1)$$

Here $m(r)$ is the function of the short-term interest rate drift, and $s(r)$ is the square of its volatility. We seek a solution of this equation in the form

$$P(r, \tau) = \sum_{n=0}^{\infty} \left(\frac{a(\tau)}{r} \right)^{\alpha+n} c_n, \quad (2)$$

where $a(\tau)$, α and c_n , $n = 0, 1, 2, \dots$, are the function and coefficients to be determined.

The corresponding derivatives used in Eq. (1) have the form

$$\begin{aligned} \frac{\partial P(r, \tau)}{\partial \tau} &= \frac{a'(\tau)}{a(\tau)} \sum_{n=0}^{\infty} (\alpha + n) \left(\frac{a(\tau)}{r} \right)^{\alpha+n} c_n, \\ \frac{\partial P(r, \tau)}{\partial r} &= \frac{1}{a(\tau)} \sum_{n=0}^{\infty} (\alpha + n) \left(\frac{a(\tau)}{r} \right)^{\alpha+n+1} c_n, \\ \frac{\partial^2 P(r, \tau)}{\partial r^2} &= \frac{1}{a(\tau)^2} \sum_{n=0}^{\infty} (\alpha + n)(\alpha + n + 1) \left(\frac{a(\tau)}{r} \right)^{\alpha+n+2} c_n. \end{aligned} \quad (3)$$

Suppose that the drift and volatility of the short-term interest rate are such that the functions $m(r)$ and $s(r)$ are polynomials of order p and q , respectively:

$$m(r) = \sum_{k=0}^p m_k r^k, \quad s(r) = \sum_{k=0}^q s_k r^k. \quad (4)$$

Now substituting expressions (2) – (4) in Eq. (1), we obtain

$$\begin{aligned}
& \sum_j (-I(j|0) (\alpha + j) a'(\tau) a(\tau)^{\alpha+j-1} c_j - I(j|-1) a(\tau)^{\alpha+j+1} c_{j+1} \\
& - I(j|1-p) \sum_{k=\text{Max}\{0, 1-j\}}^p (\alpha + j + k - 1) m_k a(\tau)^{\alpha+j+k-1} c_{j+k-1} \\
& + I(j|2-q) \sum_{k=\text{Max}\{0, 2-j\}}^q [(\alpha + j + k - 2t) (\alpha + j + k - 1) \\
& \times s_k a(\tau)^{\alpha+j+k-2} c_{j+k-2}]) \left(\frac{1}{r}\right)^{\alpha+j} = 0. \tag{5}
\end{aligned}$$

A certain complexity in the expression (5) is caused by the fact that the summation over the index j for each term starts in different ways: for the first term $j \geq 0$, for the second term $j \geq -1$, for the third term $j \geq 1 - p$, for the fourth summand $j \geq 2 - q$. Therefore, in expressions of the terms, the factors $I(j|k)$ appeared, representing indicator functions equal to one if $j \geq k$, and zero otherwise.

Equality (5) must be satisfied uniformly with respect to the variable r . In this case, since the functions r^{-j} ($j = 0, \pm 1, \pm 2, \dots$) are linearly independent, the coefficients in front of these functions in expression (5) must be zero. This leads to a system of equations for the unknown parameters α , $a(\tau)$ and c_n , $n = 0, 1, 2, \dots$, in the representation (2) of the solution of Eq. (1), if it exists in this form. Note that each term in each element of the sum (5) has a nonzero factor $a(\tau)^\alpha$, therefore, for simplicity, it can be reduced in all elements of the sum.

3 The CIR(1980) Model

Among the models of short-term rate $r(t)$ processes with zero drift, the CIR(1980) model [2] is widely known, in which the rate is generated in the general case by the diffusion process

$$dr = \sigma r^\gamma dw. \tag{6}$$

Despite the fact that the model is known for a long time, the time structure of its zero-coupon yield has not been described so far. It turns out that the proposed method for finding the time structure allows this. In Eq. (6) we take $\gamma = 1.5$ and $s \equiv 0.5\sigma^2$. Equation (1) for the price of the zero-coupon bond $P(r, \tau)$ takes the form

$$-\frac{\partial P(r, \tau)}{\partial \tau} + sr^3 \frac{\partial^2 P(r, \tau)}{\partial r^2} - rP(r, \tau) = 0, \quad P(r, 0) = 1. \tag{7}$$

We seek a solution of this equation in the form (2). The corresponding derivatives have the form (3). After substituting these expressions into Eq. (7), we obtain the following equality

$$\begin{aligned}
 & -\frac{a'(\tau)}{a(\tau)} \sum_{n=0}^{\infty} (\alpha + n) \left(\frac{a(\tau)}{r}\right)^{\alpha+n} c_n - a(\tau) \sum_{n=0}^{\infty} \left(\frac{a(\tau)}{r}\right)^{\alpha+n-1} c_n \\
 & + s a(\tau) \sum_{n=0}^{\infty} (\alpha + n)(\alpha + n + 1) \left(\frac{a(\tau)}{r}\right)^{\alpha+n-1} c_n = 0.
 \end{aligned}$$

This equality can be rewritten in a more convenient form:

$$\begin{aligned}
 & \sum_{n=0}^{\infty} \left[(\alpha + n) \frac{a'(\tau)}{a(\tau)} c_n + a(\tau) c_{n+1} - s a(\tau) (\alpha + n + 1)(\alpha + n + 2) c_{n+1} \right] \\
 & \times \left(\frac{a(\tau)}{r}\right)^{\alpha+n} + a(\tau) (1 - s \alpha (\alpha + 1)) c_0 \left(\frac{a(\tau)}{r}\right)^{\alpha-1} = 0.
 \end{aligned}$$

Since the expressions $(a(\tau)/r)^k$ as functions of the variable r for different values of k are linearly independent, and the equality must be satisfied uniformly with respect to r , then the coefficients before these expressions for different k must be zero. And we get a system of equations for the unknowns α , $a(\tau)$ and c_n , $n = 0, 1, 2, \dots$

$$s\alpha(\alpha + 1) = 1, \quad (8)$$

$$(\alpha + n) \frac{a'(\tau)}{a(\tau)^2} c_n + c_{n+1} - s(\alpha + n + 1)(\alpha + n + 2) c_{n+1} = 0, \quad n = 0, 1, 2, \dots \quad (9)$$

From the Eq. (8) the parameter α is determined by

$$\alpha = \frac{1}{2} \left(\sqrt{1 + \frac{4}{s}} - 1 \right) \equiv \frac{1}{2} \left(\frac{\sqrt{8 + \sigma^2}}{\sigma} - 1 \right) > 0. \quad (10)$$

Generally speaking, Eq. (8) has two roots: positive and negative. However, with a negative solution, as will be shown below, the price function $P(r, \tau)$ acquires properties that the price of the zero-coupon bond does not possess. Therefore, we take the root (10). Consider the Eq. (9) for $n = 0$.

$$a'(\tau) \alpha c_0 + a(\tau)^2 c_1 [1 - s(\alpha + 1)(\alpha + 2)] = 0.$$

Taking into account equality (8), it can be rewritten as

$$a'(\tau) = a(\tau)^2 \frac{2\omega}{\alpha^2}, \quad (11)$$

where for brevity we denote $\omega = c_1 / c_0$. Equation (11) is a differential equation with respect to the function $a(\tau)$. The solution of the equation has the form

$$a(\tau) = -\frac{\alpha^2}{2\omega\tau + \eta}, \quad (12)$$

up to a constant η , which, if necessary, is determined from the properties of the bond price. We note that it follows from (11) that

$$\frac{a'(\tau)}{a(\tau)^2} = \frac{2\omega}{\alpha^2}.$$

Now consider Eq. (9) for an arbitrary $n \geq 1$. It can be written as a recurrence relation that determines the coefficient c_{n+1} in terms of the coefficient c_n :

$$c_{n+1} = \frac{2(\alpha + n)\omega}{[s(\alpha + n + 1)(\alpha + n + 2) - 1]\alpha^2} c_n, \quad (13)$$

Note that

$$\frac{2(\alpha + n)}{[s(\alpha + n + 1)(\alpha + n + 2) - 1]\alpha^2} = \frac{\theta}{n + 1} \left(\frac{n + \alpha}{n + \xi} \right),$$

where for brevity is denoted $\xi = 2(\alpha + 1)$, $\theta = \xi/\alpha$. Thus, the sequence of coefficients $\{c_n, n = 0, 1, 2, \dots\}$ is as follows

$$c_0, \quad c_1 = c_0 \omega = c_0 \omega \theta \frac{\alpha}{\xi}, \quad c_2 = c_0 \frac{\omega \theta}{2} \frac{(1 + \alpha)}{(1 + \xi)},$$

$$c_3 = c_0 \frac{(\omega \theta)^2}{1 \times 2 \times 3} \frac{(1 + \alpha)(2 + \alpha)}{(1 + \xi)(2 + \xi)}, \quad \dots, \quad c_n = c_0 \frac{(\omega \theta)^{n-1}}{n!} \prod_{k=1}^{n-1} \frac{(k + \alpha)}{(k + \xi)}, \quad \dots$$

Then the solution (2) of Eq. (7) can be represented in the form

$$P(r, \tau) = c_0 \left(\frac{a(\tau)}{r} \right)^\alpha \left[1 + \left(\frac{\omega \theta a(\tau)}{r} \right) \frac{\alpha}{\xi} + \sum_{n=2}^{\infty} \left(\frac{\omega \theta a(\tau)}{r} \right)^n \frac{1}{n!} \prod_{k=0}^{n-1} \frac{(k + \alpha)}{(k + \xi)} \right].$$

We note that among the special functions there is a so-called confluent hypergeometric function (Kummer function) ${}_1F_1(x, y, z)$ (in the notation of the Wolfram Mathematica system), which is defined by

$${}_1F_1(x, y, z) = 1 + \sum_{n=1}^{\infty} \frac{z^n}{n!} \prod_{k=1}^n \frac{(x + k - 1)}{(y + k - 1)} = 1 + \frac{\Gamma(y)}{\Gamma(x)} \sum_{n=1}^{\infty} \frac{z^n}{n!} \frac{\Gamma(x + n)}{\Gamma(y + n)}.$$

Using these notations, the price $P(r, \tau)$ can be written in the form

$$P(r, \tau) = c_0 \left(\frac{a(\tau)}{r} \right)^\alpha \left(1 + \frac{\Gamma(\xi)}{\Gamma(\alpha)} \sum_{n=1}^{\infty} \frac{1}{n!} \left(\omega \theta \frac{a(\tau)}{r} \right)^n \frac{\Gamma(\alpha + n)}{\Gamma(\xi + n)} \right)$$

$$= c_0 \left(\frac{a(\tau)}{r} \right)^\alpha {}_1F_1 \left(\alpha, \xi, \omega \theta \frac{a(\tau)}{r} \right).$$

In terms of its economic properties, the bond price as a function of the maturity term τ is a continuous monotonically decreasing function that for any $r > 0$ has limits [9]

$$\lim_{\tau \rightarrow 0} P(r, \tau) = 1, \quad \lim_{\tau \rightarrow \infty} P(r, \tau) = 0.$$

These requirements can be satisfied by determining the so far undetermined constants c_0 and η by appropriate way. The final expression for the price of the zero-coupon bond becomes

$$P(r, \tau) = \frac{(1 + \alpha)\sqrt{\pi}}{2^{1+2\alpha} \Gamma(\alpha + 1.5)} \left(\frac{1}{sr\tau} \right)^\alpha {}_1F_1 \left(\alpha, 2(1 + \alpha), -\frac{1}{sr\tau} \right), \quad (14)$$

where $s \equiv 0.5\sigma^2$, $\alpha = 0.5 \left(\sqrt{1 + 4/s} - 1 \right) > 0$, and $\Gamma(x)$ is gamma function. Here it is assumed that $\alpha > 0$. When $\alpha < 0$, the gamma function $\Gamma(x)$ used in formula (14) can have undesirable properties. For example, for integer negative values of an argument, it has unbounded discontinuities, on intervals $(2\kappa, 2\kappa + 1)$, $\kappa = 0, 1, 2, \dots$, it is negative, etc., which is not corresponds to the properties of the bond price. Therefore, negative values of the parameter α are undesirable.

Typically, the term structure is in practice not represented through the bond price, but through yield. By definition, the yield to maturity of the zero-coupon bond (yield curve) $y(r, \tau)$ and the yield of the forward rates (forward curve) $f(r, \tau)$ are determined by the expressions [5]:

$$y(r, \tau) = -\frac{\ln P(r, \tau)}{\tau}, \quad f(r, \tau) = -\frac{\partial \ln P(r, \tau)}{\partial \tau} \quad (15)$$

and, unfortunately, are not presented in a compact analytical form and can only be investigated numerically.

4 The Ahn – Gao Model

Now let the polynomials $m(r)$ and $s(r)$ be such that $p = 2, q = 3$, that is $1 - p = 2 - q = -1$. Then the components of the sum (9) differ from zero only for $j \geq -1$, where the first term differs from zero only for $j \geq 0$. In this case we obtain the following system of equations:

for $j = -1$

$$-c_0 - \alpha m_2 c_0 + \alpha(\alpha + 1) s_3 c_0 = 0; \quad (16)$$

for $j = 0$

$$\begin{aligned} & -\alpha a'(\tau) a(\tau)^{-1} c_0 - a(\tau) c_1 - \sum_{k=1}^2 (\alpha + k - 1) m_k a(\tau)^{k-1} c_{k-1} \\ & + \sum_{k=2}^3 (\alpha + k - 2)(\alpha + k - 1) s_k a(\tau)^{k-2} c_{k-2} = 0; \end{aligned} \quad (17)$$

for $j = 1$

$$\begin{aligned} & -(\alpha + 1) a'(\tau) c_1 - a(\tau)^2 c_2 - \sum_{k=0}^2 (\alpha + k) m_k a(\tau)^k c_k \\ & + \sum_{k=1}^3 (\alpha + k - 1)(\alpha + k) s_k a(\tau)^{k-1} c_{k-1} = 0; \end{aligned} \quad (18)$$

for $j > 1$

$$\begin{aligned}
& -(\alpha + j)a'(\tau)a(\tau)^{j-1}c_j - a(\tau)^{j+1}c_{j+1} - \sum_{k=0}^2(\alpha + j + k - 1)m_k a(\tau)^{j+k-1}c_{j+k-1} \\
& + \sum_{k=0}^3(\alpha + j + k - 2)(\alpha + j + k - 1)s_k a(\tau)^{j+k-2}c_{j+k-2} = 0.
\end{aligned} \tag{19}$$

From the Eq. (16), which under the assumption that $c_0 \neq 0$ has the form $\alpha(\alpha + 1)s_3 = \alpha m_2 + 1$, the parameter α is determined.

$$\begin{aligned}
\alpha_1 &= \frac{1}{2s_3} \left(m_2 - s_3 - \sqrt{4s_3 + (m_2 - s_3)^2} \right), \\
\alpha_2 &= \frac{1}{2s_3} \left(m_2 - s_3 + \sqrt{4s_3 + (m_2 - s_3)^2} \right).
\end{aligned} \tag{20}$$

Since Eq. (16) is quadratic, it has two roots, which means that the solution of Eq. (1) can have two components of the form (2), a compromise between them, and also the initial condition $P(r, 0) = 1$ can affect on the choice of the coefficient c_0 .

Equation (17) is an ordinary differential equation with respect to the function $a(\tau)$. Its solution has the form

$$a(\tau) = \frac{\lambda}{\mu + \exp[(\tau + \alpha\xi c_0)(m_1 - (\alpha + 1)s_2)]}, \tag{21}$$

where for compactness we denote by $\lambda = \alpha c_0((1 + \alpha)s_2\alpha - m_1)$, $\mu = c_1(1 + (\alpha + 1)m_2 - (\alpha + 1)(\alpha + 2)s_3)$, and ξ is a constant integration of the differential equation, whose choice is made depending on the properties of the solution of Eq. (1).

Equation (18) determines the coefficient c_2 , and Eq. (19) can be considered as the basis for constructing a recurrence formula for calculating the coefficients c_{n+1} in terms of the previous coefficients c_j , $j \leq n$. Consider first the Eq. (19). It allows us to express the coefficient c_{j+1} in terms of the previous coefficients c_j , c_{j-1} , c_{j-2} by the formula

$$\begin{aligned}
c_{j+1} &= \frac{1}{a(\tau)^3[(1 + \alpha + j)(2 + \alpha + j)s_3 - (1 + \alpha + j)m_2 - 1]} \\
& \times [a(\tau)(\alpha + j)(a'(\tau) + a(\tau)(m_1 - (1 + \alpha + j)s_2))c_j \\
& - [a(\tau)(-m_0 + (\alpha + j)s_1))c_{j-1} + (\alpha + j - 2)s_0c_{j-2}](\alpha + j - 1)].
\end{aligned} \tag{22}$$

However, by the definition of the coefficients c_n in the expression (2), they must be constant coefficients independent of the variable τ . This means that in the formula (22) the right-hand side of the equality must not depend on τ . This is only if $m_0 = 0$, $s_0 = 0$, $s_1 = 0$, $s_2 = 0$. This requirement is a necessary

condition for the existence of a non-trivial solution (2), which says that a non-trivial solution does not hold for any polynomials $m(r)$ and $s(r)$ are of order 2 and 3, respectively, but only for

$$m(r) = m_1 r + m_2 r^2, \quad s(r) = s_3 r^3. \quad (23)$$

Substitution of the required necessary conditions into the formula (22) for the coefficient c_{n+1} leads to the recurrence relation

$$c_{n+1} = \frac{\alpha + n}{(1 + \alpha + n)(2 + \alpha + n)s_3 - (1 + \alpha + n)m_2 - 1} \frac{a'(\tau) + a(\tau)m_1}{a(\tau)^2} c_n. \quad (24)$$

We note that the denominator of the first factor of the right-hand side of (24) can be represented in the form

$$(1 + \alpha + n)(2 + \alpha + n)s_3 - (1 + \alpha + n)m_2 - 1 = s_3(1 + n)(\beta + n),$$

where

$$\beta = \begin{cases} \beta_1 \equiv \frac{1}{s_3} \left(s_3 - \sqrt{4s_3 + (m_2 - s_3)^2} \right) & \text{for } \alpha = \alpha_1, \\ \beta_2 \equiv \frac{1}{s_3} \left(s_3 + \sqrt{4s_3 + (m_2 - s_3)^2} \right) & \text{for } \alpha = \alpha_2. \end{cases} \quad (25)$$

When the necessary conditions are fulfilled, the function $a(\tau)$, determined by the formula (21), is somewhat simplified

$$a(\tau) = \frac{\lambda}{\mu + \exp[(\tau + \alpha\xi c_0)m_1]}, \quad (26)$$

where $\lambda = -\alpha c_0 m_1$, $\mu = (1 + (\alpha + 1)m_2 - (\alpha + 1)(\alpha + 2)s_3)c_1$. Substituting into the right-hand side of (24) an explicit expression for the function $a(\tau)$, determined by formula (15), we obtain

$$\frac{a'(\tau) + a(\tau)m_1}{a(\tau)^2} = \frac{\mu m_1}{\lambda} = \frac{s_3 \beta \omega}{\alpha},$$

where $\omega \equiv c_1/c_0$. In this case, the dependence on the variable τ on the right-hand side of formula (24) vanishes. Thus, the recurrence formula (24) for the coefficient c_{n+1} is transformed to the final form

$$c_{n+1} = \frac{\beta(\alpha + n)\omega c_n}{\alpha(1 + n)(\beta + n)}. \quad (27)$$

Now we turn to the solution of the last Eq. (18), from which it is necessary to determine c_2 . Since there are $s_0 = 0$ among the necessary conditions, Eq. (18) will coincide with Eq. (15) for $n = 1$ and therefore the coefficient c_2 is calculated from formula (26) for $n = 1$.

It turns out that if the polynomials $m(r)$ and $s(r)$ of the order 2 and 3, respectively, are determined by the expressions (23), the solution of Eq.,(1) can

be represented as the sum of two series of the type (2), each of which has the following structure

$$\left(\frac{a(\tau)}{r}\right)^\alpha c_0 \left(1 + \sum_{n=1}^{\infty} \left(\frac{a(\tau)}{r} \frac{\omega\beta}{\alpha}\right)^n \frac{1}{n!} \prod_{k=0}^{n-1} \frac{\alpha+k}{\beta+k}\right).$$

Using again the confluent hypergeometric function, the result can be compactly written in the analytical form

$$c_0 \left(\frac{a(\tau)}{r}\right)^\alpha {}_1F_1\left(\alpha, \beta, \frac{a(\tau)}{r} \frac{\omega\beta}{\alpha}\right). \quad (28)$$

As already mentioned, since Eq. (16) has two solutions (20), the solution of Eq. (1) can consist of two components of the form (2) with different sets of parameters (α, β) , whose values are determined by formulas (20) and (25):

$$\begin{aligned} P(r, \tau) &= c_{01} \left(\frac{a(\tau)}{r}\right)^{\alpha_1} {}_1F_1\left(\alpha_1, \beta_1, \frac{a(\tau)}{r} \frac{\omega\beta_1}{\alpha_1}\right) \\ &+ c_{02} \left(\frac{a(\tau)}{r}\right)^{\alpha_2} {}_1F_1\left(\alpha_2, \beta_2, \frac{a(\tau)}{r} \frac{\omega\beta_2}{\alpha_2}\right). \end{aligned} \quad (29)$$

Before concretizing the solution, we will make some preliminary analysis. First, we consider the properties of the diffusion process, given by drift and volatility, determined by the functions (1) and (23). According to the assumptions made, the process of short-term interest rate $r(t)$, corresponding to these functions, is described by equation

$$dr(t) = (m_1 r(t) + m_2 r(t)^2)dt + \sqrt{2s_3} r(t)^{3/2} dw.$$

The marginal probability density of this process has the form

$$f(r) = \frac{\delta^{2-\gamma} e^{-\delta/r}}{r^{3-\gamma} \Gamma(2-\gamma)}, \quad \delta = \frac{m_1}{s_3} > 0, \quad \gamma = \frac{m_2}{s_3} < 2, \quad s_3 > 0, \quad r \geq 0,$$

where $\Gamma(x)$ is the gamma function. Taking into account these inequalities, we note that the parameters of expression (28), according to formulas (20) and (25), take the values $\alpha_1 < 0$, $\alpha_2 > 0$, $\beta_2 > 0$, and β_1 can take positive values only when the volatility parameter is $s_3 > 4$, which practically does not occur in real cases.

As is well known, the bond price for $r > 0$ is a monotonically decreasing function with respect to the variable $\tau \in (0, \infty)$ from $P(r, 0) = 1$ to $P(r, \infty) = 0$. Therefore, expression (28) must have the same properties. The function ${}_1F_1(x, y, z)$ has suitable properties only for $x > 0$, $y > 0$, $z \in (-\infty, 0)$. Therefore, the first term in the representation (29) must be absent. In addition, for the argument z to ${}_1F_1$ to take values in the interval $(-\infty, 0)$ as τ changes in the interval $(0, \infty)$, it is necessary to define the integration constant ξ in expression (15) by the equality $\xi = \ln(\beta\omega s_3)/\alpha c_0 m_1$. Then

$$a(\tau) = \frac{\lambda}{\mu + \exp[(\tau + \alpha\xi c_0)m_1]} = \frac{-m_1}{(e^{m_1\tau} - 1)s_3}.$$

Finally, in order for the requirement $P(r, 0) = 1$, to be satisfied, it is necessary that the uncertain so far parameter c_0 be defined by the equality $c_0 = \Gamma(\beta - \alpha)/\Gamma(\beta)$. Thus, the final form of the solution (2) of Eq. (1) in the case under consideration has a final form

$$P(r, \tau) = \frac{\Gamma(\beta - \alpha)}{\Gamma(\beta)} \left(\frac{m_1}{r s_3 (e^{m_1 \tau} - 1)} \right)^\alpha {}_1F_1 \left(\alpha, \beta, \frac{-m_1}{r s_3 (e^{m_1 \tau} - 1)} \right),$$

where the parameters α and β are determined by means of formulas (20) and (25):

$$\begin{aligned} \alpha &= \frac{1}{2s_3} \left(m_2 - s_3 + \sqrt{4s_3 + (m_2 - s_3)^2} \right) > 0, \\ \beta &= \frac{1}{s_3} \left(s_3 + \sqrt{4s_3 + (m_2 - s_3)^2} \right) > 0. \end{aligned}$$

We note that this solution completely coincides with the solution obtained in another way by Ahn and Gao [3], where in the notation of these authors $m_1 = \kappa\theta - \lambda_1 > 0$, $m_2 = -\kappa - \lambda_2 < 0$, $s_3 = \sigma^2/2$. In principle, using expression (18), we can find, by formulas (15), analytical expressions for the yield curve $y(r, \tau)$ and the forward curve $f(r, \tau)$. However, these expressions are very cumbersome and more practical to use numerical methods for expressing these functions for the necessary numerical parameters.

The functions $y(r, \tau)$ and $f(r, \tau)$, defined by formulas (15) in terms of the representation $P(r, \tau)$, can be investigated only by numerical methods. True, the limiting values of these functions can be found in an analytical form:

$$\lim_{\tau \rightarrow 0} y(r, \tau) = \lim_{\tau \rightarrow 0} f(r, \tau) = r, \quad \lim_{\tau \rightarrow \infty} y(r, \tau) = \lim_{\tau \rightarrow \infty} f(r, \tau) = \alpha m.$$

As you can see, the left limit is determined only by the state of the market and does not depend on the model parameters, and the right limit is determined only by the structure of the model and does not depend on the state of the market at a certain moment in time.

5 Conclusion

The article presents models for which yield curves of zero-coupon bonds and corresponding forward curves can be found that are not related to the class of affine models. Unfortunately, models that admit such solutions are few and, in particular, include some well-known models: the CIR(1980) model [2] and the Ahn-Gao model [3]. Let us formulate the requirements for the structure of the short-term interest rate model, which would allow obtaining the term structure of the bond price in the form (2).

The parameters of the series (2) are determined by the Eq. (9), in fact, from which we obtain a system of equations with respect to the unknowns α , $a(\tau)$ and c_n , $n = 0, 1, 2, \dots$

1. To obtain a non-trivial solution (that is, for the presence of $c_n \neq 0$), it is necessary that the degrees p and q of the polynomials $m(r)$ and $s(r)$, determining the drift and volatility of the short-term interest rate, satisfy one of the following conditions: $\{p \leq 2, q = 3\}$, $\{p = 2, q \leq 3\}$, $\{p > 2, q = p + 1\}$. In these cases, the equations are found from which the positive parameter is determined.
2. Another necessary condition is related to the existence of $a(\tau)$, which does not depend on the summation index of the series (2).
3. In addition, it is necessary that the coefficients $\{c_n\}$ do not depend on the variable τ .

Simultaneous fulfillment of these necessary conditions significantly narrows the family of models for which the solution of the term structure equation (1) has the form (2).

References

1. Medvedev, G.A.: Polynomial models of yield term structure. Tomsk State Univ. J. Control Comput. Sci. **2**(39), 39–48 (2017)
2. Cox, J.C., Ingersoll, J.E., Ross, S.A.: An analysis of variable rate loan contracts. J. Financ. **35**, 389–403 (1980)
3. Ahn, D.-H., Gao, B.: A parametric nonlinear model of term structure dynamics. Rev. Financ. Stud. **12**(4), 721–762 (1999)
4. Vasicek, O.A.: An equilibrium characterization of the term structure. J. Financ. Econ. **5**, 177–188 (1977)
5. Keller-Ressel, M., Steiner, T.: Yield curve shapes and the asymptotic short rate distribution in affine one-factor models. Financ. Stochast. **12**(2), 149–172 (2008)



On Gaussian Approximation of Queueing Networks with Different Starting Load

Eugene Lebedev and Hanna Livinska^(✉)

Applied Statistics Department, Taras Shevchenko National University of Kyiv,
Volodymyrska str., 64, Kyiv 01601, Ukraine
leb@unicyb.kiev.ua, livinskaav@gmail.com
<http://applstat.univ.kiev.ua>

Abstract. In this work Markov multi-channel queueing networks with rate of the external load varying with time are considered. It is assumed that the starting load in the network can be not only a fixed constant value, but also can be asymptotically increasing in a series scheme. A many-dimensional service process of the network is introduced as the number of calls in the network nodes at the corresponding instant of time. For the service process, approximating Gaussian processes are constructed for both cases of starting load, when the network operates in heavy traffic regime. Correlation characteristics of the limit processes are written via the network parameters. It is proved that a many-dimensional Ornstein-Uhlenbeck process approximates the service process if the number of calls in the network nodes is asymptotically large at the initial instant of time.

Keywords: Multi-channel queueing network
Gaussian approximation · Heavy traffic regime

1 Introduction

In recent years, network research has acquired new practical importance as a primary tool for studying, designing and optimizing real-world systems with interacting components for which queueing networks provide a simple but extremely useful representation. For instance, the Internet can be represented as a computer network consisting of provider communication nodes, web servers, transmission stations and connected through data exchange. The role of the networks is constantly increasing in epidemiology, genetics, economics (insurance, logistics), in the study of cellular communication networks, computer viruses, computer support.

The apparatus of the theory of queueing networks is used at all levels of organization of network structures: when designing their topology, when developing protocols, when choosing switching methods and algorithms for routing information flows. Often the functioning of such networks is described by stochastic

models with parameters that are functions of time. Analysis of systems and networks with time-varying parameters is a complex mathematical problem. Until now, no universal methods for studying such models have been found. Therefore, there is need to develop methods that would be effective at least for certain types of stochastic networks.

Time-dependence of network rates yields additional difficulties in investigation of the corresponding models. This aspect motivates using nonstationary (in time) processes for their simulation. A number of publications (see, for example, [1–3, 7, 17], etc.) contains studies in this field.

The main approach to studying of queueing networks is based upon the direct method of finding expressions for the network state probabilities. It allows us to find the exact solution for multiplicative or, as they are often called, locally-balanced networks, stationary probabilities of states of which have multiplicative form. An alternative approach is to use the method of the asymptotic analysis for the researching of complex systems and networks. Different asymptotic approaches under many variants of heavy traffic conditions for studying networks are used in [1, 2, 4, 9, 10, 12, 15–17] etc.

In this paper, our approach is related to functional limit theorems for multi-channel queueing networks. We consider models of multi-channel queueing networks where an input flow is a nonhomogeneous Poisson flow whose rate depends on time. Such a flow is in good agreement with the actual flows in many situations, even if there are deviations of real flows from the Poisson one. In particular, we study the network with rate of an input flow depending on time periodically. Such dependence is typical for real-life networks. Each node operates as a multi-channel queueing system. Once the call service is completed in a node, the call is transferred to another node or it leaves the network with the corresponding probabilities. Call service times are independent random values with exponential type distributions. It is proved that under some heavy traffic assumptions on network parameters, the many-dimensional service process that is the number of calls in the nodes at the corresponding instant of time converges to a Gaussian process in the uniform topology.

Moreover, two types of starting load of the network are considered. At first, we assume that at the initial instant of time the network is empty. Actually, the formulated there result is valid not only for the completely empty network, but for the networks that are “moderately” downloaded from the start. It means that the number of calls at the initial instant is equal to the given constant value, does not depend on series number and, therefore, cannot affect the limit service process. The totally different case is when the starting load of the network is asymptotically increasing in series scheme. In such a situation the loading process can yields substantial changes in the representation of limit process. In this case we give conditions when a many-dimensional Ornstein-Uhlenbeck process can be obtained as the limit of the service process. Correlation characteristics of the Gaussian processes in both cases are written via network parameters in an explicit form.

Note, that such a network model with periodical input flow were considered in [13]. In this paper transient and stationary regimes for the network are studied. For the service process of calls, quasi-ergodic distribution is found. In transient regime, generating function of the service process is obtained, and moments of the service process are calculated. Network in heavy traffic regime is considered as well, prove of the limit theorem in case with fixed starting load, that is omitted in the present work, are given. In present work we formulate in addition our criterium for many-dimensional Gaussian processes, which enables us to prove Markovian property of the limit process in case of asymptotically large starting load. We also consider here more general case of networks with separate input flows arriving at each network node with asymptotically large starting load without the periodicity condition. Result about the correspondent service process approximation by a many-dimensional Ornstein-Uhlenbeck process is obtained.

This paper is organized as follows. In Sect. 2 we give description for the basic model of a multi-channel network with a periodical input flow when it is initially empty. The condition of a heavy traffic regime is formulated as a condition for the service parameters. Theorem about convergence of normalized service process to a Gaussian process in heavy traffic regime is given. In Sect. 3, results about convergence of the service process for the network with periodical input flow and with asymptotically large initial loading as well as some auxiliary results are presented. Additional condition for starting load is formulated in order for the limit process to be a many-dimensional Ornstein-Uhlenbeck process. In Sect. 4 the similar result for the general model with non-homogeneous Poisson input flow without periodicity condition is obtained. The conclusions and some suggestions for future research are given in Sect. 5.

2 Limit Process in Case of Fixed Starting Load

2.1 Model Description

The main model under consideration is a queueing network consisting of r service nodes. Each of the r nodes operates as a multi-channel stochastic system. If a call arrives at such a system then its service begins immediately. Service times of calls are random variables with their distribution depending on the node number.

From the outside, a periodic nonhomogeneous Poisson flow of calls $\nu(t)$, $t \geq 0$, with the leading function arrives to the network. $A(t)$, $t \geq 0$, is a positive nondecreasing right-continuous function. At first, we suppose that the input flow is a regular Poisson flow, and $A(t)$ is assumed to be an completely continuous with density function $\lambda(t)$, $t \geq 0$. It means that it can be written at the form

$$A(t) = \int_0^t \lambda(u) du.$$

Moreover, let $\lambda(t)$ be a periodic function with the period T :

$$\lambda(nT + \theta) = \lambda(\theta) \quad \text{for } n = 1, 2, \dots \quad \text{and } 0 \leq \theta < T.$$

A call arriving to the network is directed to the i -th node with probability p_{0i} , $i = 1, 2, \dots, r$. The service time in the i -th node is distributed exponentially with the rate μ_i , $i = 1, 2, \dots, r$. Once the service is completed in the i -th node, the call is directed to the j -th node with probability p_{ij} , or it leaves the network with probability $p_{ir+1} = 1 - \sum_{j=1}^r p_{ij}$. Denote by $P = \|p_{ij}\|_1^r$ the switching matrix of the network. An additional node numbered by $r+1$ is interpreted as an “exit” from the network.

According to common notation adopted in the theory of queueing networks, the above model is denoted by $[M_t|M|\infty]^r$.

Denote by $Q_i(t)$, $i = 1, 2, \dots, r$, $t \geq 0$, the number of calls in the i -th node of the network at the instant t . The service process of calls in the network of the $[M_t|M|\infty]^r$ -type will be defined as an r -dimensional process $Q(t)' = (Q_1(t), \dots, Q_r(t))$. Symbol $'$ stands for the transposed vector.

Our main purpose is to study the service process in heavy traffic regime of the network with a fixed initial load and with the initial load, that is “asymptotically large”. We identify the conditions under which this process can be approximated by the r -dimensional Ornstein-Uhlenbeck process as well.

2.2 Heavy Traffic Condition and Limit Theorem for Service Process

The heavy traffic regime for the $[M_t|M|\infty]^r$ -network with periodical input means that service rates at the network nodes depend on a series number n such that the following condition is fulfilled.

Condition 1. $\lim_{n \rightarrow \infty} n\mu_i^{(n)} = \mu_i > 0$, $i = 1, 2, \dots, r$.

In the context of Condition 1 we consider the sequence of stochastic processes:

$$\xi^{(n)}(t) = n^{-1/2}(Q^{(n)}(nt) - q^{(n)}(nt)), t \geq 0,$$

where

$$q^{(n)'}(nt) = (q_1^{(n)}(nt), \dots, q_r^{(n)}(nt)), q_j^{(n)}(nt) = \sum_{i=1}^r p_{0i} \int_0^{nt} p_j^{i(n)}(nt-u) \lambda(u) du,$$

$j = 1, \dots, r$, $p_j^{i(n)}(\tau)$ are the entries of the matrix $P^{(n)}(\tau) = \|p_j^{i(n)}(\tau)\|_1^r = \exp\{\Delta(\mu^{(n)})(P - I)\tau\}$, $\mu^{(n)'} = (\mu_1^{(n)}, \dots, \mu_r^{(n)})$; $\Delta(x) = \|\delta_{ij}x_i\|_1^r$ is a diagonal matrix with a vector $x' = (x_1, \dots, x_r)$ at the principal diagonal, $I = \|\delta_{ij}\|_1^r$ is an identity matrix.

In order to describe the limit of the sequence of stochastic processes $\xi^{(n)}(t)$, $n \rightarrow \infty$, we introduce two independent Gaussian processes

$$\xi^{(i)'}(t) = (\xi_1^{(i)}(t), \dots, \xi_r^{(i)}(t)), \quad i = 1, 2.$$

The process $\xi^{(1)}(t)$ is determined by the mean values:

$$E\xi^{(1)}(t) = 0$$

and by the correlation matrixes:

$$R^{(1)}(t) = E\xi^{(1)}(t)\xi^{(1)'}(t) - E\xi^{(1)}(t)E\xi^{(1)'}(t) = \int_0^t P'(u)\Delta(\bar{\lambda})P(u)du,$$

$$R^{(1)}(s, t) = E\xi^{(1)}(s)\xi^{(1)'}(t) - E\xi^{(1)}(s)E\xi^{(1)'}(t) = R^{(1)}(s)P(t - s), \quad s < t,$$

where $\bar{\lambda}' = (\lambda_1, \dots, \lambda_r)$, $\lambda_i = p_{0i} \int_0^T \lambda(u)du$, $P(\tau) = \exp\{\Delta(\mu)(P - I)\tau\}$.

For the process $\xi^{(2)}(t)$

$$E\xi^{(2)}(t) = 0,$$

$$R^{(2)}(t) = \int_0^t [\Delta(\bar{\lambda}'P(u)) - P'(u)\Delta(\bar{\lambda})P(u)] du,$$

$$R^{(2)}(s, t) = R^{(2)}(s)P(t - s), \quad s < t.$$

The following result give an approximation for the service process of the $[M_t|M|\infty]^r$ - network under heavy traffic Condition 1.

Theorem 1. *Let Condition 1 takes place for the $[M_t|M|\infty]^r$ - network, and let the network be empty at the initial instant $t = 0 : Q_i(0) = 0, i = 1, 2, \dots, r$. Then on any finite interval $[0, T]$ the sequence of stochastic processes $\xi^{(n)}(t), n \geq 1$, converges as $n \rightarrow \infty$ to the process $\xi^{(1)}(t) + \xi^{(2)}(t)$ in the uniform topology.*

The proof of Theorem 1 follows from some auxiliary results and can be found in [13]. It is easy observed that the limit is a Markov Gaussian process. Note, that the main feature of the limit Gaussian process $\xi(t)$, is the following: $\xi(t) = \xi^{(1)}(t) + \xi^{(2)}(t)$. The part $\xi^{(1)}(t)$ of the limit process is associated with fluctuations of an input flow and $\xi^{(2)}(t)$ is associated with fluctuations of service times.

3 Limit Theorem in Case of Asymptotically Large Starting Load

In assumptions about the above model, it was predicated that at the initial instant the network have to be empty. This assumption is not crucial, and the assertion of Theorem 1 will also held true for networks that are “moderately” loaded at the initial instant of time. This means that the number of calls initially located in each node does not depend on the series number n , so, it is limited and, accordingly, can not affect the limit service process.

The situation changes if at the initial instant of time the number of calls depends on the series number n and increases as $n \rightarrow \infty$. Then the service process of calls that were initially located at the network can converge to a non-zero limit, and this term can enter into the overall limit process.

So, now, for the $[M_t|M|\infty]^r$ -network operating in heavy traffic regime, we consider the case when the vector of initial conditions takes asymptotically large values.

In order to balance the parameters of the limit process components, we assume within this section that the $[M_t|M|\infty]^r$ -network is open. The formulation

of this network property in terms of a switching matrix leads to the following condition: the spectral radius of the switching matrix P is strictly less than 1.

We denote the solution of the balance equation for our network by $\theta' = (\theta_1, \dots, \theta_r) = \bar{\lambda}'(I - P)^{-1}$. We will require the fulfilment of the following condition instead of $Q_i^{(n)}(0) = 0$, $i = 1, 2, \dots, r$:

Condition 2. $Q_i^{(n)}(0) = \left[n\theta_i/\mu_i + \sqrt{n}\eta_i^{(0)} \right]$, $i = 1, 2, \dots, r$, where $\eta^{(0)' = (\eta_1^{(0)}, \dots, \eta_r^{(0)})$ is a fixed vector in \mathbb{R}^r , $[\cdot]$ is the integer part of a number.

Next, we study the asymptotic behavior of the sequence of stochastic processes

$$\eta^{(n)}(t) = n^{-1/2} \left(Q^{(n)}(nt) - n(\theta/\mu) \right), \quad t \geq 0, \quad n \geq 1,$$

$$(\theta/\mu)' = (\theta_1/\mu_1, \dots, \theta_r/\mu_r).$$

In this case, we introduce in addition a Gaussian process $\xi^{(3)}(t)$ that does not depend on the previously introduced processes $\xi^{(1)}(t)$ and $\xi^{(2)}(t)$. This process $\xi^{(3)}(t)$ is determined by the mean values

$$E\xi^{(3)}(t) = P'(t)\eta^{(0)},$$

and by correlation matrices

$$\begin{aligned} R^{(3)}(t) &= E\xi^{(3)}(t)\xi^{(3)'}(t) - E\xi^{(3)}(t)E\xi^{(3)'}(t) \\ &= \Delta((\theta/\mu)'P(t)) - P'(t)\Delta(\theta/\mu)P(t), \\ R^{(3)}(s, t) &= E\xi^{(3)}(s)\xi^{(3)'}(t) - E\xi^{(3)}(s)E\xi^{(3)'}(t) = R^{(3)}(s)P(t-s), \quad s < t. \end{aligned}$$

In order to constructively define the service process, we consider a Markov chain $x(t)$, $t \geq 0$, in the set of states $\{1, \dots, r, r+1\}$ with infinitesimal rates

$$a_{ij} = \begin{cases} -\mu_i(1 - p_{ii}), & \text{if } i = j = 1, \dots, r; \\ \mu_i p_{ij}, & \text{if } i \neq j, \quad i = 1, \dots, r, \quad j = 1, \dots, r, r+1; \\ 0, & \text{if } i = r+1, \quad j = 1, \dots, r, r+1; \end{cases}$$

and with the initial distribution $p'(0) = (p_1(0), \dots, p_{r+1}(0))$.

If $p_i(0) = 1$, then we mark the corresponding chain as $x^{(i)}(t)$. The state $r+1$ for the chain $x(t)$ is absorbing. Transient probabilities can be written as follows:

$$\begin{aligned} p_{ij}(t) &= P\{x(t) = j/x(0) = i\} = P\{x^{(i)}(t) = j\}, \quad i = j = 1, \dots, r, \\ P(t) &= \|p_{ij}(t)\|_1^r = \exp\{\Delta(\mu)(P - I)t\}. \end{aligned}$$

The trajectory of a call between the instant of arriving to the network through the i -th node and the exit instant of it can be described by the chain $x^{(i)}(t)$, and the absorption in the state $r+1$ is interpreted as an exit of the call from the network.

Let us connect with the chain $x^{(i)}(t)$, $t \geq 0$, an r -dimensional process of indicator type $\chi^{(i)'}(t) = \left(\chi_1^{(i)}(t), \dots, \chi_r^{(i)}(t)\right)$, $t \geq 0$, as follows:

$$\chi^{(i)}(t) = \begin{cases} e_j, & \text{if } x^{(i)}(t) = j, \quad j = 1, \dots, r; \\ e_0, & \text{if } x^{(i)}(t) = r + 1. \end{cases}$$

At this point e_0 is an r -dimensional vector with null entries, e_j is an r -dimensional vector whose j -th entry is equal to 1 while the rest are 0.

Let us denote by $\chi^{(i)k}(t)$, $t \geq 0$, $i = 1, 2, \dots, r$, $k = 1, 2, \dots$, independent stochastic processes of indicator type whose finite-dimensional distributions are the same as those of $\chi^{(i)'}(t) = \left(\chi_1^{(i)}(t), \dots, \chi_r^{(i)}(t)\right)$.

We study now the normalized service process of calls being at the nodes of the network at the initial instant $t = 0$:

$$\tilde{\eta}^{(n)}(t) = \frac{1}{\sqrt{n}} \left\{ \sum_{m=1}^r \frac{[n\Theta_m/\mu_m + \sqrt{n}\eta_m^{(0)}]}{\sum_{k=1}^r} \chi^{(m)k}(nt) - nP'(t)(\theta/\mu) \right\}.$$

Lemma 1. *On any finite time interval $[0, T]$ the sequence of stochastic processes $\tilde{\eta}^{(n)}(t)$ converges to the process $\xi^{(3)}(t)$ as $n \rightarrow \infty$ in the uniform topology.*

Proof. Taking into account that

$$\lim_{n \rightarrow \infty} E\tilde{\eta}^{(n)}(t) = P'(t)\eta^{(0)},$$

$$\lim_{n \rightarrow \infty} \left[E\tilde{\eta}^{(n)}(t)\tilde{\eta}^{(n)'}(t) - E\tilde{\eta}^{(n)}(t)E\tilde{\eta}^{(n)'}(t) \right] = \Delta [(\theta/\mu)' P(t)] - P'(t)\Delta(\theta/\mu) P(t),$$

$$\begin{aligned} \lim_{n \rightarrow \infty} \left[E\tilde{\eta}^{(n)}(s)\tilde{\eta}^{(n)'}(t) - E\tilde{\eta}^{(n)}(s)E\tilde{\eta}^{(n)'}(t) \right] \\ = \{ \Delta [(\theta/\mu)' P(s)] - P'(s)\Delta(\theta/\mu) P(s) \} P(t-s), \quad s < t, \end{aligned}$$

the conditions of the normal correlation theorem (see [5], p. 188) are fulfilled, and as a consequence we have the desired convergence of finite-dimensional distributions.

The process $\tilde{\eta}^{(n)}(t)$ has moments of any order. Therefore, in order to strengthen the convergence of finite-dimensional distributions to a convergence in the uniform topology, it is convenient to apply the criterion in [4], p. 179, with $F(t) = t$ and $\alpha = 3/4$, and to double use Chebyshev's inequality for upper bounding of the corresponding probabilities.

The lemma is proved.

Let $\tilde{Q}^{(n)'}(t) = \left(\tilde{Q}_1^{(n)}(t), \dots, \tilde{Q}_r^{(n)}(t)\right)$, $t \geq 0$, be an r -dimensional stochastic process with distribution coinciding with the distribution of service process in the $[M_t|M|\infty]^r$ -network with the periodic input flow in the case when at the initial instant the network is empty.

As a consequence of Theorem 1 we obtain the following result.

Lemma 2. *Let for the $[M_t|M]_\infty^r$ -network with a periodic input flow Condition 1 is held and spectral radius of the switching matrix P is strictly less than 1. Then the sequence of stochastic processes*

$$\tilde{\xi}^{(n)'}(t) = n^{-1/2} \left[\tilde{Q}^{(n)'}(nt) - n(\theta/\mu)'(I - P(t)) \right]$$

converges as $n \rightarrow \infty$ to $\xi^{(1)}(t) + \xi^{(2)}(t)$ in the uniform topology on any finite interval $[0, T]$.

Proof. For the above model, $\Lambda_i(t) = p_{0i} \int_0^t \lambda(u) du$, $i = 1, 2, \dots, r$, is the leading function of nonhomogeneous Poisson process arriving from the outside into the i -th node of the $[M_t|M]_\infty^r$ -network. Denote by $\bar{\Lambda}'(t) = (\Lambda_1(t), \dots, \Lambda_r(t))'$. Then we obtain the following consequence from Theorem 1:

$$n^{-1/2} \left[\tilde{Q}^{(n)'}(nt) - \int_0^{nt} [d\bar{\Lambda}(u)]' P^{(n)}(nt - u) \right] \xrightarrow{U}_{n \rightarrow \infty} \xi^{(1)}(t) + \xi^{(2)}(t),$$

where symbol \xrightarrow{U} stands for convergence in the uniform topology.

Taking into account Theorem 3.1 from [4], p. 27, Lemma 2 will be proved if we show that

$$\begin{aligned} & \lim_{n \rightarrow \infty} \sup_{t \in [0, T]} n^{-1/2} \left| \int_0^{nt} [d\bar{\Lambda}(u)]' P^{(n)}(nt - u) - n(\theta/\mu)'(I - P(t)) \right| \\ &= \lim_{n \rightarrow \infty} \sup_{t \in [0, T]} n^{-1/2} \left| \int_0^t [d\bar{\Lambda}(nv)]' P(t - v) - n(\theta/\mu)'(I - P(t)) \right| = 0. \quad (1) \end{aligned}$$

On the one hand, we have

$$\begin{aligned} \int_0^t [d\bar{\Lambda}(nv)]' P(t - v) &= n \int_0^t [d(n^{-1}\bar{\Lambda}(nv))]' P(t - v) \\ &= n \left[n^{-1}\bar{\Lambda}'(nv) - \int_0^t n^{-1}\bar{\Lambda}'(nv) dP(t - v) \right]. \end{aligned}$$

On the other hand, we can write

$$\begin{aligned} \sup_{t \in [0, T]} |n^{-1}\bar{\Lambda}(nt) - \bar{\lambda}t| &= \sup_{t \in [0, T]} \left| n^{-1} \left(\bar{\lambda}[nt]_T + \int_0^{\{nt\}_T} \bar{\lambda}(u) du \right) - \bar{\lambda}t \right| \\ &= \sup_{t \in [0, T]} n^{-1} \left| \int_0^{\{nt\}_T} \bar{\lambda}(u) du - \bar{\lambda}\{nt\}_T \right| \leq n^{-1}(1 + T)|\bar{\lambda}|, \end{aligned}$$

where $\bar{\lambda}(t) = (\lambda_1(t), \dots, \lambda_r(t))'$.

So,

$$\begin{aligned} \sup_{t \in [0, T]} n^{1/2} \left| n^{-1}\bar{\Lambda}'(nt) - \int_0^t n^{-1}\bar{\Lambda}'(nv) dP(t - v) - (\theta/\mu)'(I - P(t)) \right| \\ = O\left(\frac{1}{\sqrt{n}}\right) \quad (2) \end{aligned}$$

Since

$$\begin{aligned}\bar{\lambda}'t - \bar{\lambda}' \int_0^t v dP(t-v) &= \bar{\lambda}' \int_0^t P(t-v) dv = \bar{\lambda}' \int_0^t P(v) dv \\ &= \bar{\lambda}'(I - P)^{-1} \Delta^{-1}(\mu)(I - P(t)) = (\theta/\mu)'(I - P(t)),\end{aligned}$$

it is obvious that estimation (2) is sufficient to obtain (1).

The lemma is proved.

Under Condition 2, the normalized service process $\eta^{(n)}(t)$, $t \geq 0$, for the $[M_t|M|\infty]^r$ -network can be represented as follows:

$$\eta^{(n)}(t) \triangleq \tilde{\eta}^{(n)}(t) + \tilde{\xi}^{(n)}(t), \quad t \in [0, T].$$

Moreover, $\tilde{\eta}^{(n)}(t)$ and $\tilde{\xi}^{(n)}(t)$ are independent stochastic processes.

Then, Lemmas 1 and 2 yields the following result.

Theorem 2. *Suppose that the $[M_t|M|\infty]^r$ -network with a periodical input flow has the spectral radius of its switching matrix strictly less than 1, and let Conditions 1, 2 be satisfied. Then the sequence of stochastic processes $\eta^{(n)}(t)$ converges as $n \rightarrow \infty$ to $\xi^{(1)}(t) + \xi^{(2)}(t) + \xi^{(3)}(t)$ in the uniform topology on any finite interval $[0, T]$.*

In comparison with the statement of Theorem 1, the additional summand $\xi^{(3)}(t)$ of the limit process is associated with fluctuations of service times of the calls located in the network nodes at the initial instant $t = 0$.

It is not difficult to check that for correlation matrices we have:

$$R^{(1)}(t) + R^{(2)}(t) + R^{(3)}(t) = \Delta(\theta/\mu) - P'(t)\Delta(\theta/\mu)P(t).$$

So, Theorem 2 implies the following result.

Corollary 1. *Suppose that for the $[M_t|M|\infty]^r$ -network conditions of Theorem 2 are fulfilled. Then the sequence of stochastic processes $\eta^{(n)}(t)$ converges as $n \rightarrow \infty$ to a stochastic process $\eta^{(0)}(t)$ in the uniform topology on any finite interval $[0, T]$, and $\eta^{(0)}(t)$ is an r -dimensional Ornstein-Uhlenbeck process in transient regime with $\eta^{(0)}(0) = \eta^{(0)}$ and the following correlation characteristics:*

$$E\eta^{(0)}(t) = P'(t)\eta^{(0)},$$

$$R^{(0)}(t) = E\eta^{(0)}(t)\eta^{(0)'}(t) - E\eta^{(0)}(t)E\eta^{(0)'}(t) = \Delta(\theta/\mu) - P'(t)\Delta(\theta/\mu)P(t),$$

$$R^{(0)}(s, t) = E\eta^{(0)}(s)\eta^{(0)'}(t) - E\eta^{(0)}(s)E\eta^{(0)'}(t) = R^{(0)}(s)P(t-s), \quad s < t.$$

Note, that by the definition an r -dimensional Ornstein-Uhlenbeck process is a Markov process (see [6], [8], p. 166). Markovian property of the limit process $\eta^{(0)}(t)$ follows from the next criterium for many-dimensional Gaussian processes ([9]).

Lemma 3. Let $\xi'(t) = (\xi_1(t), \xi_2(t), \dots, \xi_r(t)) \in R_r$ be the r -dimensional Gaussian process with zero mean and correlation matrices

$$R(t) = E\xi(t)\xi'(t) - E\xi(t)E\xi'(t), \quad R(s, t) = E\xi(s)\xi'(t) - E\xi(s)E\xi'(t), \quad s < t.$$

If for some matrix A and for all s, t ($0 \leq s < t$) the functions $R(s)$ and $R(s, t)$ relate to each other by the following way:

$$R(s, t) = R(s)P(t - s), \quad \text{where } P(t) = \exp(At),$$

then the Gaussian process $\xi(t)$ is a Markov process and the conditional distribution $P(\xi(t) \in B/\xi(s) = x)$, $B \in B_{R^r}$, is Gaussian with the mean vector $P'(t - s)x$ and the correlation matrix $R(t) - P'(t - s)R(s)P(t - s)$.

The set G_A of Gaussian processes for which the condition of Lemma 3 takes place and the corresponding matrices are the same (equal A) satisfies to the closure condition: a linear combination of two independent processes from G_A belongs G_A . Thus, as a consequence from Lemma 3 we obtain the following interesting fact: the sum of two independent Markov G_A -processes is a Markov process.

4 Ornstein-Uhlenbeck Approximation Without Periodicity Condition

Now we consider a network consisting of r nodes with input flows of calls $\nu_i(t)$, $i = 1, 2, \dots, r$, $t \geq 0$, arriving into each node separately. The flow $\nu_i(t)$, that arrives at the i -th node, is a nonhomogeneous Poisson process with a leading function $\Lambda_i(t)$. Assumption about periodicity is omitted. The routing algorithm and service characteristics are the same as in the model above. Such a network will be denoted by the symbol $[\overline{M}_i|M|\infty]^r$.

For the external input flows of calls, we consider the next condition.

Condition 3. For any $T > 0$

$$\sup_{t \in [0, T]} \left| n^{-1} \overline{\Lambda}^{(n)}(nt) - \bar{\lambda}t \right| = o(n^{-1/2}),$$

where $\bar{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_r)'$, $\lambda_i \geq 0$, $i = 1, 2, \dots, r$, and $\lambda_1 + \lambda_2 + \dots + \lambda_r \neq 0$.

Obviously, when Condition 3 is satisfied, external input flows into each network node $\nu_i(t)$, $i = 1, 2, \dots, r$, $t \geq 0$, are close to stationary Poisson processes with rates λ_i on a time scale nt .

Provided that the spectral radius of the switching matrix P is strictly less than 1, there exists a solution of the balance equation

$$\theta_i = \lambda_i + \sum_{j=1}^r \theta_j p_{ji}, \quad i = 1, 2, \dots, r.$$

Next, we consider a sequence of stochastic processes

$$\eta^{(n)}(t) = n^{-1/2} \left(Q^{(n)}(nt) - n(\theta/\mu) \right), \quad n = 1, 2, \dots .$$

For $\eta^{(n)}(t)$, the similar for Theorem 2 result takes place.

Theorem 3. *Let for the $[\overline{M}_t|M|\infty]^r$ -network the spectral radius of the switching matrix P be strictly less than 1 and let Conditions 1–3 be held. Then $\eta^{(n)}(t)$ converges as $n \rightarrow \infty$ in the uniform topology on any finite interval $[0, T]$ to the r -dimensional Ornstein-Uhlenbeck diffusion process $\eta^{(0)}(t)$ ($\eta^{(0)}(0) = \eta^{(0)}$) with the drift vector*

$$A(x) = (P' - I) \Delta(\mu)x,$$

and the diffusion matrix

$$B = \Delta(\theta) (P - I) + (P' - I) \Delta(\theta).$$

It is obvious that Theorem 2 is a special case of the last theorem.

5 Conclusions

In this work we investigate asymptotic behavior of the service process of calls for the stochastic network of the $[M_t|M|\infty]^r$ -type under the assumptions that the network operates in heavy traffic regime. Firstly, the external input flow of calls is supposed to be a Poisson flow with time-dependent instant value of its rate that is periodical function. The network is considered in two different cases of starting load. Firstly, the number of calls in the initial instant is assumed to be equal to zero. Secondly, this number is assumed to increase with a series number increasing. It is proved that under formulated conditions the service process has Gaussian process as a limit in the uniform topology. It is significant, that the limit process is decomposed into the sum of independent Gaussian processes. In case of the fixed initial load it has two summands $\xi^{(1)}(t)$ and $\xi^{(2)}(t)$. In case of asymptotically large load it can be written as $\xi^{(1)}(t) + \xi^{(2)}(t) + \xi^{(3)}(t)$, where the first summand $\xi^{(1)}(t)$ is associated with fluctuations of input flow, $\xi^{(2)}(t)$ is associated with fluctuations of service times of calls arrived from the outside, and $\xi^{(3)}(t)$ is connected with fluctuations of service times of calls located in the network nodes at the initial instant $t = 0$. In case of asymptotically large load, the limit process is a many-dimensional Ornstein-Uhlenbeck diffusion process. Similar result about the service process approximation by Ornstein-Uhlenbeck process is obtained also for more general model with separated input flows without periodicity condition. For all cases functional limit theorems on convergence in uniform topology are formulated, and the approximating Gaussian processes with its characteristics in explicit form are constructed.

Some studies of similar multi-channel stochastic networks with time-dependent input flow and different types of service can be found in [9, 10, 12]. Stochastic networks with input flow controlled by Markov process was investigated in [14].

Note, that in the work jump-wise service processes are approximated by continuous Gaussian processes of simpler structure. Moreover, all convergences are proved in the uniform topology. This type of convergence of stochastic processes gives us a possibility to calculate functionals related with the processes and to solve optimization problems (see, for example, [11]).

References

1. Anisimov, V.V., Lebedev, E.A.: Stochastic Queueing Networks. Markov Models. Lybid, Kyiv (1992). (in Russian)
2. Anisimov, V.V.: Switching Processes in Queueing Models. ISTE Ltd. (2008)
3. Basharin, G.P., Bocharov, P.P., Kogan, Y.A.: Analysis of Queues in Calculating Networks. Nauka, Moscow (1989). (in Russian)
4. Billingsley, P.: Convergence of Probability Measures. Willey, Hoboken (1999)
5. Gihman, I.I., Skorohod, A.V.: Theory of Stochastic Processes, vol. 1. Nauka, Moscow (1971). (in Russian)
6. Doob, J.L.: The Brounian movement and stochastic equations. Ann. Math. **43**(2), 351–369 (1942)
7. Korolyuk, V.S., Korolyuk, V.V.: Stochastic Models of Systems. Kluwer Acad. Press, Dordrecht (1999)
8. Kovalenko, I.N., Kuznetsov, I.N., Shurenkov, N.Yu.: Stochastic Processes. Naukova Dumka, Kyiv (1983). (in Russian)
9. Lebedev, E.O., Livinska, G.V.: Gaussian approximation of multi-channel networks in heavy traffic. Commun. Comput. Inform. Sci. **356**, 122–130 (2013)
10. Lebedev, E., Chechelnitsky, A., Livinska, H.: Multi-channel network with interdependent input flows in heavy traffic. Theor. Probab. Math. Stat. **97**, 109–119 (2017). (in Ukrainian)
11. Lebedev, E.A., Makushenko, I.A.: Risk Optimisation for Multi-Channel Stochastic Network. National Library of Ukraine, Kyiv (2007)
12. Livinska, H.V.: A limit theorem for non-Markovian multi-channel networks under heavy traffic conditions. Theor. Probab. Math. Stat. **93**, 113–122 (2016)
13. Livinska, H., Lebedev, E.: On transient and stationary regimes for multi-channel networks with periodic inputs. Appl. Stat. Comput. **319**, 13–23 (2018)
14. Livinska, H., Lebedev, E.: On a multi-channel stochastic network with controlled input. In: Applied Mathematics and Computer Science, vol. 1836, pp. 020052-1–020052-4. Melville, New York (2017)
15. Moiseev, A., Nazarov, A.: Queueing network MAP - $(GI | \infty)^K$ with high-rate arrivals. Eur. J. Oper. Res. **254**(1), 161–168 (2016)
16. Moiseev, A., Nazarov, A.: Asymptotic analysis of the infinite-server queueing system with high-rate semi-Markov arrivals. In: International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, pp. 507–513 (2015)
17. Nazarov, A.A., Moiseeva, S.P.: Method of Asymptotic Analysis in Queueing Theory. Sc.-techn. litr. publ, Tomsk (2006). (in Russian)



A Retrial Queueing System with Orbital Search of Customers Lost from an Offer Zone

Ambily P. Mathew, Achyutha Krishnamoorthy, and Varghese C. Joshua^(✉)

Department of Mathematics, CMS College, Kottayam 686001, Kerala, India
{ambilypm,krishnamoorthy,vcjoshua}@cmscollege.ac.in
<http://www.cmscollege.ac.in>

Abstract. A tandem retrial queueing system with orbital search in which two self-service stations namely, the main station and the offer zone and an orbit for passive customers lost from the offer zone without joining the main station is considered. The main service station is of infinite capacity while the offer zone which works in a random environment and the orbit for passive customers are of finite capacities. Two types of customers arrive to the service stations according to a Marked Markovian Arrival Process (MMAP) with representation (D_0, D_1, D_2) . The service times in both stations are exponentially distributed. A virtual search mechanism associated with the main station will be working when the number of customers in the main station is below a pre-assigned level L . The duration of search is exponentially distributed. The condition for system stability is established. The system state distribution in the steady state is obtained. Several system performance characteristics are derived. An associated optimization problem is investigated.

Keywords: Retrial queue · Tandem queue · Main station
Offer zone · Random environment · Passive customers · Orbit

1 Introduction

Tandem queues form an important class of the queueing networks and it serves as a link between the theory of queues and queueing networks. A bibliography of articles on queueing networks with finite capacity service stations can be found in [22]. Most of the literature in this regard assume that the service stations in the tandem network are of finite capacity and the time between successive arrivals to the system are exponentially distributed. [17] gives an algorithm for solving exponential tandem queues with blocking. In [11–13] multi-stage queueing networks with correlated arrivals are considered. In Krishnamoorthy et al. [16] considered a tandem queueing model with two service stations and one of which namely, the offer zone works in a random environment. Artalejo [1, 2] gives a detailed bibliography of retrial queues. The monograph by Falin and Templeton [10] gives an introduction to the theory of retrial queues and it describes how

the theory of retrial queues can well be applied in the analysis of problems which are more realistic as well as practically important. The present paper generalizes the model described in one of our papers [16] to a retrial set-up by setting up an orbit for holding the customers who leave the system after completing their service at the offer zone.

In the present paper, we consider a tandem retrial queueing network with two service stations and an orbit for those customers who discontinued their service after a trial service. This model is the mathematical formulation of a set of real life problems persisting in the field of telecommunication. In the field of telecommunication, various service providers compete for attracting the maximum number of customers to their paid service. Some customers directly enter the paid service while some others would like to make a previous trial of service before subscribing to paid service. Service providers announce various types of offers, incentives and free trials to make the maximum number of customers to continue with their service. They try to minimize the loss of customers from the subscription of their service. Not all the customers, who utilized the offers and free trials, move on to paid service. Some may continue with the same service provider as paid customers, while some others discontinue the service temporarily after the free-trial. But the service providers have a data-base consisting of those customers who discontinued service after the free trial and that may be considered as an orbit. Having discontinued with the service for a short time, a few may have a tendency to come back to paid service, which may be considered as retrials and those retrial rates may be small when compared to direct arrivals to the paid service. So the customers in the orbit may be designated as passive customers. The service providers need a minimum number of customers in paid service for the proper functioning of their system. So whenever the number of customers in the paid service drops down to this pre-assigned value, the service providers try to bring some more customers to the paid service by means of orbital search. This search can be any of the activities like contacting those passive customers over the telephone, sending e-mails, additional cash-back offers etc. Search may result in an additional increment in the number of paid customers and whenever it reaches the optimum level, no more search has been done. Since there is some cost associated with the search, an optimum of this level to switch on the search mechanism is to be found. This problem is modelled mathematically as a tandem retrial queueing system with orbital search in which two service stations, namely the main station and the offer zone are functioning. The main service station is of infinite capacity while the offer zone is of finite capacity. The most important feature of the finite capacity offer zone in our model is that it works in a finite number of random environments, each of which lasts for a time interval whose distribution is Phase Type. In [13] and [12] servers in the same station are independent and identical. In our model the servers of the same station are identical, but the rate at which service is offered at the offer zone depends on the current environmental status of the offer zone. In addition to the retrials from the orbit, search for customers start functioning when the number of customers in the main station drops down to a preassigned level.

In classical queueing models Neuts and Ramalhoto [19] introduced the concept of search of orbital customers by the server at the end of a service completion epoch. In the case of $M/G/1$ retrial queues, search of orbital customers was introduced by Artalejo et al. [3]. Analysis of multi server queues with orbital search was done by Chakravarthy et al. in [5]. Krishnamoorthy et. al [14] investigated $M/G/1$ Retrial queues with non persistent customers and orbital search. More literature related to orbital search can be found in [6, 7, 15]. We also assume that the arrivals to the main station and the offer zone is according to a Marked Markovian Arrival Process [MMAP]. In Krishnamoorthy et al. [15] considered a queueing system with MMAP arrivals. Steady state probabilities are computed using Neuts' Matrix Geometric methods [20]. The rate matrix is computed using Logarithmic reduction Algorithm [18]. Various methods for the calculation of the equilibrium distribution of LDQBD's can be found in the papers by Neuts and Rao [21], Bright and Taylor [4] and Ramaswamy [18]. The stability condition is established and the steady state distribution is computed. Several performance measures of the system that influences the efficiency are derived. The cost functions for optimizing the level at which the search mechanism is to be switched off is derived. The control problems that optimizes the maximum capacity of the offer zone as well as the orbit are analyzed.

2 Description of the Model

We consider a tandem retrial queueing system in which, there are two self-service stations namely, the main station and the offer zone. The main station and the offer zone provides the same kind of service. But the service at the offer zone is restricted, for example, some trial service and it can not be continued for as long as they like. But after completing their service at the offer zone, the customers can decide whether to continue their service at the main station or not. There are some restrictions on the period of time they can stay in service at the offer zone. There are two types of customers in this system, say Type A and Type B. Type A customers are those customers who directly enter the main station for service and they do not try to take a trial service. Type B customers are those customers who wish to have a trial service by entering the offer zone and after their service completion at the offer zone, they can decide whether to continue their service at the main station or to leave the system. The offer zone works in a random environment and the environments at the offer zone are designed in such a way to attract the maximum number of type B customers from the offer zone to the main station and to make them get served at the main station. The service at the main station contributes a revenue to the system, while the offer zone has some kind of establishment as well as holding cost associated with it for the proper functioning. After the service completion at the offer zone, Type B customers are assumed to continue their service at the main station with probability η and with its complimentary probability $(1 - \eta)$, joins an orbit of passive customers who temporarily discontinued service but retries for service after being idle for sometime. The customers in this orbit are referred to be

passive in the sense that their retrial rates are very low compared to the arrival rates to both the stations. Let ν be the rate at which retrials from the offer zone to the main station occur and it is assumed to be lower than the fundamental rates of arrivals to both the stations. For the proper functioning of the system a minimum of L customers are to be ensured at the main station and so whenever the number of customers in service at the main station is below this level L , a virtual search mechanism associated with the main station, starts working and it go in search of customers from the orbit of passive customers. This may be by providing some additional incentives or cash back policies or some other strategies. As a result of this search, customers arrive to the main station at an exponential rate ν^* . The main station is of infinite capacity while both the offer zone and the orbit of passive customers are of finite capacities, say N and M respectively. As a result, when the offer zone is full, Type B customers directly enter the main station with probability γ and leaves the system with probability $(1 - \gamma)$. Customers arriving to the orbit when it is full, is lost from the system for ever. Non persistent customers leave the orbit at an exponential rate ζ .

In the present model both type A and type B customers arrive according to a Marked Markovian Arrival Process (MMAP) with representation (D_0, D_1, D_2) where $D_1 = pD^*$ and $D_2 = (1 - p)D^*$ for $0 \leq p \leq 1$. MMAP may be viewed as a special case of Markovian Arrival Processes or MAPs which is a more general class of point processes which takes in to account the correlation between inter-event times. It includes both Renewal as well as non-Renewal point processes. Many of the processes which we use in modelling of stochastic processes such as Poisson Processes, PH Renewal Processes, Markov Modulated Poisson Processes (MMPP) come under the class of MAP's. The MMAP governing the arrival of type A and type B customers in the present model is described as follows: Let the underlying Markov chain $\{\nu_t, t \geq 0\}$ be irreducible and let D be the generator of this Markov chain with state space $\{1, 2, 3, \dots, m\}$. At the end of a sojourn time in state i , which is exponentially distributed with a positive finite parameter λ^i , one of the following events could occur: with probability $p_{ij}(0)$ it can move to state j where $j \neq i$ without an arrival, with probability $p_{ij}(1)$ it can move to state j with an arrival of a type A customer and with probability $p_{ij}(2)$ it can move to state j with an arrival of a type B customer. Let $D_0 = d_{ij}(0)$ be the rate matrix corresponding to those transitions without an arrival. Let $D_1 = d_{ij}(1)$ be the rate matrix corresponding to the arrival of type A customer and let $D_2 = d_{ij}(2)$ be the rate matrix corresponding to the arrival of type B customer. Then the MMAP under consideration is well be described by the parameter matrices (D_0, D_1, D_2) where $D_1 = pD^*$ and $D_2 = (1 - p)D^*$ for $0 \leq p \leq 1$. $D = D_0 + D_1 + D_2$ is the infinitesimal generator of the Markov chain corresponding to the MMAP. All the off-diagonal elements of D_0 and all the elements of D_1 and D_2 are non negative. To completely specify a $MMAP(D_0, D_1, D_2)$, the initial probability vector in the Markov chain needs to be specified and we assume that the initial probability vector is the same as the stationary probability vector. That is our MMAP is a stationary MMAP. The average total arrival intensity λ is defined by $\lambda = \theta D_1 \mathbf{e}$, where θ is the invariant

vector of the stationary distribution of the Markov chain $\{\nu_t, t \geq 0\}$. The vector θ is the unique solution of the system of equations $\theta D = \mathbf{0}, \theta \mathbf{e} = 1$. where \mathbf{e} denotes a column vector of 1^s and $\mathbf{0}$ is a row vector of 0^s . The average arrival intensity λ_A and λ_B of type A and type B customers respectively are defined by $\lambda_A = \theta D_1 \mathbf{e}$ and $\lambda_B = \theta D_2 \mathbf{e}$. The squared integral (without differentiating the types of customers) coefficient of variation of intervals between successive arrivals is $c_{var} = 2\lambda\theta(-D_0)^{-1}\mathbf{e} - 1$. The squared coefficient of variation of inter-arrival times of type A customers is $c_{var(A)} = 2\lambda_A\theta[-D_0 - D_2]^{-1}\mathbf{e} - 1$ where as that of inter-arrival times of type B customers is $c_{var(B)} = 2\lambda_B\theta[-D_0 - D_1]^{-1}\mathbf{e} - 1$. The integral coefficient of correlation of two successive intervals between arrivals is given as $c_{cor} = [\lambda\theta(-D_0)^{-1}(D - D_0)(-D_0)^{-1}\mathbf{e} - 1]/c_{var}$.

The main station and the offer zone offers the same service but of the offer zone works in a random environment. We assume that there are a finite number of environments whose duration follows Phase Type distribution and the generator matrix of the Markov process leading to the PH distribution depends on the current environment of the offer zone. Let p_i where $\{i = 1, 2, 3, \dots, n\}$ is the probability that the offer zone is at environment i . Each environment of the offer zone consists of one or more offers. Let $\{1, 2, \dots, n\}$ denote the n environments of the offer zone and the duration of time the environment i works follow Phase type distribution with irreducible representation $PH(\beta_i, S_i)$ with M_i phases. The vector S_i^0 is given by $S_i^0 = -S_i \mathbf{e}$. We assume that all the customers in the offer zone are getting served in the same environment and so the offers given to those customers in service at the offer zone change with the change in the environment in which the offer zone works. After service completion at the offer zone type B customers enter the main station with probability η and enter the orbit with probability $(1 - \eta)$ provided it is not fully occupied.

3 Matrix Analytic Solution

We introduce the necessary random variables as follows: Let $N_1(t)$ denote the number of customers in the main station, $N_2(t)$ the number of customers in the offer zone, $N_3(t)$ the number of customers in the orbit, $E(t)$ the environment of the offer zone, $S(t)$ the phase of the environment of the offer zone and $A(t)$ the phase of the arrival process. $E(t)$ can take any of the values $\{1, 2, \dots, n\}$ depending on the ongoing environment of the offer zone. Then $\{N_1(t), N_2(t), N_3(t), E(t), S(t), A(t)\}$ is a Markov process and it describes the process under consideration. This model can be considered as a Level dependent Quasi-Birth-Death (LDQBD) process and a solution is obtained by Matrix Analytic Method. We define the state space of the QBD under consideration and analyze the structure of its infinitesimal generator.

The state space Ω consists of all elements of the form (i, j, k, r, s, t) where

$$i \geq 0; 0 \leq j \leq N; 0 \leq k \leq M; t = 1, 2, \dots, m; r = 1, 2, 3 \dots, n$$

For a fixed value of $r, s = 1, 2, \dots, M_r$.

Let the ordering of the elements of Ω be lexicographical. The infinitesimal generator Q of the LDQBD describing the model under consideration is of the form

$$Q = \begin{pmatrix} A_1^0 & A_0^0 & & & & \\ & A_2^1 & A_1^1 & A_0^1 & & \\ & & A_2^2 & A_1^2 & A_0^2 & \\ & & & A_2^3 & A_1^3 & A_0^3 \\ & & & & \dots & \dots \\ & & & & & \dots \end{pmatrix}$$

where A_0^i, A_1^i, A_2^i are all square matrices whose entries are block matrices of appropriate dimensions.

A_0^i represents the rate matrix corresponding to the arrival of a customer to the main station; that is transition from level $i \rightarrow i + 1$ where $i \geq 0$.

A_2^i represents the rate matrix corresponding to the departure of a customer after service completion at the main station when there are i customers in the main station; that is from level $i \rightarrow i - 1$, for $i = 1, 2, \dots$, and

A_1^i describes all transitions in which the level does not change (transitions within levels i).

In the following sequel \otimes and \oplus represent the Kronecker Sum and Kronecker product respectively. Let \mathbf{e} denote all one vector of appropriate order and I_M denote an identity matrix of order M .

The structure of the A_1^i for $i \geq 0$ are as follows:

$$A_1^i = \begin{pmatrix} E_1 & E_0 & & & & \\ E_2^1 & E_1 & E_0 & & & \\ & E_2^2 & E_1 & E_0 & & \\ & & \dots & \dots & \dots & \\ & & & \dots & \dots & \dots \\ & & & & E_2^{(N-1)} & E_1 & E_0 \\ & & & & & E_2^N & E_1^N \end{pmatrix}$$

- E_0 is the matrix representation of the rate of arrival of type B customers and it depends neither on the number of customers currently undergoing service at the main station nor the number of customers waiting in the orbit of passive customers.
- E_1 is the matrix representation of the rates corresponding to the transitions when there are i customers in the main station and j customers in the offer zone.
- E_2^j is the matrix representation of the rates at which customers leave the offer-zone after completing their service in the offer zone when there are j customers in the offer zone.

E_1 is an $(M + 1) \times (M + 1)$ matrix with sub-blocks given by

$$E_1 = \begin{pmatrix} F_1 & & & & \\ F_2 & F_1 & & & \\ & F_2 & F_1 & & \\ & & \dots & \dots & \\ & & & \dots & \dots \\ & & & & F_2 & F_1 \end{pmatrix}$$

- F_1 is the matrix representation of the transition rates corresponding to the environmental changes, phase changes of the environmental process and the phase changes of the arrival process when there are i customers in the main station and j customers in the offer zone and k customers in the orbit where $j = 1, 2, \dots, N$ and $k = 1, 2, \dots, M$
- F_2 is the matrix representation of the rates at which passive customers leave the orbit without retrying for service at the main station

F_1 is given by

$$F_1 = \begin{pmatrix} C_1 & S_1^0 \otimes p_2 \beta_2 \otimes I_m & S_1^0 \otimes p_3 \beta_3 \otimes I_m & \dots & S_1^0 \otimes p_n \beta_n \otimes I_m \\ S_2^0 \otimes p_1 \beta_1 \otimes I_m & C_2 & S_2^0 \otimes p_3 \beta_3 \otimes I_m & \dots & S_2^0 \otimes p_n \beta_n \otimes I_m \\ S_3^0 \otimes p_1 \beta_1 \otimes I_m & S_3^0 \otimes p_2 \beta_2 \otimes I_m & C_3 & \dots & S_3^0 \otimes p_n \beta_n \otimes I_m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_n^0 \otimes p_1 \beta_1 \otimes I_m & S_n^0 \otimes p_2 \beta_2 \otimes I_m & \dots & \dots & C_n \end{pmatrix}$$

For $j = 1, 2, \dots, N - 1$, if $i \leq (L - 1)$ then

$$C_i = [S_i - (i\mu + j\mu_i + k\nu + \nu^* + \zeta)] \oplus D_0$$

and if $i \geq L$ then,

$$C_i = [S_i - (i\mu + j\mu_i + k\nu + \zeta)] \oplus D_0$$

For $j = N$, if $i \leq (L - 1)$ then

$$C_i = [S_i - (i\mu + j\mu_i + k\nu + \nu^* + \zeta)] \oplus [D_0 + (1 - \gamma)D_2]$$

and for $j = N$, if $i \geq L$ then

$$C_i = [S_i - (i\mu + j\mu_i + k\nu + \zeta)] \oplus [D_0 + (1 - \gamma)D_2]$$

Let $M^* = \sum_{i=1}^n M_i$ and $M^{**} = (M + 1) \sum_{i=1}^n M_i$.

$$F_2 = \zeta I_{mM^*}$$

$$E_0 = I_{M^{**}} \otimes D_2$$

For a fixed value of j , E_2^j is a block-diagonal matrix of order $(M+1) \times (M+1)$ given by

$$E_2^j = \begin{pmatrix} O & G & & & & \\ & O & G & & & \\ & & O & G & & \\ & & & \ddots & \ddots & \\ & & & & O & G \\ & & & & O & G \end{pmatrix}$$

where

$$G = \text{diag}(I_{M_1} \otimes (1 - \eta)j\mu_1 I, I_{M_2} \otimes (1 - \eta)j\mu_2 I, \dots, I_{M_n} \otimes (1 - \eta)j\mu_n I).$$

Here $\text{diag}(a, b, c, \dots)$ represents a diagonal matrix whose diagonal entries are listed and

$$I = I_{m(M+1)(N+1)}$$

The matrix G represents the rate at which the customers enter the orbit of passive customers and the entry is restricted to a maximum number M of the passive customers in the orbit.

The matrix A_0^i corresponding to the arrival of a customer to the main station can be written as

$$A_0^i = \begin{pmatrix} U_1 & & & & & \\ & U_2^1 & U_1 & & & \\ & & U_2^2 & U_1 & & \\ & & & \ddots & \ddots & \\ & & & & U_2^N & U_1^N \end{pmatrix}$$

U_1 represents the transitions from $i \rightarrow i + 1$ without making any changes in the number of customers in the offer zone

$$U_1 = \begin{pmatrix} V_1 & & & & & \\ V_2^1 & V_1 & & & & \\ & V_2^2 & V_1 & & & \\ & & \dots & \dots & & \\ & & & \dots & \dots & \dots \\ & & & & V_2^{(M-1)} & V_1 \\ & & & & & V_2^M & V_1 \end{pmatrix}$$

where for $j = 1, 2, \dots, (N - 1)$

$$V_1 = I_{M^{**}} \otimes D_1$$

and for $j = N$

$$V_1 = I_{M^{**}} \otimes [D_1 + \gamma D_2]$$

For $k = 1, 2, \dots, M$, the matrices V_2^k represents the rate at which customers from the orbit of passive customers enter the main station.

In this case there are two possibilities depending on i , the number of customers in the main station. Whenever the number of customers is greater than or equal to L , the virtual search mechanism is in off condition and only retrials from the orbit increases the number of customers in the main station and whenever this i drops down to $(L - 1)$, the search mechanism starts search for customers from the orbit.

For a fixed i and j , if $i \geq L$ then

$$V_2^k = k\nu I_{mM^*}$$

and if $i \leq (L - 1)$ then

$$V_2^k = [k\nu + \nu^*]I_{mM^*}$$

For $j = 1, 2, \dots, N$, the matrix U_2^j gives the rates at which customers from the offer zone proceeds to the main station without discontinuing their service

$$U_2^j = \text{diag}(I_{M_1} \otimes \eta j \mu_1 I, I_{M_2} \otimes \eta j \mu_2 I, \dots, I_{M_n} \otimes \eta j \mu_n I)$$

where $\text{diag}(a, b, c, \dots, .)$ represents a diagonal matrix whose diagonal entries are listed and

$$I = I_{m(M+1)(N+1)}$$

The matrices A_2^i , representing the rates at which service completion occurs from the main station are given by

$$A_2^i = i\mu I_{(N+1)mM^{**}}$$

3.1 Stability Condition

The present model is a level dependent QBD and we apply Neuts-Rao truncation for the analysis of the model. We assume that when the number of customers in the main station exceeds a certain limit, say K , service occurs at constant rates $K\mu$. In that situation the matrices A_2^i becomes A_2^K for $i \geq K$. We also assume that the truncation level K is greater than the number L at which the search must be switched off. The infinitesimal generator Q^1 of the modified model becomes

$$Q^1 = \begin{pmatrix} A_1^0 & A_0^0 & & & & & \\ A_2^1 & A_1^1 & A_0^1 & & & & \\ & A_2^2 & A_1^2 & A_0^2 & & & \\ & & \dots & \dots & \dots & & \\ & & & A_2 & A_1 & A_0 & \\ & & & & A_2 & A_1 & A_0 \\ & & & & & \dots & \dots \\ & & & & & & \dots & \dots \end{pmatrix}$$

where $A_1 = A_1^K$, $A_2 = A_2^K$ and $A_0 = A_0^K$.

Let the matrix A be defined as $A = A_0 + A_1 + A_2$. We can see that A is an irreducible infinitesimal generator matrix of the underlying process and so there exists the stationary $1 \times (N + 1)(M + 1)mM^*$ vector π of A such that

$$\pi A = 0$$

and

$$\pi e = 1.$$

where $M^* = \sum_{r=1}^n M_r$.

The vector π can be partitioned as

$$\pi = (\pi_0, \pi_1, \pi_2, \dots, \pi_N)$$

For $i = 1, 2, \dots, N$ the vectors π_i can be partitioned as

$$\pi_i = (\pi(i, 1), \pi(i, 2), \dots, \pi(i, M))$$

whereas

$$\pi(i, j) = (\pi(i, j, 1, 1), \pi(i, j, 1, 2), \dots, \pi(i, j, 1, M_1), \dots, \pi(i, j, n, 1), \dots, \pi(i, j, n, M_n))$$

Each vector $\pi(i, j, k, l)$ is a $1 \times m$ vector denoted as

$$\pi(i, j, k, l) = (\pi(i, j, k, l, 1), \pi(i, j, k, l, 2), \dots, \pi(i, j, k, l, m))$$

where the state $\pi(i, j, k, l, m)$ is the probability of being in state (i, j, k, l, m) where i is the number of customers at the offer zone, j the number of passive customers in the orbit, k the environment of the offer zone, l the phase of the environment and r the phase of the underlying MMAP arrival process.

Let the matrix A be of the form

$$A = \begin{pmatrix} W_1^0 & W_0 & & & & & \\ W_2^1 & W_1^1 & W_0 & & & & \\ & W_2^2 & W_1^2 & W_0 & & & \\ & & \dots & \dots & \dots & & \\ & & & \dots & \dots & \dots & \\ & & & & W_2^{(N-1)} & W_1^{(N-1)} & W_0 \\ & & & & & W_2^N & W_1^N \end{pmatrix}$$

where

$$W_0 = E_0$$

for $j = 1, 2, \dots, (N - 1)$

$$W_1^j = E_1 + U_1 + K\mu I_{mM^*}$$

$$W_2^j = E_2^j + U_2^j$$

$$W_1^N = E_1^N + U_1^N + K\mu I_{mM^{**}}$$

The Markov chain with generator Q^1 is positive recurrent if and only if

$$\pi A_0 \mathbf{e} < \pi A_2 \mathbf{e}$$

3.2 Steady State Distribution

The stationary distribution of the Markov process under consideration is obtained by solving the set of equations

$$\mathbf{x}Q^1 = 0, \mathbf{x}\mathbf{e} = 1.$$

Let \mathbf{x} be the steady-state probability vector of Q^1 .

Partition this vector in conformity with Q^1 as follows:

$$\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \dots,)$$

where

$$\mathbf{x}_i = (\mathbf{x}_{i0}, \mathbf{x}_{i1}, \dots, \mathbf{x}_{iN}), i \geq 0$$

For $j = 0, 1, \dots, N$ and $k = 1, 2, \dots, M$ the vectors

$$\mathbf{x}_{ij} = (\mathbf{x}_{ij1}, \mathbf{x}_{ij2}, \mathbf{x}_{ij3}, \dots, \mathbf{x}_{ijM})$$

$$\mathbf{x}_{ijk} = (\mathbf{x}_{ijk1}, \mathbf{x}_{ijk2}, \dots, \dots, \mathbf{x}_{ijkn})$$

For $r = 0, 1, \dots, n$

$$\mathbf{x}_{ijkr} = (\mathbf{x}_{ijkr1}, \mathbf{x}_{ijkr2}, \dots, \mathbf{x}_{ijkrM_r})$$

$$\mathbf{x}_{ijkrst} = (\mathbf{x}_{ijkrst1}, \mathbf{x}_{ijkrst2}, \dots, \mathbf{x}_{ijkrstm})$$

\mathbf{x}_{ijkrst} is the probability of being in state (i, j, k, r, s, t) where

$$i \geq 0; j = 0, 1, \dots, N; k = 0, 1, 2, \dots, M;$$

$$r = 1, 2, \dots, n; s = 1, 2, \dots, M_r; t = 1, 2, \dots, m.$$

Under the stability condition the steady-state probability vector is obtained as

$$\mathbf{x}_{(K-1)+i} = \mathbf{x}_{(K-1)}R^i, i \geq 0$$

where R is the minimal non negative solution to the matrix quadratic equation

$$R^2 A_2 + R A_1 + A_0 = 0$$

and the vectors $\mathbf{x}_0, \dots, \mathbf{x}_{(K-1)}$ are obtained by solving

$$\mathbf{x}_0 A_1^0 + \mathbf{x}_1 A_2^1 = 0$$

$$\mathbf{x}_{(i-1)} A_0^{(i-1)} + \mathbf{x}_i A_1^i + \mathbf{x}_{(i+1)} A_2^{(i+1)} = 0; 1 \leq i \leq (K-2)$$

$$\mathbf{x}_{(K-2)} A_0 + \mathbf{x}_{(K-1)} [A_1^{(K-1)} + A_2 R] = 0$$

subject to the normalizing condition

$$\sum_{i=0}^{(K-2)} x_i + x_{(K-1)}(I - R)^{-1} \mathbf{e} = 1.$$

4 Some Performance Measures of the System

Some measures of performance, which helps the operators of the system to make decisions concerning the optimal values of maximum capacities N and M respectively of the offer zone and the orbit of passive customers and of the cut-off point L are evaluated. Loss of type B customers can happen mainly in two ways: The first type of loss namely, type I loss is due to the lack of space in the offer zone and this happens even before getting a service at the offer zone. The other type of loss namely, type II loss happens when the orbit is full. There is one more type of loss from the orbit of passive customers and the effect of this loss on the system can be minimized by means of orbital search if the number of customers in the main station is less than L . We can also identify the environment of the offer zone from which the maximum expected number of type B customers join the main station which in turn help us to redefine the offers. Following are some performance measures which helps us to make a detailed study about the problem under consideration.

1. Expected Number of customers in the main station

$$E[MS] = \sum_{i=0}^{\infty} ix_i \mathbf{e}$$

where \mathbf{e} is a column vector of appropriate order consisting of all ones.

2. Expected Number of customers in the offer zone

$$E[OZ] = \sum_{i=0}^{\infty} \sum_{j=0}^N \sum_{k=0}^M \sum_{r=1}^n \sum_{s=1}^{M_r} \sum_{t=1}^m j x_{ijkrst}$$

3. Expected Number of customers in the offer zone

$$E[OPC] = \sum_{i=0}^{\infty} \sum_{j=0}^N \sum_{k=0}^M \sum_{r=1}^n \sum_{s=1}^{M_r} \sum_{t=1}^m k x_{ijkrst}$$

4. Expected number of customers enter the main station as a result of search

$$E[S] = \sum_{i=0}^{(L-1)} \sum_{j=0}^N \sum_{k=0}^M \sum_{r=1}^n \sum_{s=1}^{M_r} \sum_{t=1}^m \nu^* x_{ijkrst}$$

5. Probability that a type B customer is lost from the system when the offer zone is full

$$P[L_{T_1}] = \sum_{i=0}^{\infty} \sum_{k=0}^M \sum_{r=1}^n \sum_{s=1}^{M_r} \sum_{t=1}^m (1 - \gamma) x_{iNkrst}$$

6. Probability that a type B customer is lost after service completion at the offer zone

$$P[L_{T_2}] = \sum_{i=0}^{\infty} \sum_{j=0}^N \sum_{r=1}^n \sum_{s=1}^{M_r} \sum_{t=1}^m x_{ijMrst}$$

7. Expected number of non-persistent customers lost from the orbit without joining the main station

$$E[L_{OPC}] = \sum_{i=0}^{\infty} \sum_{j=0}^N \sum_{k=0}^M \sum_{r=1}^n \sum_{s=1}^{M_r} \sum_{t=1}^m k \zeta x_{ijkrst}$$

8. Expected number of type B customers who enter the main station after service completion from environment r of the offer zone

$$E[E(r)] = \sum_{i=0}^{\infty} \sum_{j=0}^N \sum_{k=0}^M \sum_{s=1}^{M_r} \sum_{t=1}^m j \mu_r \eta x_{ijkrst}$$

for $r = 1, 2, \dots, n$

9. Expected number of type B customers lost when the offer zone is full

$$E[L_{T_1}] = \lambda_B \times P[L_{T_1}]$$

where λ_B is the fundamental rate of arrival of customers to the offer zone

10. Expected number of type B customers lost when the orbit is full

$$E[L_{T_2}] = \lambda_B \times P[L_{T_2}]$$

11. Expected number of type B customers who enter the main station after the service completion at the offer zone

$$E[OZ \rightarrow MS] = \sum_{r=0}^n E[E(r)]$$

12. Expected number of type B customer lost due to the capacity restrictions of the offer zone and the orbit

$$E[L] = E[L_{T_1}] + E[L_{T_2}]$$

13. Fraction of time the offer zone is in the r^{th} environment

$$F[r] = \sum_{i=0}^{\infty} \sum_{j=0}^N \sum_{k=0}^M \sum_{s=1}^{M_r} \sum_{t=1}^m x_{ijkrst}$$

where $r = 1, 2, \dots, n$

5 An Optimization Problem

For the economic interpretation of any queueing model, cost analysis plays an important role. In this section, we propose an optimization problem which determines the level L of the main station at which the search mechanism is to be switched off. In this case we assume that all other parameters are kept fixed.

To construct an objective function we assume that the customers undergoing service in the main station provide more revenue to the system when compared to the customers undergoing service in the offer zone. An additional revenue is provided by each customer who enter the main station. Operating cost associated with the functioning of the various environments or offers in the offer zone and holding cost associated with the working of the orbit are expenditures to the system. There is a search cost associated with each customer entering the main station by means of orbital search. The search cost is also an expenditure encountered by the system. Thus we introduce the revenue and expenditure per customer as follows:

- revenue r_1 monetary units per customer undergoing service in the main station
- revenue r_2 monetary units per customer undergoing service in the offer zone where $r_2 < r_1$
- operating cost c_1 monetary units per customer for providing various offers
- holding cost c_2 monetary units per customer waiting in the orbit
- search cost c_3 monetary units per customer entering the main station as a result of orbital search

The Expected Total Profit (**ETP**) is given by

$$(\mathbf{ETP}) = r_1E[MS] + r_2E[OZ] - c_1E[OZ] - c_2E[OPC] - c_3E[S]$$

So the objective of the service providers or the operators of the system is to determine an optimal value of ‘ L ’ for which the total expected cost (**ETP**) is maximum.

6 Conclusion

The results in this paper may be extended to tandem queueing networks consisting of more than two service stations and also to the case where the service time distributions are of so general say Phase Type distributions. Even though such a generalization essentially increases the dimensions of the state space of the Markov chain under consideration which in turn makes the computational implementations more complex and time consuming, we hope that reducing the number of environments and also the dimension of the MMAP under consideration will make it more tractable. We plan to investigate such a problem in future.

Acknowledgement. A. Krishnamoorthy and V.C. Joshua thanks the Department of Science and Technology, Government of India, for the support given under the Indo-Russian Project *INT/RUS/RSF/P-15*. A. Krishnamoorthy also thanks the UGC India for the Award of Emeritus Fellowship *No.F6-6/2017/-18/EMERITUS-2017-18-GEN-10822/(SA-II)*. Ambily P. Mathew thanks the UGC-India for the teacher fellowship sanctioned under the Faculty Development Programme [*F.No.FIP/12thplan/KLMG002TF06*].

References

1. Artalejo, J.R.: Accessible bibliography on retrial queues: progress in 2000–2009. *Math. Comput. Model.* **51**, 1071–1081 (2010)
2. Artalejo, J.R.: A classified bibliography of research on retrial queues: progress in 1990–1999. *Top* **7**, 187–211 (1999)
3. Artalejo, J.R., Joshua, V.C., Krishnamoorthy, A.: An M/G/1 retrial queue with orbital search by server. In: *Advances in Stochastic Modelling*, pp. 41–54. Notable Publications, New Jersey (2002)
4. Bright, L., Taylor, P.G.: Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. *Stoch. Mod.* **11**, 497–525 (1995)
5. Chakravarthy, S.R., Krishnamoorthy, A., Joshua, V.C.: Analysis of a multi-server retrial queue with search of customers from the orbit. *Perform. Eval.* **63**, 776–798 (2006)
6. Dudin, A.N., Krishnamoorthy, A., Joshua, V.C., Tsarenkov, G.: Analysis of BMAP/G/1 retrial system with search of customers from the orbit. *Eur. J. Oper. Res.* **157**, 169–179 (2004)
7. Dudin, A., Deepak, T.G., Joshua, V.C., Krishnamoorthy, A., Vishnevsky, V.: On a BMAP/G/1 retrial system with two types of search of customers from the orbit. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) *ITMM 2017*. CCIS, vol. 800, pp. 1–12. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_1
8. Dallery, Y., Frein, Y.: On decomposition methods for tandem queueing networks with blocking. *Oper. Res.* **41**, 386–399 (1993)
9. Falin, G.I.: A survey of retrial queues. *Queueing syst.* **7**, 127–167 (1990)
10. Falin, G.I., Templeton, J.G.C.: *Retrial Queues*. Chapman and Hall, London (1997)
11. Corral, A.G.: A tandem queue with blocking and Markovian arrival process. *Queueing syst.* **41**, 343–370 (2002)
12. Klimenok, V., Dudin, A., Vishnevsky, V.: On the stationary distribution of tandem queue consisting of a finite number of stations. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) *CN 2012*. CCIS, vol. 291, pp. 383–392. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31217-5_40
13. Klimenok, V., Dudin, A., Vishnevsky, V.: Tandem queueing system with correlated input and cross-traffic. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) *CN 2013*. CCIS, vol. 370, pp. 416–425. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38865-1_42
14. Krishnamoorthy, A., Deepak, T.G., Joshua, V.C.: An M/G/1 retrial queue with non persistent customers and orbital search. *Stoch. Anal. Appl.* **23**, 975–997 (2005)
15. Krishnamoorthy, A., Joshua, V.C., Mathew, A.P.: A retrial queueing system with abandonment and search for priority customers. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) *DCCN 2017*. CCIS, vol. 700, pp. 98–107. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66836-9_9
16. Krishnamoorthy, A., Joshua, V.C., Ambily, P.M.: MMAP/M/∞ queueing system with an offer zone working in a random environment. Accepted for presentation in Euro Conference on Queue. Thesis Israel (2018)
17. Latouche, G., Neuts, M.F.: Efficient algorithmic solutions to exponential tandem queues with blocking. *SIAM J. Algebraic Discrete Methods* **1**, 93–106 (1980)
18. Latouche, G., Ramaswami, V.: *Introduction to Matrix Analytic Methods in Stochastic Modeling*, vol. 5. SIAM (1999)
19. Neuts, M.F., Ramalhoto, M.F.: A service model in which the server is required to search for customers. *J. Appl. Prob.* **21**, 57–166 (1984)

20. Neuts, M.F.: *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Courier Corporation, New York (1981)
21. Neuts, M.F., Rao, B.M.: Numerical investigation of a multiserver retrial model. *Queueing Syst.* **7**, 169–189 (1990)
22. Perros, H.G.: A bibliography of papers on queueing networks with finite capacity queues. *Perform. Eval.* **10**, 255–260 (1989)



Perishable Queuing Inventory Systems with Delayed Feedback

Agassi Melikov¹(✉), Achyutha Krishnamoorthy², and Mammad Shahmaliyev³

¹ Institute of Control Systems, National Academy of Sciences,
B. Vahabzade street 9, Baku, Azerbaijan
agassi.melikov@gmail.com

² Department of Mathematics, Cochin University of Science and Technology,
Cochin 682022, India
achyuthacusat@gmail.com

³ Department of Computer Sciences, National Aviation Academy of Azerbaijan,
Mardakan pr. 30, Baku, Azerbaijan
mamed.shahmaliyev@gmail.com

Abstract. The three-dimensional Perishable Queuing-Inventory System (PQIS) models with positive service time and delayed feedback are studied in this paper. We assume that the customers either leave the system with/without purchasing an item or join the orbit for the decision making. We apply the (s, S) replenishment policy with the positive order lead time. The approximate formulas are developed to calculate the joint distributions and performance measures of the system. The high accuracy of the approximate formulas is confirmed by the numerical experiments.

Keywords: Perishable Queuing-Inventory Systems

Positive service time · Delayed feedback

(s, S) order replenishment policy

Finite and infinite 3D Markov Chains · Calculation methods

1 Introduction

The different models of Queuing-Inventory Systems (QIS) were widely investigated. The detailed review of Perishable and Non-Perishable QIS models could be found in [1, 2].

The classical QIS models are based on the several fundamental assumptions. The first and important one is that after the customer service completion inventory level decreases. But in reality this condition does not always hold, because some customers may refuse to purchase the item after being served due to different reasons. The model with such type of customers was first studied in [3, 4]. Later the similar models were studied in [5, 6] as well.

The second assumption in the studies of QIS models is the absence of feedback. In other words, the served customers are not considered for the repeated service call. But in real systems the served customers may return to the system

instantly (Instantaneous Feedback, IFB) or after some random period of time (Delayed Feedback, DFB) because of qualitative service. The first models of the systems with unlimited inventory (i.e. classical queuing systems) and feedback were studied in Takacs' papers [7,8]. The detailed review of the classical queuing models with feedback could be found in [9,10]. At the same time, analysis of the existing literature at the moment of writing this article showed that the QIS with Feedback (Queuing-Inventory Systems with Feedback, QISwFB) models had been hardly studied. We could only find the papers [11,12]. Let's consider these papers more detailed.

Single-channel QISwDFB with non-perishable inventory and finite queue of the primary customers (p -customers) with Poisson arrival was investigated in [11]. If the queue is full at the moment of arrival p -customers then it leaves the system without being served. The p -customers after service completion according to Bernoulli trial either joins the orbit for future service or leaves the system. The orbit has finite length and every customer after some exponentially distributed random time independently recalls for the service. The system serves the repeated customers (r -customers) if there are no p -customers and/or inventory level is zero. The r -customers requires only service, that is after the service completion of r -customers the inventory level remains unchanged. Service times of both types of customers are exponentially distributed but with different parameters. The non-preemptive service policy is assumed, so that if the r -customer is being served at the moment of arrival of p -customer, the ongoing service is not interrupted. Repetitive orbit re-joining is also allowed, that is after the service completion of r -customer according to Bernoulli trial it either re-joins the orbit or leaves the system. The (s, S) inventory replenishment policy with positive exponential lead time is applied. The three dimensional Markov Chain (3D MC) is used to describe the mathematical model of the system. The algorithm based on matrix methods [13] was introduced for the calculation of steady-state distributions. Additionally, the formulas for the calculation of the average characteristics as well as for the total cost were developed. Laplace-Stieltjes transform of waiting time for both types of customers was derived.

The QISwIFB model with perishable inventory (PQISwIFB) was studied in [12]. This paper investigates a single-channel PQISwIFB with a finite queue of p -customers that forms the MAP flow. The inventory item lifetime is finite and exponentially distributed random variable (r.v.). After the service completion, the p -customers according to Bernoulli trial either instantly joins the second queue of infinite length for the repeated service or leaves the system. At the same time, after service completion of p -customer, the system according to Bernoulli trial either takes the next p -customer or the r -customer from the second queue for the service. The r -customer after being served either instantly re-joins the second queue or leaves the system. After finishing the service of the r -customer, the system accepts the p -customers only if the inventory level is positive. Otherwise, if there are no p -customers and inventory level is zero the channel becomes idle for an exponentially distributed period of time. If during the idle period the p -customer arrives and inventory level becomes positive the channel starts to

serve the customer. Otherwise, if after the idle period no p -customer arrives and the inventory is still empty, the channel begins to serve the r -customers. Likewise in [11], the r -customers requires only the service and the inventory level remains unchanged after service completion. The system uses hybrid replenishment policy, so that if inventory level becomes equal to s then the order of size $S - s$ is placed. Also, if after the service completion of r -customer, the inventory level is equal to $i, i \leq s$ the order of size $S - i$ is placed. The lead time of order is assumed to be the phase-type distributed r.v. The system is modeled by 6D-MC and the algorithm based on matrix methods is developed to calculate the steady-state probabilities. Additionally, the formulas for the performance measures were derived and the total cost minimization problem was considered as well. It should be noted that the developed algorithm is very complex for the practical implementation and becomes less effective for the models of larger dimension.

In our paper we present new single-channel PQISwDFB model. It is similar to the model studied in [11] but with the following differences:

- We study the model with perishable inventory.
- There are three options after the service completion for the customer:
 1. Customer leaves the system without purchasing an inventory item.
 2. Customer purchases the item and leaves the system.
 3. Customer does not purchase the item and joins the orbit for “decision making”.
- r -customer may purchase inventory item as well.
- Both finite and infinite queues of p -customers are considered.
- Waiting customers become impatient when there are no items in the inventory.

These differences improve the model’s likeness to the real systems. Moreover, we present effective method for the calculation of steady-state probabilities. Also we derive the formulas for the performance measures that contains tabulated functions.

The paper is organized as follows. First, we provide the general model description and introduce the problem statement. In the next section, we develop the mathematical model of the system using 3D MC, construct the corresponding Transition matrix (Q-matrix) and derive the exact formulas for the system performance measures. Afterwards, we analyze the finite and infinite models with respect to queue length and orbit size. Finally, we provide the numerical results and give the conclusion.

2 Model Description and Problem Statement

First, let’s consider the detailed description of the model. The system continuously monitors the inventory items, so that every item becomes unusable (perishes) after some finite exponentially distributed random time. Also, we assume that the item already reserved for the servicing cannot perish.

The p -customers arrive into the system according to Poisson scheme. For the simplicity, all the inventory items are considered identical and after the service completion the inventory level decreases by a single unit if the customer purchases the item.

If at the moment of the customer arrival there are items in the inventory and the channel is idle, then the customer is taken to the service by the system. Otherwise, the arrived customer joins the queue. If the inventory level is zero at the arrival moment, the customer either joins the queue according to Bernoulli trial or leaves the system. The customers in the queue become impatient when the inventory level is zero and they independently leave the system after waiting some exponentially distributed period of time.

We consider models both with finite and infinite queue sizes. In the finite case, if at the moment of the arrival the queue is full then the customer is lost. While in the infinite case, all p -customers join the system.

There are three options after the service completion of the p -customer:

1. Customer leaves the system without purchasing an inventory item.
2. Customer purchases the item and leaves the system.
3. Customer does not purchase the item and joins the orbit for “decision making”

We assume that the customers in orbit do not have any information about queue state or inventory level. After some random time every r -customer in orbit applies for service independently, while the system does not differentiate between p -customers and r -customers. Impatience rates and service times for both types of customers are the same. Every served r -customer may re-join the orbit as well, that is the repetitive orbit joins are possible.

The r -customers in orbit are assumed to be insistent. If the queue is full or the inventory level is zero at the moment of arrival, the r -customer returns to the orbit.

The service time depends on whether the customer purchases the item or not, but it has an exponential distribution with different parameters for each case. This assumption corresponds to the real cases, because the service time needed for the customer that purchases the item is greater than for the one who does not.

For the simplicity, we use the 2-level inventory replenishment policy in our model where the lead time of the order is an exponentially distributed r.v. with finite mean.

The problem is to find the steady-state distribution of the system, determine the average queue size for both types of the customers and the average size of the orbit. Later we derive the formulas for the performance measure and perform the cost analysis of the system.

3 Calculation Methods

Let's define the following parameters of the system:

- S - the maximum inventory size
- s - the order threshold, $s < S/2$

- N - the maximum queue length for the model with limited queue size
- R - the maximum orbit size for the model with finite orbit size
- γ^{-1} - the average inventory item lifetime
- λ - the arrival rate of p -customers
- τ^{-1} - average waiting time in the queue when the inventory level is zero
- ϕ_1 - queue joining probability when the inventory level is zero
- ϕ_2 - leaving probability when the inventory level is zero, $\phi_2 = 1 - \phi_1$
- σ_1 - the probability of leaving the system without purchasing an item after the service completion
- σ_2 - the probability of purchasing an item and leaving the system after the service completion
- σ_3 - the probability of joining the orbit for “decision making” without purchasing an item after the service completion
- μ_1^{-1} - average service time of the customer not purchasing the item
- μ_2^{-1} - average service time of the customer that purchase the item
- ν^{-1} - the average lead time of the order
- η^{-1} - the average dwelling time in the orbit

Remark 1. Later the term customer will refer to both types of customers (r and p customers), unless indicated explicitly.

The model is described by 3D MC with the states (m, n, k) , where m is inventory level, n is queue size and k is the orbit size. The state space (SS) of the model is defined as follows:

$$E = \bigcup_{k=0}^R E_k, E_k \cap E_{k'} = \emptyset, k \neq k'. \quad (1)$$

where $E_k = \{(m, n, k) : m = 0, 1, \dots, S, n = 0, 1, \dots, N\}$, $k = 0, 1, 2, \dots, R$.

We conclude from (1) that SS is a set of points with integer coordinates inside the parallelepiped with the height $R + 1$ and rectangle base with the sides of length $S + 1$ and $N + 1$.

The transitions between the states inside the class E_k occur after the following events:

- arrival of p -customer
- inventory replenishment
- service completion
- inventory perishing
- leaving the system due to impatience

On the other hand, the transitions between the classes E_k and $E_{k'}$ are associated with the following events:

- joining the orbit
- r -customer arrival from the orbit

Let's denote the transition rate from the state $(m_1, n_1, k_1) \in E_{k_1}$ to the state $(m_2, n_2, k_2) \in E_{k_2}$ with $q((m_1, n_1, k_1), (m_2, n_2, k_2))$. The set of all these rates forms the generator matrix (Q-matrix) of the 3D MC.

According to the accepted service scheme and inventory replenishment policy of the model, we get the following formulas for the transition rates inside the class E_k (see Algorithm 1):

$$q((m_1, n_1, k), (m_2, n_2, k)) = \begin{cases} \lambda, & \text{if } m_2 = m_1, n_2 = n_1 + 1 \\ \mu_1 \sigma_1, & \text{if } m_2 = m_1, n_2 = n_1 - 1 \\ \mu_2 \sigma_2, & \text{if } m_2 = m_1 - 1, n_2 = n_1 - 1 \\ m_1 \gamma, & \text{if } m_2 = m_1 - 1, n_2 = n_1 = 0 \\ (m_1 - 1) \gamma, & \text{if } m_2 = m_1 - 1, n_2 = n_1 > 0 \\ \nu, & \text{if } m_2 = m_1 + S - s, n_2 = n_1 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

when $m_1 > 0$,

$$q((0, n_1, k), (m_2, n_2, k)) = \begin{cases} \lambda \phi_1, & \text{if } m_2 = 0, n_2 = n_1 + 1 \\ n_1 \tau, & \text{if } m_2 = 0, n_2 = n_1 - 1 \\ S \nu, & \text{if } m_2 = S - s, n_2 = n_1 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Algorithm 1. The calculation of Q-matrix element

```

1: function QELEM( $m_1, n_1, k_1, m_2, n_2, k_2$ )           ▷  $q((m_1, n_1, k_1), (m_2, n_2, k_2))$ 
2:   define  $q := 0$ 
3:   if  $k_2 = k_1$  and  $m_1 > 0$  then
4:     if  $m_2 = m_1$  and  $n_2 = n_1 + 1$  then  $q := \lambda$ 
5:     else if  $m_2 = m_1$  and  $n_2 = n_1 - 1$  then  $q := \mu_1 \sigma_1$ 
6:     else if  $m_2 = m_1 - 1$  and  $n_2 = n_1 - 1$  then  $q := \mu_2 \sigma_2$ 
7:     else if  $m_2 = m_1 - 1$  and  $n_2 = n_1 = 0$  then  $q := m_1 \gamma$ 
8:     else if  $m_2 = m_1 - 1$  and  $n_2 = n_1 > 0$  then  $q := (m_1 - 1) \gamma$ 
9:     else if  $m_1 \leq s$  and  $m_2 = m_1 + S - s$  and  $n_2 = n_1$  then  $q := \nu$ 
10:  else if  $k_2 = k_1$  and  $m_1 = 0$  then
11:    if  $m_2 = 0$  and  $n_2 = n_1 + 1$  then  $q := \lambda \phi_1$ 
12:    else if  $m_2 = 0$  and  $n_2 = n_1 - 1$  then  $q := n_1 \tau$ 
13:    else if  $m_2 = S - s$  and  $n_2 = n_1$  then  $q := \nu$ 
14:  else if  $k_2 \neq k_1$  and  $m_2 = m_1 > 0$  then
15:    if  $n_2 = n_1 - 1$  and  $k_2 = k_1 + 1$  then  $q := \mu_1 \sigma_3$ 
16:    else if  $n_2 = n_1 + 1$  and  $k_2 = k_1 - 1$  then  $q := k_1 \eta$ 
17:  return  $q$ 

```

when $m_1 = 0$. The transition rates between the classes E_{k_1} and E_{k_2} , $k_1 \neq k_2$ is defined as follows ($m > 0$):

$$q((m, n_1, k_1), (m, n_2, k_2)) = \begin{cases} \mu_1 \sigma_3, & n_2 = n_1 - 1, k_2 = k_1 + 1, k_1 < R \\ k_1 \eta, & n_2 = n_1 + 1, k_2 = k_1 - 1, n_1 < R. \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Let's denote the stationary probability of the state $(m, n, k) \in E$ with $p(m, n, k)$. We conclude from the formulas (2), (3) and (4) that the Q-matrix of the model is irreducible, therefore there exists the stationary distribution.

The performance measures of the system is calculated via stationary distributions. We will derive the formulas for the following performance measures: S_{av} - average inventory level, Γ_{av} - average inventory perishing intensity, RR - average reorder rate, L_s - average queue length, L_o - average number of the r -customers in the orbit, RL - average customer loss intensity.

The average inventory level, average queue length and average orbit size are defined as the mathematical expectations of the corresponding random variables:

$$S_{av} = \sum_{(m,n,k) \in E} mp(m, n, k). \quad (5)$$

$$L_s = \sum_{(m,n,k) \in E} np(m, n, k). \quad (6)$$

$$L_o = \sum_{(m,n,k) \in E} kp(m, n, k). \quad (7)$$

The average perishing rate, assuming that the reserved item for the service cannot perish, is calculated as follows:

$$\Gamma_{av} = \gamma \left(\sum_{m=1}^S m \sum_{(m,0,k) \in E} p(m, 0, k) + \sum_{m=2}^S (m-1) \sum_{(m,n,k) \in E} p(m, n, k) I(n > 0) \right). \quad (8)$$

where $I(A)$ is the indicator function of A .

The replenishment order of the inventory is placed independently whenever the inventory level reaches the threshold s :

$$RR = \gamma(s+1) \sum_{(s+1,0,k) \in E} p(s+1, 0, k) + (\mu_2 \sigma_2 + s\gamma) \sum_{(s+1,n,k) \in E} p(s+1, n, k) I(n > 0). \quad (9)$$

The customer loss intensity RL consists of three components:

1. the loss intensity of p -customers (RL_p)
2. the loss intensity because of orbit overflow (RL_o)
3. the loss intensity because of impatience of both types of customers (RL_s)

$$RL_p = \lambda \sum_{(m,N,k) \in E} p(m, N, k) + \lambda \phi_2 \sum_{(0,n,k) \in E} p(0, n, k) I(n < N). \quad (10)$$

$$RL_o = \mu_1 \sigma_3 \sum_{(m,n,R) \in E} np(m, n, R) I(mn > 0). \quad (11)$$

$$RL_s = \tau \sum_{(0,n,k) \in E} np(0, n, k). \quad (12)$$

In order to calculate the above performance measures, we need to obtain the steady-state probability distributions from the balance equations corresponding to the Q-matrix. The balance equations are the system of $(S+1) \times (N+1) \times (R+1)$ linear equations that cannot be solved numerically in a reasonable time for larger or infinite values of the parameters. Therefore, we apply the hierarchical Space Merging Algorithm (SMA) to analyze the performance measures asymptotically.

SMA can be effectively applied for the systems where the transition rates between the states of different classes E_k are very small compared to the transitions inside the class. This assumption holds for the systems where the probability of joining the orbit is far smaller than the total probability of leaving the system: $\sigma_3 \ll \sigma_2 + \sigma_1$.

Assuming the above condition we will consider four models:

1. Both queue length N and orbit size R are finite. We will provide the details of SMA and its application for this model, but provide only the final results for other cases.
2. The queue length N is finite and orbit size R is infinite.
3. Both queue length N and orbit size R are infinite.
4. The queue length N is infinite and orbit size R is finite.

3.1 Analysis of the Model with Finite Queue Length and Orbit Size

In this section we will consider the detailed step by step application of SMA for the finite model, $N < \infty$ and $R < \infty$. In the first step of the hierarchy we construct the merge function $U_1(m, n, k) = \langle k \rangle$ based on (1), where the merged state $\langle k \rangle$ represents the set of all the states inside the class E_k . The set of the all merged states is denoted by $\Omega_1 = \{\langle k \rangle : k = 0, 1, \dots, R\}$. Then we get the following approximate formula for the steady-state distributions:

$$\tilde{p}(m, n, k) \approx \rho^k(m, n) \pi_1(\langle k \rangle). \quad (13)$$

where $\rho^k(m, n)$ is the probability of the state (m, n) inside the class E_k and $\pi_1(\langle k \rangle)$ is the probability of the merged state $\langle k \rangle$, $\langle k \rangle \in \Omega_1$.

Further, based on (13) our problem is reduced to finding the probability distributions of the $R + 1$ number 2D MC-s and a single 1D MC accordingly.

Now we re-apply SMA to the obtained 2D MC-s with state spaces $E_k, k = 0, 1, \dots, R$ in order to find the corresponding $\rho^k(m, n)$ probabilities. All the 2D MC-s are identical, therefore we will consider the model with fixed k :

$$E = \bigcup_{m=0}^S E_k^m, E_k^{m_1} \cap E_k^{m_2} = \emptyset, m_1 \neq m_2. \tag{14}$$

where $E_k^m = \{(m, n, k) \in E_k : n = 0, 1, \dots, N\}, m = 0, \dots, S$. Similarly, we construct the merge function $U_2(m, n, k) = \langle m \rangle$ based on (14), where the merged state $\langle m \rangle$ represents the set of all the states inside the class E_k^m . The set of the all merged states is denoted by $\Omega_2 = \{\langle m \rangle : m = 0, 1, \dots, S\}$. Consequently, according to SMA:

$$p^k(m, n) \approx \rho_m^k(n) \pi_2^k(\langle m \rangle). \tag{15}$$

where $\rho_m^k(n)$ is the probability of the state (m, n) inside the class E_k^m and $\pi_2^k(\langle m \rangle)$ is the probability of the merged state $\langle m \rangle, \langle m \rangle \in \Omega_2$.

Further, let's consider the problem of finding the probabilities $\rho_m^k(n)$ of the split models. We conclude from the formulas (2), (3) and (4) that the transition rates between the states of the split model with state space E_k^m do not depend on the index k , therefore this index is omitted in $\rho_m^k(n)$ and $\pi_2^k(\langle m \rangle)$ onward. According to the formula (2), the probability distributions inside the all split models with the state space $E_k^m, m = 1, \dots, S$ are the same as in the classical model $M/M/1/N$ with load $a = \lambda/\mu_1\sigma_1$:

$$\rho_m(n) = a^n(1 - a)/(1 - a^{N+1}), m = 1, \dots, S. \tag{16}$$

Similarly, we get from the formula (3) that the probability distribution inside the split model with the state space E_k^0 are the same as in the Erlang model $M/M/N/N$ with the load $b = \lambda\phi_1/\tau$:

$$\rho_0(n) = \frac{\theta(b, n)}{\sum_{j=0}^N \theta(b, j)}, n = 0, 1, \dots, N. \tag{17}$$

where $\theta(i, j) = \frac{i^j}{j!}$.

After performing the mathematical transformations over the formulas (2), (3), (16) and (17) we derive the following for the transition rates between the merged states $(\langle m_1 \rangle), (\langle m_2 \rangle) \in \Omega_2$:

$$q(\langle m_1 \rangle, \langle m_2 \rangle) = \begin{cases} \Lambda_1(m_1), & \text{if } m_2 = m_1 - 1 \\ \nu, & \text{if } m_1 \leq s, m_2 = m_1 + S - s. \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

where $\Lambda_1(m_1) = m_1\gamma\rho(0) + (1 - \rho(0))(\mu_2\sigma_2 + (m_1 - 1)\gamma), m_1 = 1, 2, \dots, S$.

Further from (18) we derive (see [6]):

$$\pi_2(\langle m \rangle) = \begin{cases} \alpha_m \pi_2(\langle s+1 \rangle), & \text{if } 0 \leq m \leq s \\ \beta_m \pi_2(\langle s+1 \rangle), & \text{if } s+1 \leq m \leq S-s \\ \chi_m \pi_2(\langle s+1 \rangle), & \text{if } S-s+1 \leq m \leq S \end{cases} \quad (19)$$

where $\alpha_m = \prod_{i=m+1}^{s+1} \frac{A_1(i)}{\nu + A_1(i-1)}$, $\beta_m = \frac{A_1(s+1)}{A_1(m)}$, $\chi_m = \frac{\nu}{A_1(m)} \sum_{i=m-S+s}^S \alpha_i$, $A_1(0) = 0$.

The probability $\pi_2(\langle s+1 \rangle)$ is found from the normalizing condition: $\pi_2(\langle s+1 \rangle) = \left(\sum_{m=0}^s \alpha_m + \sum_{m=s+1}^{S-s} \beta_m + \sum_{m=S-s+1}^S \chi_m \right)^{-1}$

Consequently, after mathematical transformations we derive the following formula for the transition rates between the classes $\langle k_1 \rangle, \langle k_2 \rangle \in \Omega_1$:

$$q(\langle k_1 \rangle, \langle k_2 \rangle) = \begin{cases} A_2, & \text{if } k_2 = k_1 + 1 \\ k_1 M_2, & \text{if } k_2 = k_1 - 1 \\ 0, & \text{otherwise} \end{cases} \quad (20)$$

where $A_2 = \mu_1 \sigma_3 (1 - \rho(0))(1 - \pi_2(\langle 0 \rangle))$, $M_2 = \eta (1 - \rho(N))(1 - \pi_2(\langle 0 \rangle))$.

We conclude from (20) that the probabilities of the merged states $\pi_1(\langle k \rangle), \langle k \rangle \in \Omega_1$ are the same as in the model $M/M/R/R$ with load $c = A_2/M_2$:

$$\pi_1(\langle k \rangle) = \frac{\theta(c, k)}{\sum_{j=0}^R \theta(c, j)}, \quad k = 0, 1, \dots, R. \quad (21)$$

Finally, according to the formulas (13) and (15) the approximate steady-state probabilities of the initial 3D model is calculated as follows:

$$\tilde{p}(m, n, k) \approx \rho_m(n) \pi_2(\langle m \rangle) \pi_1(\langle k \rangle). \quad (22)$$

Further, after substituting (22) in the formulas (5)–(12) we derive the following approximate formulas for the calculation of the system performance measures:

$$S_{av} \approx \sum_{m=1}^S m \pi_2(\langle m \rangle). \quad (23)$$

$$\begin{aligned} L_s &\approx \pi_2(\langle 0 \rangle) \sum_{n=1}^N n \rho_0(n) + (1 - \pi_2(\langle 0 \rangle)) \sum_{n=1}^N n \rho(n) \\ &= b \pi_2(\langle 0 \rangle) (1 - E_B(b, N)) + (1 - \pi_2(\langle 0 \rangle)) \left(\frac{a}{1-a} - \frac{N+1}{1-a^{N+1}} a^{N+1} \right). \end{aligned} \quad (24)$$

$$L_o \approx c(1 - E_B(c, R)). \quad (25)$$

$$\begin{aligned} \Gamma_{av} &\approx \gamma \sum_{m=1}^S \pi_2(\langle m \rangle) (m\rho(0) + (m-1)(1 - \rho(0))) \\ &= \gamma \sum_{m=1}^S \pi_2(\langle m \rangle) \left(m \frac{1-a}{1-a^{N+1}} + (m-1) \frac{a-a^{N+1}}{1-a^{N+1}} \right). \end{aligned} \quad (26)$$

$$\begin{aligned} RR &\approx \pi_2(\langle s+1 \rangle) [(s+1)\gamma\rho(0) + (s\gamma + \mu_2\sigma_2)(1 - \rho(0))] \\ &= \pi_2(\langle s+1 \rangle) [(s+1)\gamma \frac{1-a}{1-a^{N+1}} + (s\gamma + \mu_2\sigma_2) \frac{a-a^{N+1}}{1-a^{N+1}}]. \end{aligned} \quad (27)$$

$$\begin{aligned} RL_p &\approx \lambda[\rho(N)(1 - \pi_2(\langle 0 \rangle)) + \rho_0(N)\pi_2(\langle 0 \rangle) + \phi_2(1 - \rho_0(N))\pi_2(\langle 0 \rangle)] = \\ &= \lambda[a^N \rho(N)(1 - \pi_2(\langle 0 \rangle)) + \pi_2(\langle 0 \rangle)(E_B(b, N) + \phi_2(1 - E_B(b, N)))]. \end{aligned} \quad (28)$$

$$\begin{aligned} RL_o &\approx \mu_1\sigma_3\pi_1(\langle R \rangle)(1 - \rho(0))(1 - \pi_2(0)) \\ &= \mu_1\sigma_3E_B(c, R)(1 - \rho(0))(1 - \pi_2(0)). \end{aligned} \quad (29)$$

$$RL_s \approx \tau\pi_2(\langle 0 \rangle) \sum_{n=1}^N n\rho_0(n) = b\tau\pi_2(\langle 0 \rangle)(1 - E_B(b, N)). \quad (30)$$

Remark 2. $E_B(x, K)$ quantities are the Erlang B-formulas for the calculation of the customer loss probability for the model $M/M/K/K$ with the load x . We provide the formulas for the case $a \neq 1$, because when $a = 1$ the formulas become even simpler: $\rho(n) = 1/(N+1)$, $n = 0, \dots, N$.

Remark 3. We conclude from the formulas (15)–(22) that the stationary distributions depend on all the load parameters of the system. At the same time, according to the formulas (23)–(30) only L_o depends explicitly on the arrival intensity of the r -customers. The reason is that according to our assumption, the probability of joining the orbit is far smaller than the total probability of leaving the system, in other words, the arrival intensity of the r -customers are far smaller than of the p -customers. Additionally, the arrival intensity of the p -customers influences the population of r -customers in the orbit, consequently, all the performance measures depend on the arrival of r -customers implicitly.

The presented methodology could be applied to PQISwDFB with the infinite queue and orbit size as well, $N = \infty$ and/or $R = \infty$. Below we skip intermediary mathematical transformations and present the resulting formulas for the steady state probabilities and the system performance measures for each case.

3.2 Analysis of the Model with Finite Queue Length and Infinite Orbit Size

Let's consider the key points and differences for the case where $N < \infty$ and $R = \infty$:

- $\rho_m(n)$ and $\rho_0(n)$ are calculated by the formulas (16) and (17) accordingly.
- The probabilities of the merged states $\pi_1(\langle k \rangle), \langle k \rangle \in \Omega_1$ are the same as in the model $M(A_2)/M(M_2)/\infty$:

$$\pi_1(\langle k \rangle) \approx \frac{c^k}{k!} e^{-c}, k = 0, 1, \dots \quad (31)$$

- Approximate formulas of the performance measures are calculated by the formulas (23)–(30), except RL_o and L_o . $RL_o = 0$ as the orbit size is infinite and loss probability due to orbit overflow is impossible. The average orbit size is calculated as follows:

$$L_o \approx c. \quad (32)$$

3.3 Analysis of the Model with Infinite Queue Length and Infinite Orbit Size

Let's consider the key points and differences for the case where $N = \infty$ and $R = \infty$:

- The probabilities of all states within the split models with the state space $E_k^m, m = 1, \dots, S$ are the same as in the classical model $M/M/1/\infty$ with the load $a = \lambda/\mu_1\sigma_1$: $\rho_m(n) = (1-a)a^n, m = 1, \dots, S$. We assume that, the ergodicity condition $a < 1$ holds true.
- The probabilities of all states within the split model with the state space E_k^0 are the same as in the Erlang model $M/M/\infty$ with the load $b = \lambda\phi_1/\tau$: $\rho_0(n) = \theta(b, n)e^{-b}, n = 1, 2, \dots$.
- The probabilities of the states of merged models are calculated by the formulas (19) and (31), where $\rho(0) = 1-a$ and $\rho(N) = 0$
- The approximate values of S_{av} and L_o are calculated by the formulas (23) and (32) accordingly. The other performance measure are calculated as follows:

$$L_s \approx b\pi_2(\langle 0 \rangle) + \frac{a}{1-a}(1 - \pi_2(\langle 0 \rangle)). \quad (33)$$

$$\Gamma_{av} \approx \gamma \sum_{m=1}^S \pi_2(\langle m \rangle)(m-a). \quad (34)$$

$$RR \approx \pi_2(\langle s+1 \rangle)((s+1)\gamma(1-a) + (s\gamma + \mu_2\sigma_2)a). \quad (35)$$

$$RL_p \approx \lambda\phi_2\pi_2(\langle 0 \rangle). \quad (36)$$

$$RL_s \approx \tau b\pi_2(\langle 0 \rangle). \quad (37)$$

3.4 Analysis of the Model with Infinite Queue Length and Finite Orbit Size

Finally, let's consider the case where $N = \infty$ and $R < \infty$:

- The state probabilities within the split models with the state space E_k^m , $m = 1, \dots, S$ and E_k^0 are the same as in $N = \infty$ and $R = \infty$ model.
- The probabilities of the states of merged models are calculated by the formulas (19) and (21), where $\rho(0) = 1 - a$ and $\rho(N) = 0$
- The approximate values of S_{av} and L_o are calculated by the formulas (23) and (25) accordingly. The other performance measure are calculated by the formulas (33)–(37), except that RL_o :

$$RL_o \approx \mu_1 \sigma_3 E_B(c, R)(1 - \rho(0))(1 - \pi_2(0)).$$

4 Numerical Results

Finally, let's consider the results of some numerical experiments for the model with the finite queue length and orbit size. Due to the limitations implied on the

Table 1. Estimation of the accuracy of the steady-state probabilities versus various norms

(S, N)	(s, R)	(λ, η)	Norms		
			$\ N\ _1$	$\ N\ _2$	$\ N\ _3$
(10,10)	(1, 2)	(55,5)	0.98964	0.01834	0.00201
	(2,3)	(60,10)	0.98955	0.02042	0.00200
	(4,4)	(65,15)	0.98989	0.01731	0.00194
(10,15)	(1,2)	(55,5)	0.98373	0.01826	0.00154
	(2,3)	(60,10)	0.98456	0.02037	0.00149
	(4,4)	(65,15)	0.98595	0.01726	0.00141
(15,10)	(2,2)	(55,5)	0.95934	0.01823	0.00173
	(5,3)	(60,10)	0.96858	0.02034	0.00154
	(7,4)	(65,15)	0.97482	0.01721	0.00138
(15,15)	(2,2)	(55,5)	0.98686	0.01312	0.00164
	(5,3)	(60, 10)	0.98900	0.01207	0.00148
	(7,4)	(65,15)	0.98996	0.01194	0.00141
(20,5)	(2,2)	(55,5)	0.98019	0.01306	0.00124
	(5,3)	(60,10)	0.98423	0.01203	0.00111
	(9,4)	(65,15)	0.98629	0.01190	0.00103
(20,10)	(2,2)	(55,5)	0.95863	0.01303	0.00129
	(5,3)	(60,10)	0.97068	0.01308	0.00110
	(9,4)	(65,15)	0.97645	0.01297	0.00099

Table 2. Estimation of the accuracy of the performance measures. EV - exact value, AV - approximate value

(S, N)	(s, R, λ, η)	S_{av}		RR		Γ_{av}		L_s		L_o		RL	
		EV	AV	EV	AV	EV	AV	EV	AV	EV	AV	EV	AV
(10,10)	(1,2,55,5)	2.257	2.641	0.536	0.509	3.279	4.010	8.993	9.053	0.524	0.583	43.979	43.716
	(2,3,60,10)	2.329	2.850	0.622	0.577	3.385	4.375	9.133	9.202	0.317	0.301	47.957	46.855
	(4,4,65,15)	2.236	2.823	0.798	0.722	3.196	4.300	9.228	9.301	0.230	0.186	52.891	51.684
(10,15)	(1,2,55,5)	2.257	2.641	0.536	0.509	3.279	4.010	13.185	13.037	0.506	0.583	43.918	41.681
	(2,3,60,10)	2.329	2.850	0.622	0.577	3.385	4.375	13.457	13.391	0.307	0.301	47.934	44.876
	(4,4,65,15)	2.236	2.823	0.798	0.722	3.196	4.300	13.635	13.624	0.224	0.186	52.876	49.696
(15,10)	(2,2,55,5)	3.396	4.165	0.549	0.511	5.434	6.926	9.071	9.146	0.524	0.583	43.872	40.477
	(5,3,60,10)	3.459	4.397	0.727	0.654	5.520	7.335	9.203	9.284	0.316	0.301	47.932	43.564
	(7,4,65,15)	3.265	4.181	0.860	0.766	5.146	6.912	9.281	9.359	0.230	0.186	52.876	48.258
(15,15)	(2,2,55,5)	3.396	4.165	0.549	0.511	5.434	6.926	13.394	13.313	0.510	0.583	42.865	43.239
	(5,3,60,10)	3.459	4.397	0.727	0.654	5.520	7.335	13.644	13.633	0.309	0.301	46.770	46.213
	(7,4,65,15)	3.265	4.181	0.860	0.766	5.146	6.912	13.776	13.791	0.225	0.186	51.871	51.174
(20,5)	(2,2,55,5)	4.387	5.390	0.507	0.473	7.371	9.336	4.450	4.535	0.532	0.582	42.806	41.568
	(5,3,60,10)	4.661	5.937	0.646	0.583	7.859	10.359	4.514	4.590	0.320	0.301	46.743	44.620
	(9,4,65,15)	4.425	5.689	0.838	0.743	7.389	9.856	4.562	4.630	0.232	0.186	51.854	49.503
(20,10)	(2,2,55,5)	4.387	5.390	0.507	0.473	7.366	9.333	9.101	9.177	0.524	0.583	42.766	40.580
	(5,3,60,10)	4.660	5.937	0.646	0.584	7.855	10.357	9.240	9.320	0.316	0.301	46.741	43.566
	(9,4,65,15)	4.425	5.689	0.838	0.743	7.387	9.854	9.320	9.399	0.230	0.186	51.854	48.297

volume of the paper we will only consider the accuracy of the SMA algorithm. We will provide comparison of steady-state probabilities and performance measures. The accuracy will be estimated using the following norms:

- Cosine similarity: $\|N\|_1 = \frac{\sum_{(m,n,k) \in E} p(m,n,k)\tilde{p}(m,n,k)}{\sqrt{\sum_{(m,n,k) \in E} (p(m,n,k))^2} \sqrt{\sum_{(m,n,k) \in E} (\tilde{p}(m,n,k))^2}}$.
- Maximum absolute difference: $\|N\|_2 = \max_{(m,n,k) \in E} |p(m,n,k) - \tilde{p}(m,n,k)|$.
- Root mean square deviation (RMSE): $\|N\|_3 = \left[\frac{1}{|E|} \sum_{(m,n,k) \in E} (p(m,n,k) - \tilde{p}(m,n,k))^2 \right]^{\frac{1}{2}}$, where $|E|$ is the cardinality of the state space E .

The exact values of steady-state probabilities are calculated from the linear system of balance equations corresponding the Q-matrix. The system parameters for numerical experiments are accepted as follows:

$$\mu_1 = 55, \mu_2 = 5, \sigma_1 = 0.3, \sigma_2 = 0.5, \phi_1 = 0.3, \nu = 1, \tau = 1.5.$$

The comparison results of the steady-state probabilities and performance measures are given in Tables 1 and 2 correspondingly. We conclude from these tables that the accuracy of the approximate approach is very accurate.

4.1 Pros and Cons

The main advantage of SMA is that it eliminates the solving of the complex systems of linear equations. Therefore it is very fast and could be easily implemented. Although there are some matrix-geometric and eigen-value based algorithms for the solution of finite and infinite MC, their implementations are more complex, additionally they may become numerically unstable and produce badly conditioned systems of linear equations. Also the most of them impose mandatory conditions on the form of Q-matrix.

The main disadvantage of SMA is that it produces approximate results, while its accuracy is very high as confirmed by the numerical experiments.

5 Conclusion

The finite and infinite 3D PQIS models with positive service time and delayed feedback are studied in this paper. It is assumed that the customers either leave the system with/without purchasing an item or join the orbit for the decision making. When the inventory level is zero, customers join the system according to Bernoulli trial, while customers in the queue become impatient. The inventory replenishment policy belongs to (s, S) class. The exact and approximate formulas are given for the calculation of the steady-state probabilities and performance measures of the system. Exact method is based on the solving of balance equations and is suitable only for the finite models. The approximate approach is based on the State Merging Algorithm of Markov Chains and is applicable for both finite and infinite systems. The high accuracy of the given formulas are proven using numerical experiments and the corresponding comparison tables are provided.

References

1. Krishnamoorthy, A., Lakshmy, B., Manikandan, R.: A survey on inventory models with positive service time. *Opsearch* **48**(2), 158–169 (2011)
2. Karaesmen, I.Z., Scheller-Wolf, A., Deniz, B.: Managing perishable and aging inventories: review and future research directions. In: Kempf, K., Keskinocak, P., Uzsoy, R. (eds.) *Planning Production and Inventories in the Extended Enterprise*. International Series in Operations Research & Management Science, vol. 151. Springer, Boston (2011). https://doi.org/10.1007/978-1-4419-6485-4_15
3. Krishnamoorthy, A., Manikandan, R., Lakshmy, B.: Revisit to queuing-inventory system with positive service time. *Ann. Oper. Res.* **233**, 221–236 (2015)
4. Krishnamoorthy, A., Manikandan, R., Shajin, D.: Analysis of a multi-server queuing-inventory system. *Adv. Oper. Res.* **2015**, 16 (2015). Article ID 747328
5. Melikov, A.Z., Shahmaliyev, M.O.: A perishable queuing-inventory system with positive service time and (S-1, S) replenishment policy. *Communications in Computer and Information Sciences*, vol. 800, pp. 83–96. Springer (2017)
6. Melikov, A.Z., Ponomarenko, L.A., Shahmaliyev, M.O.: Analysis of perishable queuing-inventory systems with different types of requests. *J. Autom. Inf. Sci.* **49**(9), 42–60 (2017)

7. Takacs, L.A.: Single-server queue with feedback. *Bell Syst. Tech. J.* **42**, 505–519 (1963)
8. Takacs, L.A.: Queuing model with feedback. *Oper. Res.* **11**(4), 345–354 (1977)
9. Melikov, A., Ponomarenko, L., Rustamov, A.: Methods for analysis of queueing models with instantaneous and delayed feedbacks. In: Dudin, A., Nazarov, A., Yakupov, R. (eds.) *ITMM 2015. CCIS*, vol. 564, pp. 185–199. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25861-4_16
10. Koroliuk, V.S., Melikov, A.Z., Ponomarenko, L.A., Rustamov, A.M.: Methods for analysis of multi-channel queueing models with instantaneous and delayed feedbacks. *Cybern. Syst. Anal.* **52**(1), 58–70 (2016)
11. Amirthakodi, M., Sivakumar, B.: An inventory system with service facility and finite orbit for feedback customers. *Opsearch.* **52**(2), 225–255 (2015)
12. Amirthakodi, M., Radhamami, V., Sivakumar, B.: A perishable inventory system with service facility and feedback customers. *Annal. Oper. Res.* **233**, 25–55 (2015)
13. Neuts, M.F.: *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. John Hopkins University Press, Baltimore (1981)



Methods of Limiting Decomposition and Markovian Summation in Queueing System with Infinite Number of Servers

Anatoly Nazarov and Diana Dammer^(✉)

Tomsk State University, Tomsk, Russia
nazarov.tsu@gmail.com, dammerdiana11@gmail.com

Abstract. In this paper, we study the process generated by the customers which are serviced in a queueing system with an infinite number of servers. Two methods referred to as the method of limiting decomposition and the method of Markovian summation are proposed, implemented and compared. The characteristic function of the probability distribution for the studied process is obtained. The numerical examples are performed for different values of queueing system characteristics.

Keywords: Queueing system · Characteristic function
Method of Markovian summation · Method of limiting decomposition

1 Introduction

At present, the methods of queueing theory [1–5] are used for modeling various real systems: production systems [6], call centers [7, 8], economic systems [9–12], telephone cellular communication [13], etc. Most of the methods used in solving these problems belong to Markovian models, where it is assumed that arrival process is a stationary Poisson one and the service time is an exponentially distributed. The observations of the real systems have shown that it is necessary to develop methods for the study of the non-Markovian queueing models [14].

Some processes occurring in these systems have been investigated, among them are the process of the number of the busy servers, the arrival process or the process of customers leaving the system. Besides, the process generated by the customers arrived in the system is of interest to the researchers. The process of this kind is formed in case of computers failure. The study has been made of this process in [15]. Also, this process is formed by the clients in the insurance company (the process of occurrence of insurance claims). It has been analyzed in [16] for an exponentially distributed service and interarrival times.

This paper focuses on the implementation of the new method of Markovian summation to analyze the process generated by the customers arrived in the queueing system with independent, identically, and generally distributed service times and stationary Poisson arrival process. Also, the aim of this paper is to

compare the results obtained by using two different methods: the method of limiting decomposition and the method of Markovian summation.

The structure of this paper is organized as follows. Section 2 gives a description of the mathematical model and the statement of the problem. Section 3 specifies the method of limiting decomposition for the problem under study. In Sect. 4, we describe the method of Markovian summation and show how it works. Section 5 presents the comparison of the results obtained by using two methods. In Sect. 6 a numerical examples are presented for the gamma-distributed service times.

2 Mathematical Model

We consider a queueing system (Fig. 1) with an infinite number of servers. It is assumed that the arrival process is a stationary Poisson process, with the rate equal to λ . The service times are independent and identically distributed with an arbitrary distribution function $B(x)$. During the service time every arrived customer generates any events where intensity equals to γ independently from other customers. These events formed by the one customer are referred to as the local d -process (derivative process). The set of the events generated by all customers are referred to as the total d -process (or d -process). The events of d -process do not generate other events.

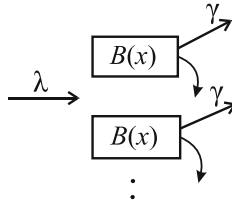


Fig. 1. The model of queueing system with d -process

The problem is to obtain the characteristic function of a number of d -process events occurred during the time T on the interval $[0, T]$ under condition that at the initial time $t = 0$ the system is free.

3 Method of Limiting Decomposition

Let us consider the method of limiting decomposition. According to this method the stationary Poisson arrival process with parameter λ is divided into N independent Poisson processes, with the rate equal to $\frac{\lambda}{N}$ under polynomial scheme. So, we can decompose the original system with an infinite number of servers and consider the set of N (for $N \rightarrow \infty$) single-line systems with losses of the arriving customers, when the server is busy (Fig. 2).

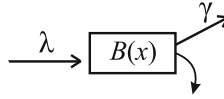


Fig. 2. The model of queueing system with losses

As d -process of the original system with an infinite number of servers is the limiting (for $N \rightarrow \infty$) sum of N independent d -processes of a single-line system with losses, then we first can obtain the probabilistic characteristics of the studied process of this single-line system. After that, we will find the probabilistic characteristics of d -process of the system with an infinite number of servers by summation.

3.1 Research of the Single-line System

Let us denote $n(t)$ number of d -process events occurred in the single-line system during the time t on the interval $[0, t]$. We assume that the system is free at the initial moment of time $t = 0$. Also, we denote $k(t)$ the state of the server:

$$k(t) = \begin{cases} 0, & \text{if the server is free,} \\ 1, & \text{if the server is busy.} \end{cases}$$

As the process $k(t)$ is not Markovian then we have to denote the additional variable $z(t)$ —a length of the interval from the moment time t to the end of the current service. And now the two-dimensional process $k(t), z(t)$ is Markovian.

Denote

$$P_0(n, t) = P\{k(t) = 0, n(t) = n\},$$

$$P_1(n, z, t) = P\{k(t) = 1, z(t) < z, n(t) = n\}.$$

Applying the formula of total probability, we can write the following equations:

$$P_0(n, t + \Delta t) = P_0(n, t) \left(1 - \frac{\lambda}{N} \Delta t \right) + P_1(n, \Delta t, t) + o(\Delta t),$$

$$P_1(n, z - \Delta t, t + \Delta t) = [P_1(n, z, t) - P_1(n, \Delta t, t)](1 - \gamma \Delta t) + P_0(n, t) \frac{\lambda}{N} \Delta t B(z) + P_1(n - 1, z, t) \gamma \Delta t + o(\Delta t).$$

After performing some transformation, we derive the following system of the Kolmogorov differential equations [17] for the probability distribution of the studied process $n(t)$:

$$\frac{\partial P_0(n, t)}{\partial t} = -\frac{\lambda}{N} P_0(n, t) + \frac{\partial P_1(n, 0, t)}{\partial z},$$

$$\frac{\partial P_1(n, z, t)}{\partial t} = \frac{\partial P_1(n, z, t)}{\partial z} - \frac{\partial P_1(n, 0, t)}{\partial z} - \gamma P_1(n, z, t) + B(z) \frac{\lambda}{N} P_0(n, t) + \gamma P_1(n - 1, z, t),$$

where it is denoted

$$\left. \frac{\partial P_1(n, z, t)}{\partial z} \right|_{z=0} = \frac{\partial P_1(n, 0, t)}{\partial z}.$$

To solve this system, we will consider the partial characteristic functions:

$$H_0(u, t) = \sum_{n=0}^{\infty} e^{jun} P_0(n, t),$$

$$H_1(u, z, t) = \sum_{n=0}^{\infty} e^{jun} P_1(n, z, t),$$

where j —imaginary unit. Thus, we can write for these partial characteristic functions the following system of equations:

$$\begin{aligned} \frac{\partial H_0(u, t)}{\partial t} &= -\frac{\lambda}{N} H_0(u, t) + \frac{\partial H_1(u, 0, t)}{\partial z}, \\ \frac{\partial H_1(u, z, t)}{\partial t} &= \frac{\partial H_1(u, z, t)}{\partial z} - \frac{\partial H_1(u, 0, t)}{\partial z} \\ &+ \gamma(e^{ju} - 1)H_1(u, z, t) + B(z)\frac{\lambda}{N}H_0(u, t), \end{aligned} \tag{1}$$

under the initial condition:

$$H_0(u, 0) = 1, \quad H_1(u, z, 0) = 0. \tag{2}$$

The solution $\{H_0(u, t), H_1(u, z, t)\}$ of the Cauchy problem (1) and (2) we will find in the following form:

$$\begin{aligned} H_0(u, t) &= 1 + \frac{1}{N}F_0(u, t) + O\left(\frac{1}{N^2}\right), \quad F_0(u, 0) = 0, \\ H_1(u, z, t) &= \frac{1}{N}F_1(u, z, t) + O\left(\frac{1}{N^2}\right), \quad F_1(u, z, 0) = 0. \end{aligned} \tag{3}$$

Thus, we can write the equations for $F_0(u, t)$, $F_1(u, z, t)$:

$$\begin{aligned} \frac{\partial F_0(u, t)}{\partial t} &= -\lambda + \frac{\partial F_1(u, 0, t)}{\partial z}, \\ \frac{\partial F_1(u, z, t)}{\partial t} &= \frac{\partial F_1(u, z, t)}{\partial z} - \frac{\partial F_1(u, 0, t)}{\partial z} + \gamma(e^{ju} - 1)F_1(u, z, t) + B(z)\lambda. \end{aligned} \tag{4}$$

Let us denote $\frac{\partial F_1(u, 0, t)}{\partial z} = h(u, t)$ then the system (4) we can write in the form:

$$\begin{aligned} \frac{\partial F_0(u, t)}{\partial t} &= h(u, t) - \lambda, \\ \frac{\partial F_1(u, z, t)}{\partial t} - \frac{\partial F_1(u, z, t)}{\partial z} &= \gamma(e^{ju} - 1)F_1(u, z, t) + B(z)\lambda - h(u, t). \end{aligned} \tag{5}$$

Solution to the second differential equation of the system (5) is determined by solving the following system of ordinary differential equations for characteristic curves [18]:

$$\frac{dt}{1} = \frac{dz}{-1} = \frac{dF_1(u, z, t)}{(e^{ju} - 1)\gamma F(u, z, t) + \lambda B(z) - h(u, t)} .$$

We will find a solution of this equation under the initial condition $F(u, z, 0) = 0$. Denoting $a(u) = (e^{ju} - 1)\gamma$, we can write:

$$F_1(u, z, t) = e^{a(u)t} \int_0^t e^{-a(u)x} [\lambda B(t + z - x) - h(u, x)] dx. \tag{6}$$

Thus, for function $h(u, t)$ we can write:

$$\begin{aligned} h(u, t) &= e^{a(u)t} \int_0^t e^{-a(u)x} \lambda B'(t - x) dx \\ &= e^{a(u)t} \int_0^t e^{-a(u)(t-y)} \lambda B'(y) dy = \lambda \int_0^t e^{a(u)y} dB(y), \end{aligned}$$

where we make the following changes to the variable: $t - x = y$. Thus, we have obtained:

$$h(u, t) = \lambda \int_0^t e^{a(u)y} dB(y). \tag{7}$$

Let us denote

$$\lim_{z \rightarrow \infty} F_1(u, z, t) = F_1(u, t).$$

Performing the transition $z \rightarrow \infty$ in the formula (6) and taking into account (7), we will obtain:

$$F_1(u, T) = \frac{\lambda}{a(u)} \left\{ e^{a(u)T} (1 - B(T)) - 1 + \int_0^T e^{a(u)x} dB(x) \right\}, \tag{8}$$

where $a(u) = (e^{ju} - 1)\gamma$.

From the first equation of the system (5) and taking into account (7), we can write:

$$F_0(u, T) = -\lambda T + \lambda \int_0^T (T - x) \exp\{(e^{ju} - 1)\gamma x\} dB(x). \tag{9}$$

Summing up (8) and (9) and denoting

$$F(u, t) = F_0(u, t) + F_1(u, t),$$

we obtain

$$F(u, T) = -\lambda T + \lambda \int_0^T (T - x) \exp[(e^{ju} - 1)\gamma x] dB(x) + \frac{\lambda}{(e^{ju} - 1)\gamma} \times \left\{ (1 - B(T)) \exp[(e^{ju} - 1)\gamma T] - 1 + \int_0^T \exp[(e^{ju} - 1)\gamma x] dB(x) \right\}. \tag{10}$$

Taking into account (3), we can write

$$H(u, T) = H_0(u, T) + H_1(u, T) = 1 + \frac{1}{N} F(u, T),$$

where $H(u, T)$ —characteristic function of the number of d -process events occurred during time T on the interval $[0, T]$ into the single-line system with loses.

3.2 Research of the System with an Infinite Number of Servers

We denote $\tilde{H}(u, T)$ —the characteristic function of the number of the total d -process events occurred during time T on the interval $[0, T]$ into the original system with an infinite number of servers. Thus, we can write for $\tilde{H}(u, T)$ the following expression:

$$\tilde{H}(u, T) = \lim_{N \rightarrow \infty} \left(1 + \frac{1}{N} F(u, T) \right)^N = \exp\{F(u, T)\}, \tag{11}$$

where $F(u, T)$ is defined by the formula (10). Rewrite it back to notation $a(u)(e^{ju} - 1)\gamma$:

$$F(u, T) = -\lambda T + \lambda \int_0^T (T - x) e^{a(u)x} dB(x) + \frac{\lambda}{a(u)} \left\{ (1 - B(T)) e^{a(u)T} - 1 + \int_0^T e^{a(u)x} dB(x) \right\}. \tag{12}$$

The equalities (11) and (12)—are result of a study of the d -process problem into the system with an infinite number of servers, with an arbitrary distribution function of the service times and a stationary Poisson arrival process.

The method of limiting decomposition considered above can only be applied to the systems with stationary Poisson arrival process. However, this method cannot be used to study the systems with other kinds of arrival processes, e.g. Markovian modulated Poisson process or renewal arrival process, because these arrival processes are not stochastic independent.

Thus, to explore the queueing system and to solve our problem we offer other method referred to as the method of Markovian summation.

4 Method of Markovian Summation

Considering d -process into the system with an infinite number of servers on the interval of time $[0, T]$, let us denote $\xi(t)$ —the number of d -process events generated on the interval $[0, T]$ (precisely on the interval $[t, T]$) by the customer arrived at time t . Also, we denote $n(t)$ —the total number of d -process events formed on the interval $[0, T]$ by all the customers arrived in the system during the time t on the interval $[0, t]$.

It is obvious, if $t = T$ then value $n(T)$ is equal to the number of d -process events occurred during the time T on the interval $[0, T]$ under condition that at initial moment $t = 0$ the system is free. Herewith, the value $n(t)$ is not equal to the number of d -process events occurred during the time t on the interval $[0, t]$, because d -process events formed by arrived customers occur throughout the interval $[0, T]$, the time after moment t included. The study of the process $n(t)$ is necessary to find the probability distribution of the value $n(T)$ equal to the sum of the number d -process events occurred on the interval $[0, T]$.

As the random variables $\xi(t)$ for different moments t of the arriving customers are independent and the random process $n(T)$ is Markovian, then the proposed method will be called the method of Markovian summation of the values $\xi(t)$.

4.1 Probability Distribution of the Value $\xi(t)$

We denote for the arriving customer at the moment $t \in [0, T]$ $r(i, t) = P\{\xi(t) = i\}$ and obtain this probability distribution.

The arrived customers can be of two kinds (Fig. 3).

On the time axis the interval boundaries and the moments t_1 and t_2 of two arriving customers are marked. For the customer arrived at the moment t_1 the

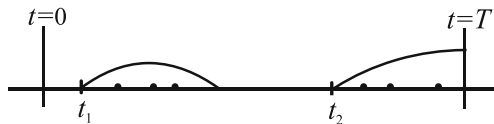


Fig. 3. The model of formation d -process

service time ends before the moment $t = T$. For the customer arrived at the moment t_2 the service time ends after the moment $t = T$. In the first case all d -process events generated by this customer belong to the interval $[0, T]$. For the customer of the second kind the d -process events generated by this customer belong to the interval $[t, T]$.

Let us fix the value x of the service time of the customer of the first kind. Then, the number of the d -process events generated by this customer is Poisson distributed, with rate equal to γx . For the number of the d -process events generated by the customer of the second kind the rate of the Poisson distribution is equal to $\gamma(T - t)$.

Applying the formula of total probability to the distribution $r(i, t)$, we can write:

$$r(i, t) = \int_0^{T-t} \frac{(\gamma x)^i}{i!} e^{-\gamma x} dB(x) + (1 - B(T - t)) \frac{(\gamma(T - t))^i}{i!} e^{-\gamma(T-t)}. \tag{13}$$

Its characteristic function can be represented as follows:

$$g(u, t) = M\{e^{ju\xi(t)}\} = \sum_{i=0}^{\infty} e^{ju i} r(i, t) \\ = \int_0^{T-t} \exp\{(e^{ju} - 1)\gamma x\} dB(x) + (1 - B(T - t)) \exp\{(e^{ju} - 1)\gamma(T - t)\}. \tag{14}$$

4.2 Kolmogorov Equations

For the Markovian process $n(t)$ let us denote $P(n, t) = P\{n(t) = n\}$ and write the equalities:

$$P(n, t + \Delta t) = P(n, t)(1 - \lambda\Delta t) + \lambda\Delta t \sum_{i=0}^n P(n - i, t)r(i, t) + o(\Delta t).$$

After performing some transformation, we derive the following system of the Kolmogorov differential equation for the probability distribution of the process $n(t)$:

$$\frac{\partial P(n, t)}{\partial t} = -\lambda P(n, t) + \lambda \sum_{i=0}^n P(n - i, t)r(i, t). \tag{15}$$

Let us denote $H(u, t)$ the characteristic function of the distribution $P(n, t)$

$$H(u, t) = M\{e^{jun(t)}\} = \sum_{n=0}^{\infty} e^{jun} P(n, t).$$

Taking into account the system (15), we will write the equation

$$\frac{\partial H(u, t)}{\partial t} = \lambda H(u, t)(g(u, t) - 1).$$

Now, we can write the characteristic function for $t = T$ under initial condition $H(u, 0) = 1$:

$$H(u, T) = \exp \left\{ \lambda \int_0^T (g(u, t) - 1) dt \right\}, \tag{16}$$

where $g(u, t)$ is defined by equality (15).

The obtained characteristic function (16) as well as the function (11) define the probability distribution $P(n, T)$ of the number of d -process events occurred during the time T on the interval $[0, T]$ in the system with an infinite number of servers. It is obvious, these functions should be the same. Let us make sure.

5 The Comparison of the Characteristic Functions

The characteristic functions (11) and (16) are the exponential ones depending on the $F(u, T)$ and $\lambda \int_0^T (g(u, t) - 1) dt$. The function $F(u, T)$ is defined by the equality (12) and has the following form:

$$F(u, T) = -\lambda T + \lambda \int_0^T (T - x) e^{a(u)x} dB(x) + \frac{\lambda}{a(u)} \left\{ (1 - B(T)) e^{a(u)T} - 1 + \int_0^T e^{a(u)x} dB(x) \right\}. \tag{17}$$

Denote $G(u, T) = \lambda \int_0^T (g(u, t) - 1) dt$, where $g(u, t)$ is defined by the expression (14). Rewrite its using the designation $(e^{ju} - 1)\gamma = a(u)$:

$$G(u, T) = -\lambda T + \lambda \int_0^T g(u, t) dt = -\lambda T + \int_0^T \left\{ (1 - B(T - t)) e^{a(u)(T-t)} + \int_0^{T-t} e^{a(u)x} dB(x) \right\} dt.$$

Let us make the change to the variable $T - t = y$ in (17). We will obtain

$$G(u, T) = -\lambda T + \lambda \int_0^T (T - x) e^{a(u)x} dB(x) + \frac{\lambda}{a(u)} \left\{ (1 - B(T)) e^{a(u)T} - 1 + \int_0^T e^{a(u)x} dB(x) \right\}. \tag{18}$$

Since the functions $F(u, T)$ and $G(u, T)$ from (17) and (18) are equal to each other then the characteristic functions $\tilde{H}(u, T)$ from (11) and $H(u, T)$ from (16) are match.

6 Numerical Example

In this section, we present the numerical results for the various parameters of the studied model. Let the service time has gamma distribution with a shape parameter equal to α and an inverse scale parameter equal to β . The interarrival time has an exponential distribution with parameter λ , the intensity of occurrence of d -process events equal to γ . Shown on the graphs (Figs. 4, 5 and 6) are distributions $P(n, T)$ of the number of d -process events occurred during time T on the interval $[0, T]$ for the various values of parameters: $\lambda, \gamma, T, \alpha, \beta$:

Table 1 contains the expected value (a) and variance (D) of the number of d -process events occurred during time T on the interval $[0, T]$ for the following values of parameters: $\lambda = 2, T = 5, N = 60$. The values of parameters $\alpha, \gamma, \beta = \alpha$ will be changed in a series of experiments.

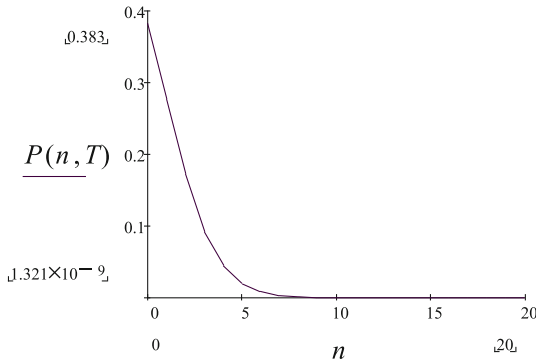


Fig. 4. $\lambda = 3, \gamma = 1, T = 1, \alpha = 2, \beta = \alpha, N = 20$

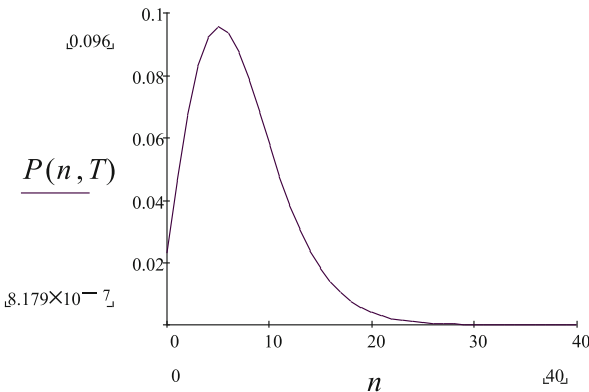


Fig. 5. $\lambda = 2, \gamma = 1, T = 5, \alpha = 0.5, \beta = \alpha, N = 40$

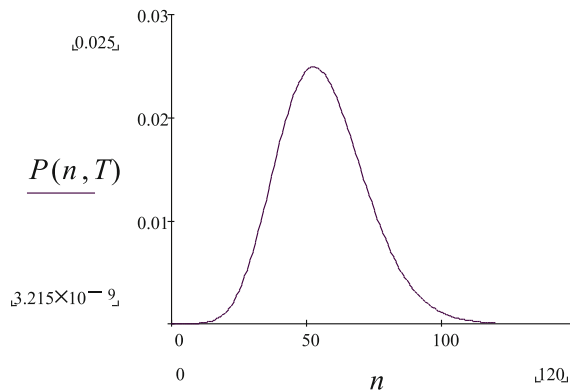


Fig. 6. $\lambda = 2, \gamma = 2, T = 15, \alpha = 1, \beta = \alpha, N = 120$

Table 1. The expected value and variance of the number of d -process events

	$\gamma = 0.2$	$\gamma = 1$	$\gamma = 2$
$\alpha = 0.5$	$a = 1.437, D = 2.040$	$a = 7.159, D = 20.014$	$a = 14.315, D = 65.679$
$\alpha = 1$	$a = 1.603, D = 2.090$	$a = 8.013, D = 20.202$	$a = 16.030, D = 64.855$
$\alpha = 2$	$a = 1.700, D = 2.140$	$a = 8.500, D = 19.504$	$a = 16.998, D = 60.940$

7 Conclusions

Thus, in this paper we have researched the d -process generated by the arrived customers using two methods. The method of limiting decomposition is used to study the system with a stationary Poisson arrival process. However, the proposed method of Markovian summation can be successfully generalized for the study d -process in the system with the more common arrival processes, e.g. Markovian modulated Poisson process or renewal arrival process. The results of the study of the processes as d -process can be used to analyze the activity of some economic or production systems. The future investigations will be focused on the study d -process in the systems with Markovian modulated Poisson process and arbitrary service time distribution function.

References

1. Bocharov, P.P., D'Apice, C., Pechinkin, A.V., Salerno, S.: Queueing Theory. VSP, Utrecht, Boston (2004)
2. Narayan Bath, U.: An Introduction to Queueing Theory: Modeling and Analysis in Applications. Birkauer, Boston (2008)
3. Asmussen, S.: Applied Probability and Queues. Stochastic Modelling and Applied Probability. Springer, New-York (2003). <https://doi.org/10.1007/b97236>

4. Shortle, J.F., Thompson, J.M., Gross, D., Harris, C.M.: *Fundamentals of Queueing Theory*. Wiley, Hoboken, USA (2018)
5. Kleinrock, L.: *Queueing Systems*, vol. 1. Wiley Interscience, New York (1975)
6. Balsamo, S., De Nitti Persone, V., Inverardi, P.: A review on queueing network models with finite capacity queues for software architectures performance prediction. *Perform. Eval.* **51**(2), 269–288 (2003)
7. Borst, S., Mandelbaum, A., Reiman, M.I.: Dimensioning large call centers. *Oper. Res.* **52**, 17–34 (2004)
8. Brian, H.F., Adan, I.J.B.F.: An infinite-server queue influenced by a semi-Markovian environment. *Queueing Syst.* **61**, 65–84 (2009)
9. Dammer, D.: Research of mathematical model of insurance company in the form of queueing system with unlimited number of servers considering “Implicit Advertising”. In: Dudin, A., Nazarov, A., Yakupov, R. (eds.) *ITMM 2016. Communications in Computer and Information Science*, vol. 564. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25861-4_14
10. Dammer, D.: Research of mathematical model of insurance company in the form of queueing system in a random environment. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) *ITMM 2016. Communications in Computer and Information Science*, vol. 800. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_17
11. Jidkova, L.A., Moiseeva, S.P.: Mathematical model of customers flow of a two-product trading company in the form of queueing system with repeated access to blocks. *News Tomsk Polytech. Univ.* **322**, 5–9 (2013). (in Russian)
12. Dammer, D.D., Nazarov, A.A.: Research of the mathematical model of the insurance company in form of the infinite queueing system by using method of asymptotic analysis. In: *Proceedings of 7th Ferghan conference “Limit theorems and its applications”*, Namangan, pp. 191–196 (2015). (in Russian)
13. Lee, W.C.Y.: *Mobile Cellular Telecommunications: Analog and Digital System*, 2nd edn. McGraw-Hill, New York (1995)
14. Nazarov, A.A., Moiseev, A.N.: *Queueing Systems and Networks with Unlimited Number of Servers*. NTL, Tomsk (2015). (in Russian)
15. Cox, D.R., Lewis, P.A.W.: *The statistical analysis of series of events*. Methuen and Co. Ltd., London (1966)
16. Dammer, D.D.: A mathematical model of insurance company in the form of a queueing system with an unlimited number of servers considering one-time insurance payments. In: Dudin, A., Nazarov, A. (eds.) *ITMM 2016. Communications in Computer and Information Science*, vol. 638, pp. 34–43. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44615-8_3
17. Nazarov, A.A., Terpugov, A.F.: *Queueing Theory*. NTL, Tomsk (2010). (in Russian)
18. Elsgolts, L.E.: *Differential Equations and Calculus of Variations*. Science, Moscow (1969). (in Russian)



Multi-channel Queuing Systems with Markovian Impatience

Yury I. Ryzhikov(✉)

Institute for Informatics and Automation of the Russian Academy of Sciences,
39, 14-th Line VO, St. Petersburg 199178, Russian Federation
ryzhbox@yandex.ru

Abstract. The iterative Takahashi—Takami method is adjusted to calculate distribution of the number of requests in the multi-phase systems with H_2 - service time and exponential distribution of the requests' "patience". The method of calculating the moments of waiting and sojourn time distributions for "successful" requests is also offered. The results are compared with the ones obtained from the simulation model. Application of the method is shown to calculate the successful request's sojourn time distribution in the queueing network.

Keywords: Queueing theory · Multi-phase systems · Iteration
Impatient requests

1 Introduction

Among many applications of the queuing theory, situations with *impatient customers* who have random restrictions on the request's sojourn time play a significant role. In telecommunications and military it could be some moving equipment with a limited time of staying in the zone of reach, in emergency situations people rescue, in court—lengthy legal procedures with deadlines, in medicine—critical patients whose conditions deteriorate rapidly in anticipation of emergency assistance, etc.

The simplest problem of this type—Markovian system with exponentially distributed service time [1–3]—has a very limited *practical* value. The attempt taken in [4] to generalize the approach for $M/H_2/n - H_2$ model proved ineffective due to the fast growth of the problem dimension.

A reasonable compromise would be a $M/H_2/n - M$ model. Specifics of implementation of an iterative method [5–7] are discussed below—with additional calculation of the system's sojourn time for "successful" requests, their ratio with respect to input flow, as well as calculating non-productive system losses due to incomplete service. This software model is in essence the only one that allows to generalize the problem on the *queueing networks*—thanks to the possibility to disregard the accumulated requests' patience due to its Markovian property.

2 Iterative Method for the Model $M/H_2/n - M$

The considered system receives a Poisson flow of requests of the intensity λ . The H_2 - servicing can be presented as an exponentially distributed for requests of two types, selected with probabilities y_1 and y_2 , with intensities μ_1 and μ_2 respectively. Any request's sojourn time in the system, regardless of its location (in the channel or in the queue), is limited by a random variable exponentially distributed with parameter γ .

Shown on Fig. 1 is a fragment of the diagram of transitions between microstates of the system $M/H_2/3 - M$ by outgoing, presented for 2-nd, 3-rd and 4-th layers.

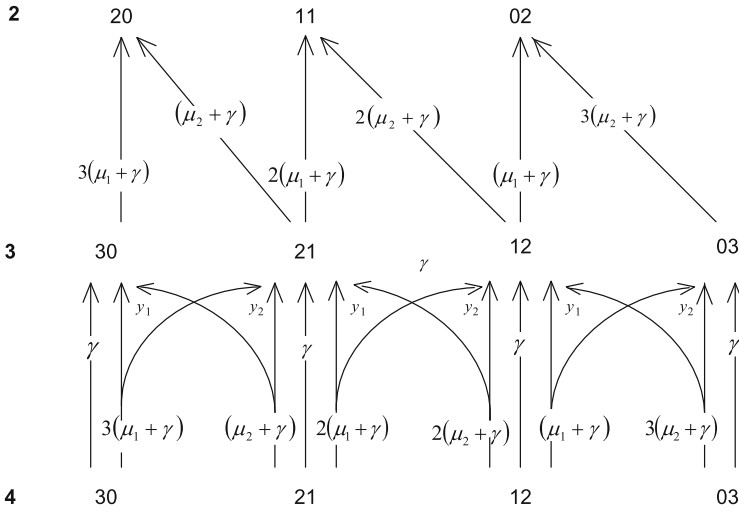


Fig. 1. Fragment of the withdrawal chart

Each layer corresponds to the number of system requests shown to the left (in our example—2, 3, 4). Code combinations like (2,1) indicate the distribution by type of requests being served; additional multipliers $\{y_i\}$ placed at the ends of arrows—the probabilities of selecting requests of the appropriate type from the queue. On layers $j > 3$ the difference will consist only in the intensity of additional *vertical* transitions ($\gamma, 2\gamma, 3\gamma, \dots$), reflecting exit of the impatient requests from a queue of the current length.

Let S_j be a set of all possible system microstates in which exactly j of requests are being served, and let m_j be a number of elements in S_j . Further, in accordance with the diagram of transitions for the selected model, let's build the matrices of the intensities of infinitesimal transitions:

- $A_j[m_j \times m_{j+1}]$ —in S_{j+1} (on arrival requests),
- $B_j[m_j \times m_{j-1}]$ —in S_{j-1} (full completion of service request),

$D_j[m_j \times m_j]$ —exit from the states of layer j (the matrices sizes are indicated in the square brackets). Calculation of these matrices in case of H_2 approximation of the service time distribution is easily programmed.

Let's introduce vectors-strings $g_j = \{g_{j,1}, g_{j,2}, \dots, g_{j,m_j}\}$ of probabilities for the system to be in the state (j, i) , $j = 0, 1, \dots$. Now it is possible to write down the vector-matrix equations of the transition balance

$$\begin{aligned} g_0 D_0 &= g_1 B_1, \\ g_j D_j &= g_{j-1} A_{j-1} + g_{j+1} B_{j+1}, \quad j = 1, 2, \dots \end{aligned} \tag{1}$$

Now we describe the general scheme of the iterative calculation of the stationary vectors of probabilities. Assume $t_j = \gamma_j/p_j$, where p_j is the cumulative probability of presence exactly j requests in the system, and define

$$x_j = p_{j+1}/p_j, \quad z_j = p_{j-1}/p_j. \tag{2}$$

With the *bottom-up* passage of the layers in the iteration number m the system of equations (1) can be rewritten with respect to vectors of conditional probabilities of the microstates normalized to 1 within a layer:

$$\begin{aligned} t_0^{(m)} D_0 &= x_0 t_1^{(m)} B_1, \\ t_j^{(m)} D_j &= z_j t_{j-1}^{(m-1)} A_{j-1} + x_j t_{j+1}^{(m)} B_{j+1}, \quad j = 1, 2, \dots \end{aligned} \tag{3}$$

Using vectors-columns $\mathbf{1}_j = \{1, 1, \dots, 1\}^T$ of size σ_j , the additional system conditions (3) for normalizing components to 1 can be written for all j

$$t_j \mathbf{1}_j = 1 \tag{4}$$

and the balance of the total intensities of transitions between adjacent layers

$$t_j^{(m)} A_j \mathbf{1}_{j+1} = x_j t_{j+1}^{(m)} B_{j+1} \mathbf{1}_j. \tag{5}$$

In the case of the system with an unlimited queue, the calculation algorithm for the set of vectors $\{t_j\}$ and the numbers $\{x_j\}$ and $\{z_j\}$ satisfying ratios (3)–(5), relies on the existence of a limit vector of conditional probabilities $t = \lim_{j \rightarrow \infty} t_j$, which is a consequence of the stabilization of transition matrices at $j > n$. The algorithm is based on a sequential approximation to the desired characteristics for a bounded set of indices $j = 0, \overline{N}$.

Let's rewrite the equations of the system (3) for $j \geq 1$ as

$$t_j^{(m)} = z_j \beta'_j + x_j \beta''_j, \tag{6}$$

where

$$\begin{aligned} \beta'_j &= t_{j-1}^{(m-1)} A_{j-1} D_j^{-1}, \\ \beta''_j &= t_{j+1}^{(m)} B_{j+1} D_j^{-1}. \end{aligned} \tag{7}$$

In this and subsequent formulas the products of matrices can be calculated before the start of iterations. In particular, their products by $\mathbf{1}_j$ are equal to the row sums, and the products

$$t_{j-1}^{(m)} A_{j-1} \mathbf{1}_j = \lambda. \quad (8)$$

One of the central ideas of the Takahashi—Takami method is the assumption about stabilization of conditional probabilities vectors $\{t_j\}$ for $j \rightarrow \infty$ confirmed by calculations. It allows to close the calculation scheme by assuming that for a layer with sufficient large number $j = N$

$$\beta''_N = t_{N-1}^{(m-1)} B_{N+1} D_N^{-1}. \quad (9)$$

This assumption was a consequence of the transition matrices stabilization already at $j = n + 1$. In our case, $\{B_j\}$ and $\{D_j\}$ are stabilized only at $j \rightarrow \infty$, but due to increasing decline of “impatient” requests, the cumulate layer probabilities $\{p_j\}$, having other equal conditions, will decrease much faster, and hence the errors from the above assumption will play a lesser role. Therefore, we will still use the condition (9) for the boundary layer N . The acceptability of this assumption can be verified by repeating the calculation for an increased value of N .

What is left to specify is how to calculate $\{z_j\}$ and $\{x_j\}$. We rewrite (5) accounting (7):

$$(z_j \beta'_j + x_j \beta''_j) A_j \mathbf{1}_{j+1} = z_j t_{j+1}^{(m)} B_{j+1} \mathbf{1}_j.$$

Hence we have the proportionality

$$z_j = c x_j$$

with a factor

$$c = \frac{t_{j+1}^{(m)} B_{j+1} \mathbf{1}_j - \beta''_j A_j \mathbf{1}_{j+1}}{\beta'_j A_j \mathbf{1}_{j+1}}.$$

Since all the products $A_j \mathbf{1}_{j+1}$ (the row sums of the matrices of requests arrival intensities) in the considered case of Poissonian incoming flow are equal to λ , the last formula can be represented as

$$c = \frac{t_{j+1}^{(m)} B_{j+1} \mathbf{1}_j - \lambda \beta''_j \mathbf{1}_j}{\lambda \beta'_j \mathbf{1}_j}. \quad (10)$$

Substitution of (8) in (6) and multiplying both parts of the result by $\mathbf{1}_j$ give

$$1 = t_j^{(m)} \mathbf{1}_j = x_j^{(m)} (c \beta'_j + \beta''_j) \mathbf{1}_j.$$

So,

$$x_j^{(m)} = 1 / [(c \beta'_j + \beta''_j) \mathbf{1}_j]. \quad (11)$$

A convenient criterion for terminating iterations is the condition

$$\max_j |x_j^{(m)} - x_j^{(m-1)}| \leq \varepsilon.$$

Due to rather obvious considerations, it is convenient to choose initial conditional distributions of the number of served requests of each type to be binomial with probabilities proportional to $\{y_i/\mu_i\}$.

After iterations termination, we can now move on to find cumulant probabilities. Assuming $p_0 = 1$, we sequentially calculate

$$p_{j+1} = p_j x_j, \quad j = \overline{0, N-1}, \tag{12}$$

and then normalize them to 1. We recall that sufficiency of the chosen N is determined by the smallness of the last probabilities.

Let’s compare the implementation of the described method and the corresponding simulation model for a three-channel system with H_2 - service time distribution (coefficient of variation $v = 2$)—Table 1:

Table 1. Probabilities of system states

j	Simulation	Calculation	j	Simulation	Calculation
0	8.578e-2	8.556e-2	10	5.577e-3	5.640e-3
1	1.986e-1	1.981e-1	11	2.838e-3	2.812e-3
2	2.256e-1	2.252e-1	12	1.319e-3	1.330e-3
3	1.650e-1	1.654e-1	13	6.093e-4	5.973e-4
4	1.178e-1	1.181e-1	14	2.611e-4	2.553e-4
5	8.096e-2	8.101e-2	15	1.188e-4	1.040e-4
6	5.286e-2	5.295e-2	16	4.774e-5	4.042e-5
7	3.275e-2	3.284e-2	17	1.886e-5	1.502e-5
8	1.924e-2	1.929e-2	18	4.876e-6	1.822e-6
9	1.064e-2	1.072e-2	19	3.149e-6	5.966e-7

Simulation was carried out before acceptance for service of 2 million requests. Therefore, the agreement of the calculated and statistical probabilities of order 10^{-4} and more, for which the number of observations exceeded 100, should be considered very good.

3 Distribution of the Waiting Time

The biggest problem in the systems with impatient requests calculation is the definition of their temporary’s characteristics. Here, the known formula for the moments of waiting time distribution

$$w_j = q_{[j]}/\lambda^j, \quad j = 1, 2, \dots, \tag{13}$$

in which $\{q_{[j]}\}$ are the factorial moments of queue length distribution, as shown by simulation experiments, gives poor accuracy.

Let us calculate

- vectors-rows $g_k = p_k * t_k$ of the stationary probabilities of microstates, $k = 0, 1, \dots$,
- diagonal matrices of total up-transition intensities with elements $\{\sigma_{k,i}\}$,
- diagonal matrices $\{U_k(s)\}$ of the Laplace-Stieltjes transformations (LST) of the distributions of up-transitions duration by the corresponding total intensities $\{\sigma_k\}$ with elements $\{\sigma_{k,i}/(\sigma_{k,i} + s)\}$,
- the product $\tilde{U}_n(s)$ of the matrix $U_n(s)$ by a unit vector-column,
- matrices $\{T_k\}$ with elements $\{b_{k,i,j}/\sigma_{k,i}\}$ of the probabilities of transitions on the overlying layer.

In addition, we replace the diagonal matrix $U_n(s)$ by the same name vector-column. It is not difficult to see that the LST of waiting time outputs directly on service from n -th layer is

$$\omega_n(s) = g_n \tilde{U}_n(s).$$

For the $(n + 1)$ -th layer we have

$$\omega_{n+1}(s) = g_{n+1} [U_{n+1}(s) T_{n+1}] \tilde{U}_n(s),$$

for the $(n + 2)$ —

$$\omega_{n+2}(s) = g_{n+2} [U_{n+2}(s) T_{n+2} U_{n+1}(s) T_{n+1}] \tilde{U}_n(s),$$

etc. Given the above rules for forming factors $F_k(s) = U_k(s) T_k$ included in these formulas, we can immediately define matrices $\{F_k\}$ as the sets of elements type $\{b_{k,i,j}/(\sigma_{k,i} + s)\}$. Summing up the results for all possible starting layers, we get the final formula for the LST of waiting time distribution:

$$\omega(s) = \left[\sum_{k=0}^{\infty} g_{n+k} \prod_{i=0}^k F_{n+i}(s) \right] \tilde{U}_n(s). \tag{14}$$

In this formula, the *inverted* product symbol is used to specify the inverse order of cofactors (be reminded that the multiplication of matrices in general is non-commutative). The initial value is $F_n(s) = I$.

Having calculated a table of LST values in a neighbourhood of zero, we can construct its approximation by the Newton interpolation polynomial, and obtain the moments of waiting time distribution by multiple differentiation of the latter.

All is left is to consider the possibility of “impatience” for the same labeled request. Because for each request, the probability to endure the time u is equal to $e^{-\gamma u}$, the distribution function of a successful waiting

$$W^+(t) = \int_0^t e^{-\gamma u} w(u) du.$$

Accordingly, the LST of this distribution

$$\omega^+(s) = \int_0^\infty e^{-st} d \left[\int_0^t e^{-\gamma u} w(u) du \right] dt.$$

The derivative of integral by the parameter in this case is $e^{-\gamma t} w(t)$. Hence,

$$\omega^+(s) = \int_0^\infty e^{-st} e^{-\gamma t} w(t) dt = \omega(s + \gamma).$$

Thus, LST of the *successful* waiting should be calculated according to (14) with replacing the argument s by $s + \gamma$.

The moments of successful waiting, obtained in this way, should be divided by the probability of a successful wait, including zero, that is, by

$$\pi_w = \sum_{k=0}^{n-1} p_k + \omega^+(\gamma).$$

We compare the numerical results obtained by this technique and by means of simulation (2 million of served requests). For a three-channel system with the intensity $\lambda = 1.5$ of the incoming Poissonian flow, average service time $b_1 = 4.0$, the service variation factor $v_b = 2.0$ and impatience intensity $\gamma = 0.2$ the results are summarized in Table 2.

Table 2. Moments of the distribution of successful waiting

Method	w_1^+	w_2^+	w_3^+
Imitation	0.483	1.048	3.319
Calculation	0.479	1.039	3.287

4 Distribution of a Successful Request Sojourn Time

When the request from a queue has been extracted, the assumption of the permissible patience having Markovian distribution allows to count down its patience *anew*. After all, we are only interested in “successful” requests which received complete servicing. Suppose the distribution of the latter be two-phase hyper-exponential with parameters $\{y_m, \mu_m\}$. Then the j -th moment of the time of successful servicing

$$b_j^+ = \int_0^\infty \left[\int_0^\theta t^j b(t) dt \right] \gamma e^{-\gamma \theta} d\theta = \int_0^\infty \left[\int_0^\theta t^j \sum_{m=1}^2 y_m \mu_m e^{-\mu_m t} dt \right]. \quad (15)$$

It can be shown that

$$\begin{aligned}
 b_j^+ &= j! \sum_{m=1}^2 \frac{y_m}{\mu_m^j} - j! \sum_{m=1}^2 \frac{y_m}{\mu_m^j} \cdot \sum_{i=0}^j \frac{\gamma}{\mu_m + \gamma} \left(\frac{\mu_m}{\mu_m + \gamma} \right)^i \\
 &= j! \sum_{m=1}^2 \frac{y_m}{\mu_m^j} \left[1 - \frac{\gamma}{\mu_m + \gamma} \sum_{i=0}^j \left(\frac{\mu_m}{\mu_m + \gamma} \right)^i \right].
 \end{aligned} \tag{16}$$

In accordance with (15), the zero moment of the successful servicing can be considered as a probability π_s of the such. Substituting $j = 0$ in (15), we get

$$\pi_s = \sum_{m=1}^2 u_m \left[1 - \frac{\gamma}{\mu_m + \gamma} \right] = \sum_{m=1}^2 \frac{y_m \mu_m}{\mu_m + \gamma}. \tag{17}$$

It is of interest to estimate the volume of wasted service. The average service time of the interrupted request

$$\bar{\tau} = \int_0^{\infty} \theta \bar{B}(\theta) \gamma e^{-\gamma \theta} d\theta,$$

where $\bar{B}(\theta)$ is the complementary distribution function of the full service duration. In our problem

$$\begin{aligned}
 \bar{\tau} &= \int_0^{\infty} \theta \left[\sum_{m=1}^2 y_m e^{-\mu_m \theta} \right] \gamma e^{-\gamma \theta} d\theta \\
 &= \gamma \sum_{m=1}^2 y_m \int_0^{\infty} \theta e^{-(\mu_m + \gamma) \theta} d\theta = \gamma \sum_{m=1}^2 \frac{y_m}{(\mu_m + \gamma)^2}.
 \end{aligned} \tag{18}$$

Total losses per unit of time will be

$$g = \lambda \pi_w (1 - \pi_s) \bar{\tau}.$$

The moments of $\{v_j^+\}$ of a successful stay in the system are calculated via convolution of $\{w_j^+\}$ and $\{b_j^+\}$, and the probability of a successful stay in the system

$$\pi_v = \pi_w \pi_s.$$

5 Calculation of a Network with Impatient Requests

The assumption that permissible patience has the Markovian distribution allows us to apply the results of previous section to the calculation of networks with hyper-exponential servicing, using their flow-equivalent decomposition. Since in our case the intensity of the exiting successful flow differs from the intensity of the incoming one, it is necessary to make the following changes in the usual scheme of open network calculation:

1. There should be no cyclic routes in the network.
2. The numbering and, respectively, the order of the nodes calculation must be determined on the basis of a preceding relationship—for example, using the well-known Floyd algorithm.
3. The nodes with impatience must be calculated using the above method. The intensities of the output flow for such nodes must be calculated via multiplying the incoming intensities by the corresponding probability π_i .

The network sojourn time calculation deserves a special consideration. When talking about most critical applications, it isn't enough to know the average sojourn time—such cases usually raise the question of the highest moments and/or calculation of the distribution function. Appropriate technique [8] is based on the construction of the LST for the network sojourn time distribution via the “nodal” LST, a routing matrix, and its' subsequent numerical differentiation at zero.

6 Conclusion

The main results of this work are as follows:

1. A diagram of transitions between microstates of the model $M/H_2/n - M$ is proposed taking into account exponentially distributed patience of all requests located in the system. On its basis, the rules are corrected to calculate the matrices $\{B_j\}$ and $\{D_j\}$ of the transition intensities.
2. Permissibility of using the formula (8) was justified, which allows to limit the number of accounted layers.
3. The stationary probabilities of system states, moments of successful waiting and sojourn time were obtained and compared with their analogs received by simulation.
4. The formulas for calculating average losses from interrupted service per unit of time and the intensity of the flow of successful requests were proposed
5. Application of these results to the calculation of the *queueing networks* service with impatient requests was demonstrated.

Acknowledgments. The work described in the paper was supported by state project 0073-2018-0003.

References

1. Takagi, H.: Waiting time in the M/M/m/(m+c) queue with impatient customers. *Int. J. Pure Appl. Math.* **90**(4), 519–559 (2014)
2. Aktekin, T., Soyer, R.: Bayesian analysis of queues with impatient customers: applications to call centers. *Nav. Res. Logist.* **59**, 441–456 (2012)
3. Boot, N.K., Tijm, H.: A multiserver queueing system with impatient customers. *Manag. Sci.* **45**(3), 444–448 (1999)

4. Ryzhikov, Yu.I., Ulanov, A.V.: Calculation of the hyperexponential queueing system $M/H_2/n - H_2$ with requests impatient in the queue. Bull. Tomsk State Univ.: Manag. Comput. Technol. Comput. Sci. **2**(27), 47–53 (2014). (in Russian)
5. Ryzhikov, Yu.I.: Iterative method for calculating multi-channel queueing systems - the basics, modifications and limiting opportunities. In: Proceedings of the 9th Russian Multiconference on Control Problems. Information Technologies in Management. SPb.: “Concern Elektropribor”, pp. 224–233 (2016). (in Russian)
6. Seelen, L.P.: An algorithm for Ph/Ph/c Queues. Eur. J. Oper. Res. **23**, 118–127 (1986)
7. Takahashi, Y., Takami, Y.: A numerical method for the steady-state probabilities of a GI/G/c queueing system in a general class. J. Oper. Res. Soc. Jpn. **19**(2), 147–157 (1976)
8. Ryzhikov, Yu.I.: An algorithmic approach to queueing problems: monograph. A.F. Mojaysky VKA (Military Space Academy), St.-Petersburg (2013). (in Russian)



Optimal State Estimation of Semi-synchronous Event Flow of the Second Order Under Its Complete Observability

Luydmila Nezhelskaya and Diana Tumashkina^(✉)

National Research Tomsk State University, Tomsk, Russia
ludne@mail.ru, diana1323@mail.ru

Abstract. We consider the optimal estimation problem for the states of a semi-synchronous event flow of the second order with two states; it is one of the adequate mathematical models for an incoming stream of claims (events) in modern digital integral servicing networks, telecommunication systems, satellite communication networks. We find an explicit form for posterior probabilities of the flow states. The decision about the flow state is made with the maximal a posteriori criterion.

Keywords: Semi-synchronous event flow · Optimal states estimation
Posterior probabilities · Maximal a posteriori criterion

1 Introduction

In modern times information flows of messages are functioning in telecommunication systems, satellite communication networks and global computer networks. Doubly stochastic flows of events are their adequate mathematical models [1, 2]. These flows are characterized by double randomness, namely: the moments when events occur are random and the intensity of the flow is a random process.

A semi-synchronous event flow of the second order is the object of studying in this work. A semi-synchronous flow is an integration of the following types of flows: a synchronous flow, where the transition from state to state depends directly on the occurrence of the event [3], and asynchronous, where the transition from state to state does not depend on whether the event has occurred or not [4].

The main problems in the studying of doubly stochastic event flows are problems that are realized by observing the moments when events occur: (1) estimating the states of an event flow [5–7]; (2) estimating flow parameters [8–10].

We emphasize that the mathematical models of doubly stochastic flows of events, in particular the model of the semi-synchronous event flow of the second order is considered in this paper, are the most characteristic and appropriate flow models in real telecommunication systems and networks [11, 12].

We propose an algorithm for optimal estimation of the states of the flow under consideration by the method of maximum a posterior probability in this paper. The application of this method is due to the fact that a posterior probability is a characteristic possessing the most complete information about the process being investigated, that is contained in the sample of observations, and also because the method of maximum a posterior probability provides a minimum of the total probability of error in making the decision [13]. This article is a direct development of [14, 15].

2 Problem Setting

We consider the stationary operation mode of a semi-synchronous doubly stochastic event flow of the second order (hereinafter flow), the accompanying random process of which is a piecewise constant process $\lambda(t)$ with two states S_1 and S_2 . Hereinafter, the i th state of the process is understood as the state S_i , $i = 1, 2$.

The duration of the interval between the flow events at the first state is determined by the random variable $\eta = \min(\xi^{(1)}, \xi^{(2)})$, where random variable $\xi^{(1)}$ has distribution function $F_1^{(1)}(t) = 1 - e^{-\lambda_1 t}$, random variable $\xi^{(2)}$ has distribution function $F_1^{(2)}(t) = 1 - e^{-\alpha_1 t}$; $\xi^{(1)}$ and $\xi^{(2)}$ are independent random variables.

At the moment of the flow event occurrence, the process $\lambda(t)$ transits from the first state to the second either with probability $P_1^{(1)}(\lambda_2|\lambda_1)$, or with probability $P_1^{(2)}(\lambda_2|\lambda_1)$, depending on what value the random variable η has taken. At the moment of the flow event occurrence, the process $\lambda(t)$ remains at the first state either with probability $P_1^{(1)}(\lambda_1|\lambda_1)$, or with probability $P_1^{(2)}(\lambda_1|\lambda_1)$, depending on what value the random variable η has taken. Here $P_1^{(1)}(\lambda_2|\lambda_1) + P_1^{(1)}(\lambda_1|\lambda_1) = 1$, $P_1^{(2)}(\lambda_2|\lambda_1) + P_1^{(2)}(\lambda_1|\lambda_1) = 1$. The duration of the interval between the flow events at the first state is random variable with distribution function $F(t) = 1 - e^{-(\lambda_1 + \alpha_1)t}$.

The time during which the process $\lambda(t)$ remains at the second state is random variable with distribution function $F_2(t) = 1 - e^{-\alpha_2 t}$. During the time when the process $\lambda(t)$ is in the second state, there is a Poisson event flow with parameter λ_2 .

Hereinafter, it is assumed that the state S_i (i th state) of the process $\lambda(t)$ takes place if $\lambda(t) = \lambda_i$, $i = 1, 2$; $\lambda_1 > \lambda_2 \geq 0$.

The infinitesimal characteristics matrices for the process $\lambda(t)$ are as follows

$$\mathbf{D}_0 = \begin{vmatrix} -(\lambda_1 + \alpha_1) & 0 \\ \alpha_1 & -(\lambda_2 + \alpha_2) \end{vmatrix},$$

$$\mathbf{D}_1 = \begin{vmatrix} \lambda_1 P_1^{(1)}(\lambda_1|\lambda_1) + \alpha_1 P_1^{(2)}(\lambda_1|\lambda_1) & \lambda_1 P_1^{(1)}(\lambda_2|\lambda_1) + \alpha_1 P_1^{(2)}(\lambda_2|\lambda_1) \\ 0 & \lambda_2 \end{vmatrix}.$$

Elements of the matrix \mathbf{D}_1 are the intensities of the process transitions from state to state with an event occurrence. Nondiagonal elements of the matrix \mathbf{D}_0

are the intensities of transitions from state to state without an event. In turn, the diagonal elements of the matrix \mathbf{D}_0 are the intensities of the process exit from its states taken with the opposite sign.

An example of one of the realizations of the process $\lambda(t)$ and the event flow are shown on Fig. 1, where t_1, t_2, \dots denote the moments when events occur in the flow.

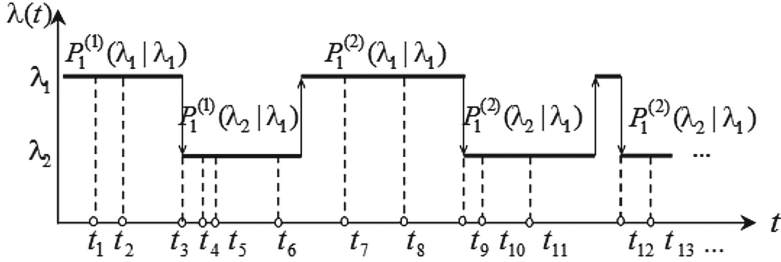


Fig. 1. Semi-synchronous event flow of the second order

Assertion. $\lambda(t)$ is a Markov process.

Proof. It is not difficult to show that the time during which the process $\lambda(t)$ remains at the first state is a random variable is distributed according to the exponential law with the distribution function $F_1(t) = 1 - e^{-[\lambda_1 P_1^{(1)}(\lambda_2 | \lambda_1) + \alpha_1 P_1^{(2)}(\lambda_2 | \lambda_1)]t}$.

In turn, the time during which the process $\lambda(t)$ remains at the second state is a random variable is distributed according to the exponential law $F_2(t) = 1 - e^{-\alpha_2 t}$. This implies the statement of the assertion.

Since the process $\lambda(t)$ is unobservable in principle, and we can only observe time moments t_1, t_2, \dots when events occur in the flow, then $\lambda(t)$ is a hidden Markov process or an unobservable accompanying Markov process.

We have to estimate the state of the process $\lambda(t)$ (flow) at time moment t , when the observations have stopped by observations t_1, t_2, \dots of the event flow over the time interval (t_0, t) , where t_0 denotes the beginning of observations. We can let $t_0 = 0$ without loss of generality.

To make the decision regarding the state of the process $\lambda(t)$ at time moment t , we have to determine posterior probabilities $w(\lambda_i | t) = w(\lambda_i | t_1, \dots, t_m, t) = P(\lambda(t) = \lambda_i | t_1, \dots, t_m, t)$, $i = 1, 2$, of the fact that at time moment t the value of the process $\lambda(t) = \lambda_i$ (m is the number of events in time t), here $w(\lambda_1 | t) + w(\lambda_2 | t) = 1$. The optimal estimation is as follows: if $w(\lambda_i | t) \geq w(\lambda_j | t)$, $i, j = 1, 2$, $i \neq j$, then the estimation of the state of the process is $\hat{\lambda}(t) = \lambda_i$, otherwise $\hat{\lambda}(t) = \lambda_j$, $i, j = 1, 2$.

3 Optimal Estimation Algorithm for the States of the Semi-synchronous Event Flow of the Second Order

We will consider time intervals $(t_k, t_{k+1}), k = 1, 2, \dots$, between neighboring events in the flow. Denote t as the decision making moment, here $0 < t < t_1$ or $t_k < t < t_{k+1}, k = 1, 2, \dots$

Lemma 1. *On time intervals $(0, t_1)$ and $(t_k, t_{k+1}), k = 1, 2, \dots$, the posterior probability $w(\lambda_1|t)$ satisfies the Riccati differential equation*

$$\frac{dw(\lambda_1|t)}{dt} = (\lambda_1 - \lambda_2 + \alpha_1)w^2(\lambda_1|t) - (\lambda_1 - \lambda_2 + \alpha_1 + \alpha_2)w(\lambda_1|t) + \alpha_2. \quad (1)$$

Proof. To derive the formulas for posterior probabilities we use the method of obtaining recurrence relations for posterior probabilities as described in [13]: we first consider discrete observations divided by sufficiently small time intervals Δt and then make the limit transition as Δt tends to zero.

We suppose that the time is discrete and changes with step $\Delta t : t^{(n)} = n\Delta t, n = 0, 1, \dots$. We introduce a two-dimensional process $(\lambda^{(n)}, r_n), (\lambda^{(n)}, r_n) = (\lambda(n\Delta t), r_n(\Delta t)) = (\lambda(n\Delta t), r(n\Delta t) - r((n-1)\Delta t))$ where $\lambda^{(n)} = \lambda(n\Delta t)$ is the value of process $\lambda(t)$ at time moment $t^{(n)} = n\Delta t (\lambda^{(n)} = \lambda_i, i = 1, 2); r_n = r_n(\Delta t)$ is the number of events in the flow occurred on the interval $((n-1)\Delta t, n\Delta t)$ of length $\Delta t, r_n = 0, 1, \dots$. Note that this two-dimensional process is Markovian.

We denote by $\mathbf{r}_m = (r_0, r_1, \dots, r_m)$ the sequence of the number of events in time from 0 to $m\Delta t$ on intervals $((n-1)\Delta t, n\Delta t)$ of length $\Delta t, n = 0, 1, \dots, m$, where r_0 is the number of events on the interval $(-\Delta t, 0)$; since there are no observations on this interval, so we can set an arbitrary value to it, say $r_0 = 0$. We denote by $\boldsymbol{\lambda}^{(m)} = (\lambda^{(0)}, \lambda^{(1)}, \dots, \lambda^{(m)})$ the sequence of unobservable values of the process $\lambda(n\Delta t)$ at time moments $n\Delta t, n = 0, 1, \dots, m$, where $\lambda^{(0)} = \lambda(0) = \lambda_i, i = 1, 2$. We denote by $(\lambda^{(m)}|\mathbf{r}_m)$ the posterior probability of the value $\lambda^{(m)}$ under the condition that we have observed a realization \mathbf{r}_m .

For the Markov random process $(\lambda^{(n)}, r_n)$, a recurrent relation is obtained in [16] for the posterior probabilities $w(\lambda^{(m+1)}|\mathbf{r}_{m+1})$ and $w(\lambda^{(m)}|\mathbf{r}_m)$

$$\begin{aligned} w(\lambda^{(m+1)}|\mathbf{r}_{m+1}) &= \\ &= \frac{\sum_{\lambda^{(m)}=\lambda_1}^{\lambda_2} w(\lambda^{(m)}|\mathbf{r}_m)p(\lambda^{(m+1)}, r_{m+1}|\lambda^{(m)}, r_m)}{\sum_{\lambda^{(m)}=\lambda_1}^{\lambda_2} \sum_{\lambda^{(m+1)}=\lambda_1}^{\lambda_2} w(\lambda^{(m)}|\mathbf{r}_m)p(\lambda^{(m+1)}, r_{m+1}|\lambda^{(m)}, r_m)}, \end{aligned} \quad (2)$$

where $p(\lambda^{(m+1)}, r_{m+1}|\lambda^{(m)}, r_m)$ is the transition probability for the process $(\lambda^{(n)}, r_n)$ in one step Δt from state $(\lambda^{(m)}, r_m)$ to state $(\lambda^{(m+1)}, r_{m+1})$. Due to the fact that the two-dimensional process under consideration is Markovian and also because of the construction of the semi-synchronous event flow of the second order, the transition probability can be written as

$$p(\lambda^{(m+1)}, r_{m+1}|\lambda^{(m)}, r_m) = p(\lambda^{(m+1)}|\lambda^{(m)})p(r_{m+1}|\lambda^{(m)}, I(\lambda^{(m)})), \quad (3)$$

where the indicator

$$I(\lambda^{(m)}) = \begin{cases} \lambda^{(m+1)}, & \text{if } \lambda^{(m)} = \lambda_1, \\ 0, & \text{if } \lambda^{(m)} = \lambda_2. \end{cases}$$

Taking into account (3) and that $w(\lambda^{(m)}|\mathbf{r}_m) = w(\lambda^{(m)}|\mathbf{r}_m(t)) = w(\lambda^{(m)}|t)$, $w(\lambda^{(m+1)}|\mathbf{r}_{m+1}) = w(\lambda^{(m+1)}|\mathbf{r}_{m+1}(t + \Delta t)) = w(\lambda^{(m+1)}|t + \Delta t)$ we can rewrite (2) as

$$\begin{aligned} & w(\lambda^{(m+1)}|t + \Delta t) = \\ & = \frac{\sum_{\lambda^{(m)}=\lambda_1}^{\lambda_2} w(\lambda^{(m)}|t)p(\lambda^{(m+1)}, \lambda^{(m)})p(r_{m+1}|\lambda^{(m)}, I(\lambda^{(m)}))}{\sum_{\lambda^{(m)}=\lambda_1}^{\lambda_2} \sum_{\lambda^{(m+1)}=\lambda_1}^{\lambda_2} w(\lambda^{(m)}|t)p(\lambda^{(m+1)}, \lambda^{(m)})p(r_{m+1}|\lambda^{(m)}, I(\lambda^{(m)}))}. \end{aligned} \quad (4)$$

Let in (4) $\lambda^{(m+1)} = \lambda_1$ for definiteness. The value r_{m+1} takes only two values $r_{m+1} = 0, 1$ due to the definition of the semi-synchronous event flow of the second order. Situations where $r_{m+1} = 2, 3, \dots$ has probability equal $o(\Delta t)$. We consider the behavior of posterior probability $w(\lambda_1|t)$ on the interval (t_k, t_{k+1}) between neighboring events of the flow, i.e. consider the case of the absence of events on the observation interval $(t, t + \Delta t)$. Then $r_{m+1} = 0$ and taking into account the matrix \mathbf{D}_0 the transition probabilities in the recurrence relation (4) on the interval $(t, t + \Delta t) = (m\Delta t, (m + 1)\Delta t)$ take the following form

$$\begin{aligned} & p(\lambda^{(m+1)} = \lambda_i|\lambda^{(m)} = \lambda_i)p(r_{m+1} = 0|\lambda^{(m)} = \lambda_i, \lambda^{(m+1)} = \lambda_i) = \\ & = 1 - \lambda_i\Delta t - \alpha_i\Delta t + o(\Delta t), \quad i = 1, 2, \\ & p(\lambda^{(m+1)} = \lambda_1|\lambda^{(m)} = \lambda_2)p(r_{m+1} = 0|\lambda^{(m)} = \lambda_2) = \alpha_2\Delta t + o(\Delta t), \\ & p(\lambda^{(m+1)} = \lambda_2|\lambda^{(m)} = \lambda_1)p(r_{m+1} = 0|\lambda^{(m)} = \lambda_1, \lambda^{(m+1)} = \lambda_2) = 0. \end{aligned}$$

Substituting these expressions into (4), we obtain

$$w(\lambda_1|t + \Delta t) = \frac{w(\lambda_1|t) - (\lambda_1 + \alpha_1)w(\lambda_1|t)\Delta t + \alpha_2w(\lambda_2|t)\Delta t + o(\Delta t)}{1 - \Delta t \left[(\lambda_1 + \alpha_1)w(\lambda_1|t) + \lambda_2w(\lambda_2|t)\Delta t + \frac{o(\Delta t)}{\Delta t} \right]}. \quad (5)$$

Taking into account the fact that $(1 - x)^{-1} = 1 + x + o(x)$ where $x > 0$ is a sufficiently small quantity, we can rewrite (5) as

$$\begin{aligned} & w(\lambda_1|t + \Delta t) - w(\lambda_1|t) = \\ & = -[(\lambda_1 + \alpha_1)w(\lambda_1|t) - \alpha_2w(\lambda_2|t) - (\lambda_1 + \alpha_1)w^2(\lambda_1|t) - \lambda_2w(\lambda_1|t)w(\lambda_2|t)]\Delta t + o(\Delta t). \end{aligned}$$

Taking into account that $w(\lambda_2|t) = 1 - w(\lambda_1|t)$, dividing both sides of the last equality by Δt and passing to the limit for $\Delta t \rightarrow 0$, we find (1). *Lemma 1 is proved.*

Lemma 2. *The posterior probability $w(\lambda_1|t)$ at the time a semi-synchronous flow of the second order event t_k , $k = 1, 2, \dots$, occurs is given by the following conversion formula:*

$$w(\lambda_1|t_k + 0) = \frac{[\lambda_1 P_1^{(1)}(\lambda_1|\lambda_1) + \alpha_1 P_1^{(2)}(\lambda_1|\lambda_1)]w(\lambda_1|t_k - 0)}{\lambda_2 + [\lambda_1 + \alpha_1 - \lambda_2]w(\lambda_1|t_k - 0)}, k = 1, 2, \dots \quad (6)$$

Proof. Let $r_{m+1} = 1$, i.e. a flow event occurs on the interval $(t, t + \Delta t)$ at time moment t_k , $t < t_k < t + \Delta t$. Thus, we have two adjacent intervals (t, t_k) and $(t_k, t + \Delta t)$, whose durations are $\Delta t' = t_k - t$ and $\Delta t'' = t + \Delta t - t_k$ respectively. Then $w(\lambda^{(m)}|t) = w(\lambda^{(m)}|t_k - \Delta t')$, $w(\lambda^{(m+1)}|t + \Delta t) = w(\lambda^{(m+1)}|t_k + \Delta t'')$ take place and (4) takes the following form for $\lambda^{(m+1)} = \lambda_1$

$$\begin{aligned} & w(\lambda_1|t_k + \Delta t'') = \\ &= \frac{\sum_{\lambda^{(m)}=\lambda_1}^{\lambda_2} w(\lambda^{(m)}|t_k - \Delta t')p(\lambda_1|\lambda^{(m)})p(r_{m+1} = 1|\lambda^{(m)}, I(\lambda^{(m)}))}{\sum_{\lambda^{(m)}=\lambda_1}^{\lambda_2} \sum_{\lambda^{(m+1)}=\lambda_1}^{\lambda_2} w(\lambda^{(m)}|t_k - \Delta t')p(\lambda^{(m+1)}, \lambda^{(m)})p(r_{m+1} = 1|\lambda^{(m)}, I(\lambda^{(m)}))}. \quad (7) \end{aligned}$$

Taking into account the matrix \mathbf{D}_1 on the interval $(t, t + \Delta t) = (m\Delta t, (m + 1)\Delta t)$ we can rewrite transition probabilities in (7) as

$$\begin{aligned} p(\lambda^{(m+1)} = \lambda_i|\lambda^{(m)} = \lambda_1)p(r_{m+1} = 1|\lambda^{(m)} = \lambda_1, \lambda^{(m+1)} = \lambda_i) \\ &= [\lambda_1 P_1^{(1)}(\lambda_i|\lambda_1) + \alpha_1 P_1^{(2)}(\lambda_i|\lambda_1)]\Delta t + o(\Delta t), \quad i = 1, 2, \\ p(\lambda^{(m+1)} = \lambda_2|\lambda^{(m)} = \lambda_2)p(r_{m+1} = 1|\lambda^{(m)} = \lambda_2) &= \lambda_2\Delta t + o(\Delta t), \\ p(\lambda^{(m+1)} = \lambda_1|\lambda^{(m)} = \lambda_2)p(r_{m+1} = 1|\lambda^{(m)} = \lambda_2) &= 0. \end{aligned}$$

Substituting the expressions for the transition probabilities into (7), we obtain

$$\begin{aligned} w(\lambda_1|t_k + \Delta t'') &= \{[\lambda_1 P_1^{(1)}(\lambda_1|\lambda_1) + \alpha_1 P_1^{(2)}(\lambda_1|\lambda_1)]w(\lambda_1|t_k - \Delta t')\Delta t + o(\Delta t)\} \times \\ &\times \{[(\lambda_1 + \alpha_1)w(\lambda_1|t_k - \Delta t') + \lambda_2 w(\lambda_2|t_k - \Delta t')]\Delta t + o(\Delta t)\}^{-1}. \end{aligned}$$

Taking into account that $w(\lambda_2|t_k - \Delta t) = 1 - w(\lambda_1|t_k - \Delta t')$ dividing the numerator and denominator of the last equality by Δt and passing to the limit for $\Delta t \rightarrow 0$ ($\Delta t'$ and $\Delta t''$ tend to zero simultaneously), we find (6). *Lemma 2 is proved.*

Remark. The point t_k , which is a solution of Eq. (1), is the point of discontinuity of the first kind for the probability $w(\lambda_1|t)$. The probability $w(\lambda_1|t_k + 0)$ depends on the value $w(\lambda_1|t_k - 0)$, where $w(\lambda_1|t_k - 0)$ is the value of probability $w(\lambda_1|t)$ at the time moment t_k when t changes on the interval (t_{k-1}, t_k) , $k = 1, 2, \dots$. Thus, probability $w(\lambda_1|t_k + 0)$ contains information about the entire prehistory of flow observations starting from the time moment t_0 of the beginning of observations until the moment t_k of the k th event occurrence of the flow.

Let $\pi_1(t|t^0)$ be a prior probability that the process $\lambda(t) = \lambda_1$ at time moment t under the condition that the flow functioning started at the time moment t^0 . Then the differential equation for the introduced probability takes place

$$\pi_1'(t|t^0) = -[\lambda_1 P_1^{(1)}(\lambda_2|\lambda_1) + \alpha_1 P_1^{(2)}(\lambda_2|\lambda_1) + \alpha_2] \pi_1(t|t^0) + \alpha_2,$$

integrating which and taking into account the initial condition $\pi_1(t^0|t^0) = \pi$ and considering the obtained solution under the stationary operation mode ($t \rightarrow \infty$ or $t^0 \rightarrow -\infty$), we find the prior final probability of the first state of the process $\lambda(t)$:

$$\pi_1 = \frac{\alpha_2}{\lambda_1 P_1^{(1)}(\lambda_2|\lambda_1) + \alpha_1 P_1^{(2)}(\lambda_2|\lambda_1) + \alpha_2}. \quad (8)$$

Lemmas 1 and 2 yield the following theorem.

Theorem 1. *On time intervals $(0, t_1)$ and (t_k, t_{k+1}) , $k = 1, 2, \dots$, the posterior probability $w(\lambda_1|t)$ follows the following explicit formula:*

$$\begin{aligned} w(\lambda_1|t) &= \\ &= \frac{w_1[1 - w(\lambda_1|t_k + 0)] - [w_1 - w(\lambda_1|t_k + 0)]e^{-(\lambda_1 - \lambda_2 + \alpha_1)(1 - w_1)(t - t_k)}}{[1 - w(\lambda_1|t_k + 0)] - [w_1 - w(\lambda_1|t_k + 0)]e^{-(\lambda_1 - \lambda_2 + \alpha_1)(1 - w_1)(t - t_k)}}, \end{aligned} \quad (9)$$

where $w_1 = \frac{\alpha_2}{\lambda_1 - \lambda_2 + \alpha_1}$, $t_k < t < t_{k+1}$, $k = 0, 1, \dots$; $w(\lambda_1|t_k + 0)$, $k = 1, 2, \dots$, is given by (6); $w(\lambda_1|t_0 + 0) = \pi$, π is defined in (8).

Proof. Integrating Eq. (1) with the initial condition $w(\lambda_1|t) = w(\lambda_1|t_k + 0)$ at time moment t_k of the event occurrence of the flow, we obtain the assertion of the theorem. *Theorem is proved.*

Thus, formulas (6), (8), (9) let us construct the optimal estimation algorithm for the states of the semi-synchronous event flow of the second order, in other words, the algorithm to compute probabilities $w(\lambda_1|t)$, $w(\lambda_2|t) = 1 - w(\lambda_1|t)$ and the algorithm to make a decision regarding the state of the process $\lambda(t)$ at any time moment t :

- (1) according to formula (8), compute the prior probability of the first state of the process under consideration $w(\lambda_1|t_0 + 0) = w(\lambda_1|t_0 = 0) = \pi_1$ at time moment t ;
- (2) according to formula (9), for $k = 0$ compute the probability $w(\lambda_1|t)$ at time moment t , $t_0 < t < t_1$, where t_1 is the moment when the first event occurs in the flow, here π_1 is the initial condition at this step;
- (3) for $k = 0$ compute the probability $w(\lambda_1|t_1) = w(\lambda_1|t_1 - 0)$ at time moment t_1 by following formula

$$w(\lambda_1|t_1) = \frac{w_1[1 - \pi_1] - [w_1 - \pi_1]e^{-(\lambda_1 - \lambda_2 + \alpha_1)(1 - w_1)(t_1 - t_0)}}{[1 - \pi_1] - [w_1 - \pi_1]e^{-(\lambda_1 - \lambda_2 + \alpha_1)(1 - w_1)(t_1 - t_0)}};$$

- (4) according to formula (6) for $k = 1$ compute the posterior probability $w(\lambda_1|t_1 + 0)$ at time moment t_1 which is the initial value for $w(\lambda_1|t)$ at the next step of the algorithm;
- (5) according to formula (9), for $k = 1$ compute the probability $w(\lambda_1|t)$ at time moment t , $t_1 < t < t_2$, where t_2 is the moment when the second event occurs in the flow;
- (6) for $k = 1$ compute the probability $w(\lambda_1|t_2) = w(\lambda_1|t_2 - 0)$ at time moment t_2 by following formula

$$w(\lambda_1|t_2) = \frac{w_1[1 - w(\lambda_1|t_1 + 0)] - [w_1 - w(\lambda_1|t_1 + 0)]e^{-(\lambda_1 - \lambda_2 + \alpha_1)(1 - w_1)(t_2 - t_1)}}{[1 - w(\lambda_1|t_1 + 0)] - [w_1 - w(\lambda_1|t_1 + 0)]e^{-(\lambda_1 - \lambda_2 + \alpha_1)(1 - w_1)(t_2 - t_1)}};$$

- (7) k increases by one and so on.

As we compute posterior probability $w(\lambda_1|t)$ we can make a decision regarding the state of process $\lambda(t)$ at any time moment t : if $w(\lambda_1|t) \geq w(\lambda_2|t)$ then we estimate $\hat{\lambda}(t) = \lambda_1$, otherwise $\hat{\lambda}(t) = \lambda_2$.

4 Results of Numerical Calculations

We developed the algorithm for computing the posterior probability to obtain numerical results. The algorithm consists of two stages. Also statistical experiments were made.

The simulation of the semi-synchronous event flow of the second order is performed directly at the first stage of the implementation algorithm. Posterior probabilities are calculated at the second stage based on the obtained sample of the moments of event occurrence of the flow, and also estimates of the considered event flow states are constructed.

Figure 2 shows an example of one of the realizations of the process $\lambda(t)$ and its estimation at the time interval of $T = 10$ units of time (simulation time) for the following set of parameters: $\lambda_1 = 2$, $\lambda_2 = 0,8$, $\alpha_1 = 2$, $\alpha_2 = 0,8$ and the probabilities $P_1^{(1)}(\lambda_1|\lambda_1) = P_1^{(2)}(\lambda_1|\lambda_1) = 0,4$, $P_1^{(1)}(\lambda_2|\lambda_1) = P_1^{(2)}(\lambda_2|\lambda_1) = 0,6$.

Figure 3 shows the trajectory of the posterior probability behavior $w(\lambda_1|t)$ for the case under consideration.

These statistical experiments, which consist of following steps, were made to find the frequency of making erroneous decisions about the state of the process $\lambda(t)$:

- (1) for the fixed set of parameters simulate the semi-synchronous event flow of the second order at time interval of T (i th experiment);
- (2) according to formulas (6), (9) compute the probability $w(\lambda_1|t)$ for the given time interval;
- (3) construct the estimation of the process $\lambda(t)$ at any time moment t for the given time interval;
- (4) compute the value d_i (for i th experiment) that is total length of time intervals, where estimation value $\hat{\lambda}(t)$ does not coincide with the actual value of the process $\lambda(t)$;

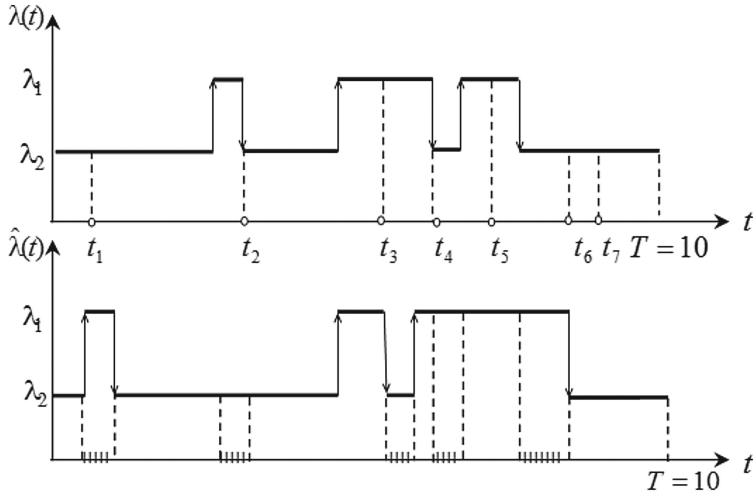


Fig. 2. The process $\lambda(t)$ trajectory and the estimation $\hat{\lambda}(t)$ trajectory

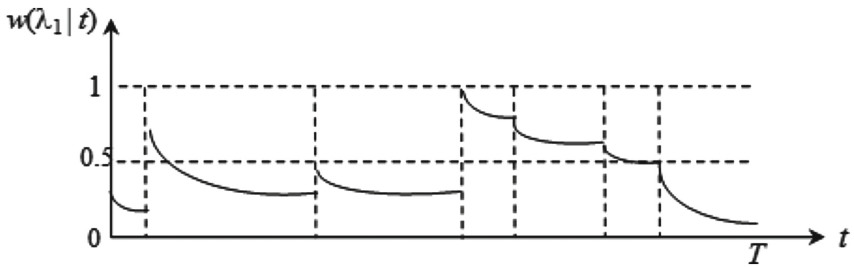


Fig. 3. Trajectory of the posterior probability $w(\lambda_1|t)$

- (5) compute the fraction of wrong decisions $\hat{p}_i = d_i/T$ (estimation of the total error probability of the state estimation of the process $\lambda(t)$ for the given time interval);
- (6) repeat 1-5 steps N times ($i = \overline{1, N}$).

The result of the above algorithm is a sample of fractions of wrong decisions in N experiments \hat{p}_i , $i = \overline{1, N}$, by which we find a sample mean of the total probability of the erroneous solution $\hat{P}_{er} = \frac{1}{N} \sum_{i=1}^N \hat{p}_i$ and a sample variance $\hat{D} = \frac{1}{N-1} \sum_{i=1}^N (\hat{p}_i - \hat{P}_{er})^2$.

The results of the first statistical experiment are given in Tables 1, 2, 3, 4 and 5.

We consider the dependence of \hat{P}_{er} , \hat{D} change for $N = 100$, probabilities $P_1^{(1)}(\lambda_1|\lambda_1) = P_1^{(2)}(\lambda_2|\lambda_1) = 0,4$, $P_1^{(1)}(\lambda_2|\lambda_1) = P_1^{(2)}(\lambda_1|\lambda_1) = 0,6$ and flow

Table 1. The results of the first statistical experiment for $\lambda_1 = 3$

T	100	300	500	700
\hat{P}_{er}	0,1901	0,2002	0,1899	0,1908
\hat{D}	0,0023	0,0021	0,0019	0,0020

Table 2. The results of the first statistical experiment for $\lambda_1 = 4$

T	100	300	500	700
\hat{P}_{er}	0,1700	0,1877	0,1777	0,1699
\hat{D}	0,0018	0,0017	0,0020	0,0018

Table 3. The results of the first statistical experiment for $\lambda_1 = 5$

T	100	300	500	700
\hat{P}_{er}	0,1666	0,1709	0,1670	0,1507
\hat{D}	0,0015	0,0019	0,0020	0,0017

Table 4. The results of the first statistical experiment for $\lambda_1 = 6$

T	100	300	500	700
\hat{P}_{er}	0,1309	0,1397	0,1299	0,1101
\hat{D}	0,0014	0,0010	0,0013	0,0011

Table 5. The results of the first statistical experiment for $\lambda_1 = 7$

T	100	300	500	700
\hat{P}_{er}	0,1001	0,0897	0,0721	0,0892
\hat{D}	0,0011	0,0009	0,0002	0,0010

parameters $\lambda_2 = 0,8$, $\alpha_1 = 2$, $\alpha_2 = 0,8$ for the values $\lambda_1 = 3, 4, 5, 6, 7$ and simulation time $T = 100, 300, 500, 700$ units of time in this experiment.

Analysis of the numerical results given in Tables 1, 2, 3, 4 and 5 shows that the estimation of the unconditional probability of an erroneous solution \hat{P}_{er} is sufficiently stable for the simulation time $T \geq 100$ units of time for all variants of computation, and it is also clear that the estimate \hat{P}_{er} decreases with increasing of the parameter λ_1 , which is quite normal due to better states distinguishability in this case. We note that the optimal estimation algorithm for the states of the semi-synchronous event flow of the second order provides a sufficiently acceptable estimate of the unconditional probability of an erroneous solution for these parameter values, and the sample variance of this estimate is small.

Table 6. The results of the second statistical experiment for $\lambda_2 = 0, 8, \alpha_1 = 2, \alpha_2 = 0, 8$

λ_1	2	4	6	8
\hat{P}_{er}	0,3009	0,2189	0,1387	0,0350
\hat{D}	0,0029	0,0021	0,0009	0,0008

Table 7. The results of the second statistical experiment for $\lambda_1 = 2, \lambda_2 = 0, 8, \alpha_2 = 0, 8$

α_1	2	4	6	8
\hat{P}_{er}	0,2961	0,1998	0,1197	0,0410
\hat{D}	0,0035	0,0024	0,0013	0,0013

Table 8. The results of the third statistical experiment for $\lambda_2 = 0, 1, \alpha_1 = \alpha_2 = 1$

λ_1	0,2	0,4	0,6	0,8
\hat{P}_{er}	0,1437	0,1206	0,1095	0,0629
\hat{D}	0,0015	0,0013	0,0008	0,0008

Table 9. The results of the third statistical experiment for $\lambda_1 = 1, \alpha_1 = \alpha_2 = 1$

λ_2	0,2	0,4	0,6	0,8
\hat{P}_{er}	0,0769	0,0992	0,1120	0,1307
\hat{D}	0,0008	0,0009	0,0013	0,0014

Table 10. The results of the third statistical experiment for $\lambda_1 = 1, \lambda_1 = 0, 1, \alpha_2 = 1$

α_1	0,2	0,4	0,6	0,8
\hat{P}_{er}	0,0499	0,0753	0,1096	0,1260
\hat{D}	0,0005	0,0008	0,0012	0,0018

We present the results of the second statistical experiment for the fixed simulation time $T = 100$ units of time and values $N = 100, P_1^{(1)}(\lambda_1|\lambda_1) = P_1^{(2)}(\lambda_1|\lambda_1) = 0, P_1^{(1)}(\lambda_2|\lambda_1) = P_1^{(2)}(\lambda_2|\lambda_1) = 1$ in Tables 6 and 7. The estimations \hat{P}_{er}, \hat{D} are obtained for $\lambda_2 = 0, 8, \alpha_1 = 2, \alpha_2 = 0, 8$ with parameter changing $\lambda_1 = 2, 4, 6, 8$ in Table 6; for $\lambda_1 = 2, \lambda_2 = 0, 8, \alpha_2 = 0, 8$ with parameter changing $\alpha_1 = 2, 4, 6, 8$ in Table 7.

Analyzing the numerical results obtained in Tables 6 and 7, we can do the following conclusions. The estimate \hat{P}_{er} decreases with increasing of the parameter λ_1 (Table 6) which is normal, since the conditions for the states distinguishability of the process $\lambda(t)$ improve; the estimate \hat{P}_{er} also decreases with increasing of the parameter α_1 (Table 7), since the conditions for the states distinguishability

of the process $\lambda(t)$ improve (the process $\lambda(t)$ is predominantly at the second state).

We present the results of the third statistical experiment for the fixed simulation time $T = 100$ units of time and values $N = 100$, $P_1^{(1)}(\lambda_1|\lambda_1) = P_1^{(2)}(\lambda_2|\lambda_1) = 0, 4$, $P_1^{(1)}(\lambda_2|\lambda_1) = P_1^{(2)}(\lambda_1|\lambda_1) = 0, 6$ in Tables 8, 9 and 10. The estimations \hat{P}_{er} , \hat{D} are obtained for $\lambda_2 = 0, 1$, $\alpha_1 = \alpha_2 = 1$ with parameter values $\lambda_1 = 0, 2, 0, 4, 0, 6, 0, 8$ in Table 8; $\lambda_1 = 1$, $\alpha_1 = \alpha_2 = 1$ with parameter values $\lambda_2 = 0, 2, 0, 4, 0, 6, 0, 8$ in Table 9; for $\lambda_1 = 1$, $\lambda_2 = 0, 1$, $\alpha_2 = 1$ with parameter values $\alpha_1 = 0, 2, 0, 4, 0, 6, 0, 8$ in Table 10.

Analyzing the numerical results obtained in Tables 8, 9 and 10, we can do the following conclusions. The estimate \hat{P}_{er} decreases with increasing of the parameter λ_1 (Table 8). The latter is due to the fact that the frequency of transitions from the first state to the second state of the process $\lambda(t)$ increases with increasing the parameter λ_1 , which has a positive effect for the states distinguishability conditions. The estimate \hat{P}_{er} increases with increasing of the parameter λ_2 (Table 9), which is also due to the convergence of the values λ_1 and λ_2 , which affects negatively on the states distinguishability conditions in this case. Increasing of the estimate \hat{P}_{er} is observed with increasing the parameter α_1 for its small values (Table 10). This is explained by the frequency of the change of states. Thus, we can conclude that the probability of error \hat{P}_{er} decreases for large values of the quantities $|\lambda_1 - \lambda_2|$ and $|\alpha_1 - \alpha_2|$ with constant other flow parameters.

5 Conclusion

We present formulas for the computation of posterior probabilities $w(\lambda_1|t)$, $w(\lambda_2|t)$ in this paper. The optimal estimation algorithm for the states of the semi-synchronous event flow of the second order has been developed on the basis of the formulas that was found, which provides a minimum of the total error probability of the decision making. The algorithm to compute posterior probabilities and the algorithm for estimating the flow states were implemented based on a sample of the moments of event occurrence obtained by simulating the flow under consideration. The algorithms were implemented by C# programming language in Visual Studio 2013. Statistical experiments were carried out, the numerical results of which do not contradict the physical interpretation.

References

1. Basharin, G.P., Kokotushkin, V.A., Naumov, V.A.: On the equivalent substitutions method for computing fragments of communication networks. *Izv. Akad. Nauk USSR. Tekhn. Kibern.* **6**, 92–99 (1979)
2. Neuts, M.F.: A versatile Markov point process. *J. Appl. Probab.* **16**, 764–779 (1979)
3. Gortsev, A.M., Nezhelskaya, L.A.: Estimation of the parameters of the synchronous double-stochastic event flow by the method of moments. *Tomsk State Univ. J. Control Comput. Sci.* **1**, 24–29 (2002)

4. Gortsev, A.M., Leonova, M.A., Nezhelskaya, L.A.: Joint probability density of the duration of intervals of the generalized asynchronous event flow with unextendable dead time. *Tomsk State Univ. J. Control Comput. Sci.* **21**(4), 14–25 (2012)
5. Gortsev, A.M., Nezhelskaya, L.A., Shevchenko, T.I.: Estimation of the states of an MC-stream of events in the presence of measurement errors. *Russ. Phys. J.* **36**(12), 1153–1167 (1993)
6. Nezhelskaya, L.A.: Optimal state estimation of the semi-synchronous event flow under conditions of its partial observability. *Tomsk State Univ. J. Control Comput. Sci.* **269**, 95–98 (2000)
7. Leonova, M.A., Nezhelskaya, L.A.: The probability of an error estimating the states of a generalized asynchronous event flow. *State Univ. J. Control Comput. Sci.* **2**, 88–101 (2012)
8. Gortsev, A.M., Kalyagin, A.A., Nezhelskaya, L.A.: Maximum likelihood estimation of dead time period in the generalized semi-synchronous event flow. *State Univ. J. Control Comput. Sci.* **1**, 27–37 (2015)
9. Gortsev, A.M., Kalyagin, A.A., Nezhelskaya, L.A.: Joint probability density of the duration of the intervals of the generalized semi-synchronous event flow with unextendable dead time. *State Univ. J. Control Comput. Sci.* **27**(2), 19–29 (2014)
10. Gortsev, A.M., Leonova, M.A., Nezhelskaya, L.A.: Comparison of MP and MM estimates of the dead time period in the generalized asynchronous event flow. *State Univ. J. Control Comput. Sci.* **25**(4), 32–42 (2013)
11. Dudin, A.N., Klimenok, V.I.: *Queueing Systems with Correlated Flows*. Belarus Gos. Univ, Minsk (2000)
12. Basharin, G.P., Gaidamaka, Y.V., Samouylov, K.E.: Mathematical theory of teletraffic and its application to the analysis of multiservice communication of next generation networks. *Autom. Control Comput. Sci.* **47**(2), 11–21 (2013)
13. Khazen, E.M.: *Methods of Optimal Statistical Decisions and Optimal Control*. Sovetskoe Radio, Moscow (1968)
14. Nezhelskaya, L.A., Tumashkina, D.A.: Simulation model of a semi-synchronous flow of the second order. In: *Proceedings of Tomsk State University, Series of Physics and Mathematics*, vol. 299, pp. 109–114 (2016)
15. Nezhelskaya, L.A., Tumashkina, D.A.: Optimal state estimation of a semi-synchronous event flow of the second order. In: *Proceedings of Tomsk State University, Series of physics and mathematics*, vol. 301, pp. 97–105 (2017)
16. Gortsev, A.M., Nezhelskaya, L.A.: Optimal nonlinear filtering of markov event flow with switching. *Technol. Commun. Means. Ser. Commun. Syst.* **7**, 46–54 (1989)



Modelling of Output Flows in Queuing Systems and Networks

Gurami Tsitsiashvili^{1,2}(✉) and Marina Osipova^{1,2}

¹ IAM FEB RAS, Vladivostok, Russia
guram@iam.dvo.ru, mao1975@list.ru

² Far Eastern Federal University, Vladivostok, Russia

Abstract. A simplification of Burke theorem proof [1] and its generalizations for queuing systems and networks are considered. The proof simplification is based on the fact that points in output flow take place in moments when Markov process of customers number in queuing system has jumps down. First steps in this direction were made in [2]. But here we improved proves of main results and consider queuing systems in random environment. In such way it is possible to obtain a property of the mutual independence of the flow into disjoint periods of time and to calculate intensity of output flow. In this case Poisson input flow with randomly varying intensity may be represented as Poisson flow with average intensity also. If this flow is independent with service process then it is possible to simplify significantly consideration of queuing systems in random environment. These assumptions may be applied to a consideration of multiphase type networks [3] which are convenient in analysis of queuing models with retrial queues [4–8].

Keywords: An output Poisson flow · The Jackson network
A random environment · A directed graph · A non-return set of nodes

1 Introduction

This paper is devoted to analysis of output flows in queuing systems and networks. In first part of the paper we consider simplification of Burke theorem proof [1] and its generalizations. The proof simplification is based on the fact that points in output flow take place in moments when Markov process of customers number in queuing system has jumps down. In such way it is possible to obtain a property of the mutual independence of the flow into disjoint periods of time. Then it is possible knowing the process of customers number distribution to calculate intensities of such jumps down and so to calculate intensity of output flow. This approach allows to obtain different corollaries for output flows in open and close queuing networks.

Such consideration may be applied not only to output flows but to input flows also. In this paper it is shown that for some stochastic models Poisson input flow with randomly varying intensity coincides by distribution with Poisson flow with

average intensity. This fact allows to analyse and to calculate distributions of processes in open queuing network with finite number of nodes, infinite number of servers in nodes, exponential distributions of service times and Poisson input flow with randomly varying intensity. A presence of infinite number of servers in the network nodes [4–6] together with the statement that the cardinality of counting set of counting sets is counting set also allows to transform initial queuing network into queuing network of multiphase type [3] so that in each node a customer may be served no more than once. A transformation of the Jackson network into the multiphase type network is closely connected with models of retrial queues [7, 8].

It is proved that all flows of so transformed network in stationary regime are Poisson. Synergetic effects in this network are analysed using a replacement of infinite number of servers by finite number of them. Synergetic effect means that if number of servers in nodes and intensity of input flow increase in $n \rightarrow \infty$ times then probability of queues existence on finite time interval tends to zero.

This investigation is based on the Burke theorem [1] that in stationary regime output flow of multiserver system $M|M|n|\infty$ is Poisson.

2 New Proof of Burke Theorem and Its Corollaries

In [1] the following statement is proved: in queuing system $M|M|n|\infty$ in stationary state, the output flow has the same distribution as the input flow. Recently, however, interest in the study of flows in queuing systems is increased. Now it is necessary to give a more compact and convenient for generalizations proof of this theorem.

A random sequence of points will be called a Poisson flow with continuously differentiable intensity $\lambda(t)$, $t \geq 0$, if the following conditions are satisfied [9, p. 12, 13], [10, p. 20, 35 – 37]:

- (a) the probability of the existence of the point of flow on the time interval $[t, t + h)$ does not depend on the location of the points of the flow up to the time t (this property is called lack of follow-through and expresses the mutual independence of the flow into disjoint periods of time);
- (b) the probability that a flow point appears in the semi-interval $[t, t + h)$ is $\lambda(t)h + o(h)$, $h \rightarrow 0$;
- (c) the probability of occurrence of two or more flow points in the range $[t, t + h)$ is $o(h)$, $h \rightarrow 0$.

Let the system $A_n = M|M|n|\infty$ of the Poisson input flow has an intensity $\lambda > 0$, and the service time has an exponential distribution with the parameter $\mu > 0$, $1 \leq n < \infty$. Denote $P_{k,n}(t)$, $k \geq 0$, distribution of the number of customers in the system at the time t .

Theorem 1. *The output flow in queuing system A_n is Poisson with intensity*

$$a(t) = \sum_{0 < k} \mu P_{k,n}(t) \min(k, n).$$

Proof. Let the output flow $T_n = \{0 \leq t_1 < t_2 < \dots\}$ be A_n described by a random function $y_n(t)$ equal to the number of points of this flow on the segment $[0, t)$. Denote $x_n(t)$ the number of customers in the system A_n at the time t . It is known that a random process $x_n(t)$ is Markov process (of death and birth of [10, Chap. II, Sect. 1]), with each point of the T_n flow corresponding to the time of the jump down process $x_n(t)$. Therefore, the output flow T_n satisfies the condition (a). In turn, the condition (b) follows from the equalities:

$$\begin{aligned} P(y_n(t+h) = y_n(t) + 1) &= \sum_{k=1}^n P(y_n(t+h) = y_n(t) + 1/x_n(t) = k)P_{k,n}(t) \\ &\quad + P(y_n(t+h) = y_n(t) + 1/x_n(t) > n) \sum_{k>n} P_{k,n}(t) \\ &= \sum_{k=1}^n P_{k,n}(t)(k\mu h + o_k(h)) + \sum_{k>n} P_{k,n}(t)(n\mu h + o_0(h)) = a(t)\mu h + o(h), \end{aligned}$$

where for $h \rightarrow 0$ we have $\frac{o_k(h)}{h} \rightarrow 0, k = 0, \dots, n,$

$$o(h) = \sum_{k=1}^n P_{k,n}(t)o_k(h) + \sum_{k>n} P_{k,n}(t)o_0(h), \quad o_0(h)/h \rightarrow 0.$$

Thus, the output flow T_n satisfies the condition (b). Check of condition (c) is quite obvious.

Theorem 2. *In queuing system A_n , when the ergodicity condition $\lambda < \mu$ is satisfied and the process $x_n(t)$ is stationary, the output flow is Poisson with intensity λ .*

Proof. Denote $P_{k,n}, k = 0, 1, \dots,$ stationary probabilities of ergodic process $x_n(t)$. The system of Kolmogorov-Chapman equalities for $P_{k,n}, k = 0, 1, \dots,$ is following:

$$0 = -P_{0,n}\lambda + P_{1,n}\mu_1, \quad 0 = -P_{k,n}(\lambda + \mu_k) + P_{k-1,n}\lambda + P_{k+1,n}\mu_{k+1}, \quad (1)$$

with $\mu_k = \min(k, n)\mu, k = 1, 2, \dots$. Prove by an induction that from Formulas (1) we have

$$0 = -P_{k,n}\lambda + P_{k+1,n}\mu_{k+1}, \quad k = 0, 1, \dots \quad (2)$$

Indeed for $k = 0$ this statement is a corollary of the first equation in Formulas (1). Assume that the equality (2) is true for $k = i$, then from equations in (1) and induction assumption we have for $k = i + 1$:

$$\begin{aligned} 0 &= [-P_{i+1,n}(\lambda + \mu_{i+1}) + P_{i,n}\lambda + P_{i+2,n}\mu_{i+2}] + [-P_{i,n}\lambda + P_{i+1,n}\mu_{i+1}] \\ &= -P_{i+1,n}\lambda + P_{i+2,n}\mu_{i+2}. \end{aligned}$$

Consequently the equations (2) are true for all $k = 0, 1, \dots$. Summarize equalities (2) by $k = 0, 1, \dots$, we obtain

$$\lambda = \sum_{k \geq 0} P_{k+1,n} \mu_{k+1}. \tag{3}$$

So from Theorem 1 we have the statement of Theorem 2.

Remark 1. Using the scheme of the proof of Theorem 2, it is possible to extend the results to output flows of systems with limited queue, with priority service, with unreliable servers [10, Sect. 7].

3 Poisson Flows in Stationary Queuing Networks

Consider an open queuing network (Jackson network [11]) S with a Poisson input flow of intensity λ_0 , consisting of a finite number of nodes $k = 0, 1, \dots, m$ with exponentially distributed service times. The dynamics of the movement of customers in the network is set by the route matrix $\Theta = \|\theta_{i,j}\|_{i,j=0}^m$, where $\theta_{i,j}$ is the probability of customer transition after service in the i -th node to j -th node, $\theta_{0,0} = 0$, where the node 0 is an external source and at the same time a drain for customers leaving the network. The i node contains $l_i < \infty$ servers, the service time of which has an exponential distribution with the parameter μ_i , $i = 1, \dots, m$.

Assume that route matrix $\Theta = \|\theta_{i,j}\|_{i,j=0}^m$ is indecomposable, i.e.

$$\forall i, j \in \{0, \dots, m\} \exists i_1, \dots, i_r \in \{0, \dots, m\} : \theta_{i,i_1} > 0, \theta_{i_1,i_2} > 0, \dots, \theta_{i_r,j} > 0.$$

Then for a fixed $\lambda_0 > 0$, the system of linear algebraic equations for intensities of fluxes coming from nodes of S

$$\lambda_k = \lambda_0 \theta_{0,k} + \sum_{t=1}^m \lambda_t \theta_{t,k}, \quad k = 1, \dots, m \tag{4}$$

has the only solution $(\lambda_1, \dots, \lambda_m)$ $\lambda_1 > 0, \dots, \lambda_m > 0$, [12, p. 13].

The system (4) is called the system of balance relations and plays an important role in the formulation and the proof of the product Jackson theorem [11], widely used in queuing theory. If

$$\lambda_i < l_i \mu_i, \quad i = 1, \dots, m,$$

then the discrete Markov process $(n_1(t), \dots, n_m(t))$, $t \geq 0$, describing the number of customers in the network nodes has a limiting distribution $P_S(k_1, \dots, k_m)$, independent of initial conditions and representable in the form

$$P_S(k_1, \dots, k_m) = \prod_{i=1}^m P_i(k_i),$$

where $P_i(k_i)$ is the limiting distribution of the number of customers in a stand-alone l_i -channel queuing system with Poisson input flow of intensity λ_i , $i = 1, \dots, m$.

In [13] network S is mapped to a directed graph G with edges corresponding to positive elements of the route matrix. Let's call the vertex set $U \subseteq \{0, 1, \dots, m\}$ irrevocable if from any node not included in U , there is no edge to the node belonging to U . Then all flows passing through the edges from the node set U to the node set $\{0, 1, \dots, m\} \setminus U$, are independent and Poisson.

Theorem 3. *Flow T_S^i , $i = 1, \dots, m$, coming out of node i of open queuing network S , with stationary process $(n_1(t), \dots, n_m(t))$, $t \geq 0$, is Poisson with intensity λ_i .*

Proof. Indeed, the points of the flow T_S^i , exiting the i , node are the moments of jumps down the $n_i(t)$ component of the discrete Markov process $(n_1(t), \dots, n_m(t))$, $t \geq 0$. Hence the flow T_S^i satisfies the condition (a). Conditions (b), (c) are checked similarly to the proof of Theorem 1. Note that the limit probability that the i node contains k_i of customers is $P_i(k_i)$, and the flow rate T_S^i is λ_i , $i = 1, \dots, m$.

Theorem 4. *Flows T_S^i , $i = 1, \dots, m$, are independent.*

Proof. From Theorem 3 and independence of stationary random variables $n_j(t)$, $j = 1, \dots, m$, it follows that the union

$$T_S = \bigcup_{j=1}^m t_S^j$$

of flows leaving the nodes of open queuing network S is also Poisson flow with intensity $\lambda_\Sigma = \sum_{j=1}^m \lambda_j$. And each point of the combined flow T_S belongs to the flow T_S^i with probability $\frac{\lambda_i}{\lambda_\Sigma}$. Hence the flows T_S^i , $i = 1, \dots, m$, are independent.

Remark 2. Theorems 3, 4 enhance the results of the article [13], removing restrictions on the independent Poisson flows considered in it.

Consider now a closed queueing network \bar{S} , consisting of a finite number of nodes $i = 1, \dots, m$. The i node contains $l_i < \infty$ servers, the service time on which has an exponential distribution with the parameter μ_i , $i = 1, \dots, m$. A finite number N of customers move along network \bar{S} . The dynamics of the customers movement in the network is specified by the matrix $\bar{\Theta} = \|\bar{\theta}_{i,j}\|_{i,j=1}^m$, where $\bar{\theta}_{i,j}$ is the probability of transition after service of customer in the i th node to j -th one.

Let the route matrix $\bar{\Theta}$ be indecomposable, i.e.

$$\forall i, j \in \{1, \dots, m\} \exists i_1, \dots, i_r \in \{1, \dots, m\} : \bar{\theta}_{i,i_1} > 0, \bar{\theta}_{i_1,i_2} > 0, \dots, \bar{\theta}_{i_r,j} > 0.$$

Then for a fixed $\lambda_1 > 0$, the system of linear algebraic equations

$$\lambda_k = \sum_{t=1}^m \lambda_t \bar{\theta}_{t,k}, \quad k = 1, \dots, m \tag{5}$$

has a unique solution of $(\lambda_1, \dots, \lambda_m)$ with $\lambda_1 > 0, \dots, \lambda_m > 0$, [12, p. 13].

For a closed queueing network \bar{S} with N customers discrete Markov process $(\bar{n}_1(t), \dots, \bar{n}_m(t))$, $t \geq 0$, describing the number of customers in the network nodes has a limit distribution of $P_{\bar{S}}(k_1, \dots, k_m)$, independent of the initial conditions and presented in the form

$$P_{\bar{S}}(k_1, \dots, k_m) = \frac{\prod_{i=1}^m P_i(k_i)}{\sum_{k_1, \dots, k_m: k_1 + \dots + k_m = N} \prod_{i=1}^m P_i(k_i)}, \quad k_1 + \dots + k_m = N.$$

Hence, the stationary probability $\pi_i(k_i)$ that in a node i of the network \bar{S} there is k_i customers satisfies the equality

$$\pi_i(k_i) = \sum_{k_j, 1 \leq j \neq i \leq m, \sum_{1 \leq j \neq i \leq m} k_j = N - k_i} P_{\bar{S}}(k_1, \dots, k_m), \quad k_i = 0, \dots, N.$$

Theorem 5. *The flow $T_{\bar{S}}^i$, leaving the i node of the closed queueing network \bar{S} with the total number of customers N , being in a stationary state, is Poisson with intensity $\sum_{k_i=1}^N \min(k_i, l_i) \mu_i \pi_i(k_i)$, $i = 1, \dots, m$.*

Proof. Indeed, the points of the flow $T_{\bar{S}}^i$, exiting the node i , are the moments of jumps down the component $\bar{n}_i(t)$ of the discrete Markov process $(\bar{n}_1(t), \dots, \bar{n}_m(t))$, $t \geq 0$. Consequently, the flow $T_{\bar{S}}^i$ satisfies condition (a). Conditions (b), (c) are proved similarly to the proof of Theorem 1.

4 Queuing System $M|M|1|_{\infty}$ with Random Intensities of Input Flow and Service

Consider queuing system $A_1 = M|M|1|_{\infty}$ with a service intensity of $\mu(t)$ and a Poisson input flow Λ with an intensity of $\lambda(t)$, which are randomly changed by the following rules. Let the time axis $t \geq 0$ be split into half-intervals

$$[T_0 = 0, T_1 = T_0 + \xi_1), [T_1, T_2 = T_1 + \xi_2), \dots,$$

where ξ_1, ξ_2, \dots are independent random variables with distribution

$$P(\xi_k > t) = \exp(-\sigma t), \quad t \geq 0, \quad k = 1, 2, \dots$$

with parameter $\sigma > 0$.

We introduce a discrete Markov chain $n_l, l = 1, \dots$, with a set of states $\{1, \dots, N\}$ and an irreducible transition matrix $\|\theta_{i,j}\|_{i,j=1}^N$. Markov chain $n_l, l = 1, \dots$, has a unique (with positive components) solution (ψ_1, \dots, ψ_N) of the system of Kolmogorov-Chapman stationary equations

$$\psi_i = \sum_{j=1}^N \psi_j \theta_{j,i}, \quad i = 1, \dots, N. \tag{6}$$

We now introduce the Markov process $n(t), t \geq 0$, such that $n(t) = n_l, t \in [T_{l-1}, T_l), l = 1, \dots$. It is obvious that the stationary distribution (ψ_1, \dots, ψ_N) of the Markov chain $n_l, l = 1, \dots$, is a stationary distribution of the Markov process $n(t), t \geq 0$. Indeed denote

$$\psi_i(t) = p(n(t) = i), \quad i = 1, \dots, N, \tag{7}$$

then the Kolmogorov-Chapman system of equations for Markov process $n(t)$ has the form

$$\dot{\psi}_i(t) = -\sigma \psi_i(t) + \sigma \sum_{j=1}^N \psi_j(t) \theta_{j,i}, \quad i = 1, \dots, n,$$

so the system of Kolmogorov-Chapman stationary equations coincides with (6). Let's call such a queuing system as $M|M|1|\infty$ in a random environment.

Suppose that on each half-interval $[T_{k-1}, T_k)$ the input flow to the $M|M|1|\infty$ system is Poisson with intensity $\lambda(t) = \lambda_{n_l}, t \in [T_{l-1}, T_l), l = 1, 2, \dots$, and the service intensity satisfies the relations $\mu(t) = \mu_{n_l}, t \in [T_{l-1}, T_l), l = 1, 2, \dots$, where $\lambda_1, \dots, \lambda_N, \mu_1, \dots, \mu_N$ are some positive numbers. It is worthy to remark that in this system the input flow and the process of service (random sequence of service times) are dependent random objects.

Theorem 6. *The stationary output flow in the system $M|M|1|\infty$ in a random environment is Poisson with an average intensity $a = \sum_{j=1}^N \psi_j \lambda_j$.*

Proof. Consider Markov random process $(x(t), n(t)), t \geq 0$, whose first component sets the number of customers in the system $M|M|1|\infty$ and write for its stationary probabilities $p_{i,j}$ Kolmogorov-Chapman equations:

$$\lambda_i p_{0,i} = -\sigma p_{0,i} + \mu_i p_{1,i} + \sigma \sum_{j=1}^N p_{0,j} \theta_{j,i}, \quad i = 1, \dots, N,$$

$$(\lambda_i + \mu_i + \sigma) p_{k,i} = \lambda_i p_{k-1,i} + \mu_i p_{k+1,i} + \sigma \sum_{j=1}^N p_{k,j} \theta_{j,i}, \quad i = 1, \dots, N, \quad k = 1, 2, \dots \tag{8}$$

We introduce the following notation at $i = 1, \dots, N$:

$$A_{0,i} = -\lambda_i p_{0,i} + \mu_i p_{1,i}, \quad A_{k,i} = -(\lambda_i + \mu_i) p_{k,i} + \lambda_i p_{k-1,i} + \mu_i p_{k+1,i}, \quad k = 1, 2, \dots,$$

$$B_{k,i} = -\sigma p_{k,i} + \sigma \sum_{j=1}^N p_{k,j} \theta_{j,i}, \quad k = 0, 1, \dots$$

Then the equations (8) may be rewritten as

$$0 = A_{k,i} + B_{k,i}, \quad i = 1, \dots, N, \quad k = 0, 1, \dots \quad (9)$$

Denote $C_{k,i} = \sum_{r=0}^k A_{r,i}$, $D_{k,i} = \sum_{r=0}^k B_{r,i}$, then by Formulas (9) we have

$$0 = C_{k,i} + D_{k,i}, \quad i = 1, \dots, N, \quad k = 0, 1, \dots \quad (10)$$

Obviously, the following relations are fulfilled:

$$\frac{1}{\sigma} \sum_{i=1}^N B_{k,i} = - \sum_{i=1}^N p_{k,i} + \sum_{i=1}^N \sum_{j=1}^N p_{k,j} \theta_{j,i} = - \sum_{i=1}^N p_{k,i} + \sum_{j=1}^N \sum_{i=1}^N p_{k,j} \theta_{j,i} = 0.$$

and consequently

$$\sum_{i=1}^N D_{k,i} = 0. \quad (11)$$

By induction of k we can obtain equalities by analogy with Theorem 2 proof:

$$C_{k,i} = -\lambda_i p_{k,i} + \mu_i p_{k+1,i}, \quad i = 1, \dots, N, \quad k = 0, 1, \dots \quad (12)$$

Summing up the equations (10) by $i = 1, \dots, N$, $k = 0, 1, \dots$, and using Formulas (11), (12), we obtain:

$$0 = - \sum_{i=1}^N \lambda_i \sum_{k=0}^{\infty} p_{k,i} + \sum_{i=1}^N \mu_i \sum_{k=0}^{\infty} p_{k+1,i}. \quad (13)$$

The second term in Formula (13) is the intensity of a of the output Poisson flow in a given queuing system. In turn, by virtue of formulas (7), (13) we obtain that the intensity

$$a = \sum_{j=1}^N \psi_j \lambda_j. \quad (14)$$

Remark 3. By methods of Theorem 1 proof it is easy to obtain that the flow Λ is Poisson with intensity $a = \sum_{j=1}^N \psi_j \lambda_j$. Indeed, let us consider the Markov process $(y(t), n(t))$, $t \geq 0$, where $y(t)$ is the number of customers of the input flow that came to the system up to t . This process has the following transient intensities: the transition intensity $(m, i) \rightarrow (m, j)$ equals $\sigma \theta_{i,j}$, the intensity

of the transition $(m, i) \rightarrow (m + 1, i)$ equals to λ_i , $i, j = 1, \dots, N$, $m = 0, 1, \dots$. So the jump intensity $y(t) \rightarrow y(t) + 1$ equals

$$\sum_{1 \leq j \leq N, 0 \leq m} p(y(t) = m, n(t) = j) \lambda_j = \sum_{j=1}^N p(n(t) = j) \lambda_j = \sum_{j=1}^N \psi_j \lambda_j = a.$$

Thus, the random flow Λ by distribution coincides with the Poisson flow of average intensity a .

Remark 4. The statement of Remark 3 allows to obtain criterion's of ergodicity, to derive formulas for stationary distributions, to analyse output flows for manifold queuing systems with independent input flow Λ and sequences of service times: open queuing network of Jackson type, queuing systems with failures, queuing systems with feedbacks [4].

5 Transformation of Open Queuing Network into Multiphase Type Queuing Network

Following [3] demonstrate how to transform open queuing network into multiphase type queuing network. Consider open queuing network S with finite number of nodes $U = \{0, 1, \dots, m\}$ and input flow Λ . As the flow Λ and service times of customers in different nodes are independent then it is convenient to consider the flow Λ as Poisson flow with average intensity $\lambda_0 = a$. Paths of customers in the network S are defined by the route matrix $\Theta = \|\theta_{i,j}\|_{i,j=0}^m$, $\theta_{0,0} = 0$, consisting of probabilities $\theta_{i,j}$ of customers transitions from the node i to the node j after a service in the node i . The node 0 is a source of customers arriving the network and a container of customers departing the network. Here $\theta_{0,i}$ is the probability that input flow customer moves to the node i and $\theta_{i,0}$ is the probability that customer departs network after service in the node i . In the node k of the network S there is infinite number of identical servers with service times which has the distribution

$$F_k(t) = 1 - \exp(-\mu_k t), \quad t \geq 0, \quad \mu_k, \quad 0 < \mu_k < \infty, \quad k = 1, \dots, m.$$

Transform the network S into the following network S^* . Each node k , $0 \leq k \leq m$, is divided into infinite number of nodes (k, j) , $1 \leq j$. Here nodes with $1 \leq k \leq m$ are nodes with infinite numbers of servers and nodes with $k = 0$ absorb customers departing the network. A customer arriving the network with the probability $\theta_{0,k}$ moves to the node $(k, 1)$. The node $(0, 1)$ is sham because $\theta_{0,0} = 0$ and so customers do not visit it. Then after a service in the node (p, j) , $1 \leq p \leq m$, $1 \leq j$, customer with the probability $\theta_{p,q}$ moves to the node $(q, j + 1)$ and with the probability $\theta_{p,0}$ moves to the node $(0, j + 1)$ - departs the network, $1 \leq p, q \leq m$, $1 \leq j$. Consequently initial network S is transformed into the network S^* with the nodes set $U^* = \{(k, j), 1 \leq j, 0 \leq k \leq m\}$. Graphically the network S^* is represented in Fig. 1.

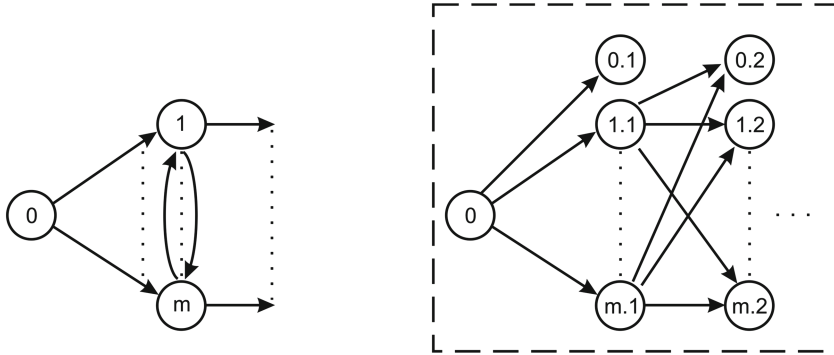


Fig. 1. Transformation of Jackson network (leftward) into multiphase type network (rightward).

The network S^* is constructed similar to retrial queues systems [4–8]. Transformation of the network S into the network S^* does not change paths and service times of customers.

In the network S^* a system of balance equations for stationary intensities of flows arriving the nodes of the set U^* may be solved by recurrent relations

$$\lambda_{k,1} = \lambda_0 \theta_{0,k}, \quad \lambda_{k,j+1} = \sum_{p=1}^m \lambda_{p,j} \theta_{p,k}, \quad 0 \leq k \leq m, \quad 1 \leq j, \quad (15)$$

and its synergetic effects may be analysed in suggestion that each node with infinite number of servers in multiphase type network is replaced by node with large by finite number of servers.

6 Conclusion

It is worthy to devote special attention to an application of Remark 4 to queuing systems and networks with retrial queues. Such systems appear in manifold modern applied problems [4–8]. In this section we connect a representation of the input flow Λ as Poisson flow with average intensity and a consideration of networks with infinite number of servers in their nodes [4–6]. For this purpose we use a transformation of such networks into multiphase type networks [3]. In multiphase type networks it is possible to assume that each customer may be serviced a fixed number of times also not arbitrary ones. This suggestion together with the representation of the input flow Λ as Poisson flow with average intensity and with an assumption that the flow Λ and service process are independent allow to consider models more close to applications.

Acknowledgment. This paper is partially supported by Russian Fund for Basic Researches, project 17-07-00177.

References

1. Burke, P.J.: The output of a queuing system. *Oper. Res.* **4**, 699–704 (1956)
2. Tsitsiashvili, G.S., Osipova, M.A.: Generalization and extension of Burke theorem. *Reliab.: Theor. Appl.* **13**(1), 59–62 (2018)
3. Tsitsiashvili, G.S., Osipova, M.A., Losev, A.S., Kharchenko, Y.N.: Jackson network as network of multiphase type. *Int. Math. Forum* **12**(7), 303–310 (2017)
4. Moiseeva, S., Zadiranova, L.: Feedback in infinite-server queuing systems. In: Vishnevsky, V., Kozyrev, D. (eds.) *DCCN 2015. CCIS*, vol. 601, pp. 370–377. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30843-2_38
5. Moiseeva, S.P., Zakhoroľ'naya, I.A.: Athemathical model of parallel retrial queueing of multiple requests. *Optoelectron. Instrum. Data Process.* **47**(6), 567–572 (2011)
6. Moiseev, A., Nazarov, A.: Queueing network $MAP/(GI/\infty)^K$ with high-rate arrivals. *Eur. J. Oper. Res.* **254**, 161–168 (2016)
7. Nobel, R.D., Tijms, H.C.: Optimal control for an MX/G/1 queue with two service modes. *Eur. J. Oper. Res.* **113**(3), 610–619 (1999)
8. Nobel, R.D., Tijms, H.C.: Waiting-time probabilities in the “M/G/1” retrial queue. *Stat. Neerl.* **60**(1), 73–78 (2006)
9. Khinchin, A.Y.: *Works on the Mathematical Queueing Theory*. Physmatlit, Moscow (1963)
10. Ivchenko, G.I., Kashtanov, V.A., Kovalenko, I.N.: *Queueing Theory*. Vishaya Shkola, Moscow (1982)
11. Jackson, J.R.: Networks of Waiting Lines. *Oper. Res.* **5**(4), 518–521 (1957)
12. Basharin, G.P., Tolmachev, A.L.: Queueing network theory and its applications to the analysis of information-computing networks. *Itogi Nauki Tech. Ser. Teor. Veroiatn. Mat. Stat. Teor. Kibern.* **21**, 3–119 (1983). (in Russian)
13. Beutler, F.J., Melamed, B.: Decomposition and customer streams of feedback networks of queues in equilibrium. *Oper. Res.* **26**(6), 1059–1072 (1978)



A Multi-server Queueing System with Backup Servers

Valentina Klimenok, Alexander Dudin^(✉), and Uladzimir Shumchenia

Department of Applied Mathematics and Computer Science, Belarusian State
University, 220030 Minsk, Belarus
{klimenok,dudin}@bsu.by, uladzimir.shumchenia@gmail.com

Abstract. We consider a multi-server queueing system that can be useful for solving the problem of reaching a trade-off between energy saving considerations and quality of customer's service by the use of so-called backup servers. A backup server joins to the service of a customer in case the timer installed at the service beginning moment on the main server expires. Such service organization allows to avoid too much delays in the system in conditions of reasonable energy savings. The system under consideration can also be considered as a model of an unreliable system where in the case of a failure of a main server a customer is serviced by a back-up server. In this case, the time set on the timer is interpreted as the time before the breaking-down of a main server. The behavior of the system is described by two-dimensional continuous time Markov chain which is successfully analysed in this paper.

Keywords: Multi-server queueing system · Backup servers
Stationary performance measures · sojourn time

1 Introduction

The problems associated with energy saving in many real systems, in particular, in data processing centres for cloud computing, can be solved by redundancy, with further adaptive connection of backup servers. In systems with heterogeneous information, for example, in call centers, backup servers assigned for priority information can greatly improve the quality of service. In unreliable data transmission systems the availability of redundant channels allows to improve transmission quality. Due to the stochastic nature of processing and transmission of information, mathematical modeling of systems with redundancy within the framework of queueing theory is actual. We mention only some publications in this field.

Papers [1–3] are devoted to tandem queueing systems with reserved channels for priority customers at the second station. These systems can be used i.e. for modeling call centers where customers of different types and different degrees of importance get service. In the papers [4–7], mathematical models of hybrid

communication systems consisting of unreliable Free Space Optics channel and backup reliable radio channel were investigated.

The authors of papers [8–10] consider the problem energy saving in the data center under an acceptable level of customer service defined by service level agreements. In the papers [9] and [10], two classes of servers are considered: main servers and backup servers. When the number of customers in the system increases to some nonnegative integer (threshold), the backup servers connect to the service of customers. When the number of customers decreases to some other threshold, the backup servers gets back to the standby. The time required to switch on the reserve block is taken into account. In the system discussed in the paper [8], there are several reserve blocks and there is instantaneous switching-on. To activate and deactivate reserve blocks the authors use the multi-threshold policy. The author come to the conclusion that the benefit of using multiple blocks of backup servers instead of a single block is negligible. A model similar to [9] and [10] (but with momentary switching-on of the block of backup servers) was considered in [11].

In this paper, we consider a multi-server queuing system that can be also used to solve the problem of finding a trad-off between energy saving and quality of service by using so-called backup servers. A backup server joins to the service of a customer in case customers's service time on the main servers exceeds some limit defined by a random value. In this case, we will say that the timer has expired. Such service organization allows to avoid too much delays in the system in conditions of reasonable energy savings. The system under consideration can also used for modeling an unreliable queue where in the case of a breaking-down of a main server a customer is serviced by a back-up server. In this case, the limiting time is interpreted as the time before the breaking-down of a main server. The analogous queue was considered in [13] under an assumption that the input flow is described by a Batch Markovian Arrival Process and service time has a Phase type distribution. Here we suppose that an arrival flow is the stationary Poisson one and service times are exponentially distributed. In this partial case the results have more simple and tractable form. At the same time, these results do not obviously follow from the results of [13]. This was our motivation for a separate consideration of the system with backup servers in case of exponential distributions.

2 Model Description

We consider a multi-server queueing system with an infinite buffer and two class of servers: N identical independent main servers and R identical independent backup (reserve) servers, $1 \leq R \leq N$. Customers arrive into the system in the stationary Poisson flow with the intensity λ .

The service time of a customer by a main server is exponentially distributed with the parameter μ .

An arriving customer occupies an idle main server, if any. Otherwise, the customer goes to the end of the queue and selected for service in accordance with

the FIFO discipline. At the time when a customer occupies the main server, a timer is set on this server. If the service time of the customer by the main server is not over and the timer has expired, an idle backup server, if any, joins to the service of the customer. The timer is defined as exponentially distributed random value with parameter τ . From the moment of joining the backup server, the residual service time has exponential distribution with parameter $\tilde{\mu}$.

If at the epoch of the timer expiration the service of a customer does not finish and all backup servers busy, then with probability p the customer leaves the system forever and with the probability $1 - p$ the timer on this server is set again and the service continues, $0 \leq p \leq 1$.

3 Process of the System States

Let, at time t ,

- i_t be the number of customers in the system, $i_t \geq 0$;
- r_t be the number of busy backup servers, $r_t = 0, 1, \dots, \min\{i_t, R\}$.

The operation of the queue under consideration is described by a regular irreducible continuous-time Markov chain $\xi_t = \{i_t, r_t\}$, $t \geq 0$, with state space

$$\Omega = \{(0)\} \cup \{(i, r), i > 0, r = 0, 1, \dots, \min\{i, R\}\}.$$

In the following, we will assume that the states of the chain ξ_t , $t \geq 0$, are enumerated in the lexicographic order.

Lemma 1. *Infinitesimal generator Q of the Markov chain ξ_t , $t \geq 0$, has the following three-diagonal block structure:*

$$Q = \begin{pmatrix} \mathcal{H}_0^{(0,0)} & \mathcal{H}_0^{(0,1)} & O & O & O \cdots & O & O & O & O \cdots \\ \mathcal{H}_0^{(1,0)} & \mathcal{H}_0^{(1,1)} & \mathcal{H}_0^{(1,2)} & O & O \cdots & O & O & O & O \cdots \\ O & \mathcal{H}_0^{(2,1)} & \mathcal{H}_0^{(2,2)} & \mathcal{H}_0^{(2,3)} & O \cdots & O & O & O & O \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \cdots \\ O & O & O & O & O \cdots & \mathcal{H}_0^{(N-1,N)} & O & O & O \cdots \\ O & O & O & O & O \cdots & \mathcal{H}_0^{(N,N)} & H_1^{(N)} & O & O \cdots \\ O & O & O & O & O \cdots & Q_{-1} & Q_0 & Q_1 & O \cdots \\ O & O & O & O & O \cdots & O & Q_{-1} & Q_0 & Q_1 \cdots \\ O & O & O & O & O \cdots & O & O & Q_{-1} & Q_0 \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where

$$\mathcal{H}_0^{(i,i-1)} = \begin{pmatrix} i\mu & 0 & \dots & 0 & 0 \\ \tilde{\mu} & (i-1)\mu & \dots & 0 & 0 \\ 0 & 2\tilde{\mu} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & (i-1)\tilde{\mu} & \mu \\ 0 & 0 & \dots & 0 & i\tilde{\mu} \end{pmatrix}, \quad 0 \leq i \leq R,$$

$$\mathcal{H}_0^{(i,i)} = -\lambda I$$

$$- \begin{pmatrix} i(\mu + \tau) & i\tau & \dots & 0 & 0 \\ 0 & (i-1)(\mu + \tau) + \tilde{\mu} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & (\mu + \tau) + (i-1)\tilde{\mu} & \tau \\ 0 & 0 & \dots & 0 & i\tilde{\mu} \end{pmatrix},$$

$$0 \leq i \leq R,$$

$$\mathcal{H}_0^{(i,i+1)} = (\lambda I_{i+1} \mid O_{(i+1) \times \min\{1, R-i\}}), \quad 0 \leq i \leq R,$$

$$\mathcal{H}_0^{(i,i-1)} = \begin{pmatrix} i\mu & 0 & \dots & 0 & 0 \\ \tilde{\mu} & (i-1)\mu & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & (i-R+1)\mu & 0 \\ 0 & 0 & \dots & R\tilde{\mu} & (i-R)(\mu + p\tau) \end{pmatrix}, \quad R < i \leq N,$$

$$\mathcal{H}_0^{(i,i)} = -\lambda I$$

$$- \begin{pmatrix} i(\mu + \tau) & i\tau & \dots & 0 & 0 \\ 0 & (i-1)(\mu + \tau) + \tilde{\mu} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & (i-R+1)(\mu + \tau) & (i-R+1)\tau \\ 0 & 0 & \dots & 0 & (i-R)(\mu + p\tau) \end{pmatrix},$$

$$R < i \leq N,$$

$$\mathcal{H}_0^{(i,i+1)} = \lambda I_{R+1}, \quad R < i < N.$$

$$H_1^{(N)} = \lambda I_{R+1},$$

$$Q_{-1} = \begin{pmatrix} N\mu & 0 & \dots & 0 & 0 \\ \tilde{\mu} & (N-1)\mu & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & (N-R+1)\mu & 0 \\ 0 & 0 & \dots & R\tilde{\mu} & (N-R)(\mu + p\tau) \end{pmatrix},$$

$$Q_0 = -\lambda I$$

$$- \begin{pmatrix} N(\mu + \tau) & N\tau & \dots & 0 & 0 \\ 0 & (N-1)(\mu + \tau) + \tilde{\mu} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & (N-R+1)(\mu + \tau) & (N-R+1)\tau & \\ 0 & \dots & 0 & (N-R)(\mu + p\tau) \end{pmatrix},$$

$$Q_1 = \lambda I_{R+1}.$$

Proof. To prove the lemma, we clarify the probabilistic sense of the blocks of the generator. In the following, we will say that a main server “works in regime 0”, if it works without support of a backup server and “works in regime 1”, if it works with support of a backup server. We also assume that all states of the chain under consideration corresponding to i customers in the system form a level \mathbf{i} , $i \geq 0$.

The entries of the matrix $\mathcal{H}_0^{(i,i-1)}$ are the intensities of the chain $\xi_t, t \geq 0$, transition from the level \mathbf{i} to the level $\mathbf{i}-1$. Such a transition occurs at the service completion epoch at one of $(i-r)$ main busy servers working in regime 0 with intensity $(i-r)\mu$, or at the service completion epoch at one of r main busy servers working in regime 1 with intensity $r\tilde{\mu}$. If there are no available backup servers when the timer expires, the number of customers in the system also decrease by one. In this case the customer in service leaves the system forever with intensity $p(i-r)\tau$.

The entries of the matrix $\mathcal{H}_0^{(i,i)}$ are the intensities of the chain $\xi_t, t \geq 0$, transitions without changing the number of busy servers. If the number of busy backup servers is equal to r , the corresponding transition intensity is equal to $r\tau$. If the timer expires and all backup servers are not available, the number of customer in the system also does not change. In this case, the interrupted customer goes to the queue with intensity $(1-p)(i-R)\tau$.

The entries of the matrix $\mathcal{H}_0^{(i,i+1)}$ are the intensities of the chain $\xi_t, t \geq 0$, transitions from the level \mathbf{i} to the level $\mathbf{i}+1$. These transitions occur if an arriving customer sees an idle main server and immediately occupies this server. The intensities of such transitions are defined by the matrix λI_{i+1} .

The block $H_1^{(N)}$ contains the intensities of the chain $\xi_t, t \geq 0$, transitions from the level \mathbf{N} to the level $\mathbf{N}+1$. These transitions occur if an arriving customer sees N main servers busy and forces to go to the queue. The intensities of such transitions are defined by the matrix λI_{R+1} .

The block Q_{-1} is formed by the intensities of the chain $\xi_t, t \geq 0$, transitions from the level $\mathbf{i}, \mathbf{i} \geq N+2$, to the level $\mathbf{i}-1$. The corresponding transition is equal to $(N-r)\mu$ if the service ends at one of r busy main servers working in regime 0, and equal to $r\tilde{\mu}$ if the service ends at one of r busy servers working in regime 1. If there are no available backup servers when the timer expires, the number of customers in the system also decrease by one. In this case the interrupted customer leaves the system. The intensity of such a transition is equal to $p(N-R)\tau$.

The block Q_0 is formed by the intensities of the chain $\xi_t, t \geq 0$, transitions from the level $\mathbf{i}, \mathbf{i} > N$, to the same level. Such a transition occurs without changing the number of busy servers. If the transition causes by the timer expiration, the transition intensity is $(N-r)\tau$. If the timer expires and all backup servers are not available, the number of customers in the system does not also change with probability $1-p$. In this case the interrupted customer returns to the queue. The corresponding transition intensity is equal to $(1-p)(N-R)\tau$.

The block Q_1 is formed by the intensities of the chain $\xi_t, t \geq 0$, transitions from the level $\mathbf{i}, \mathbf{i} > N$, to the level $\mathbf{i}+1$. Such transitions occur when customers

arrive at the system. In this case, the intensities of transitions are given by the matrix λI_{R+1} .

Corollary 1. *The Markov chain $\xi_t, t \geq 0$, is a quasi-birth-and-death (QBD) process with many boundary levels, see [14].*

The proof of the corollary evidently follows from the definition of QBD given in [14] and the structure of the generator Q .

4 Ergodicity Condition. Stationary Distribution

Theorem 1. *The Markov chain $\xi_t, t \geq 0$, is ergodic if and only if the following inequality*

$$\lambda < \sum_{r=0}^R x_r(N-r)\mu + \sum_{r=1}^R x_r r \tilde{\mu} + p x_R(N-R)\tau. \tag{1}$$

holds. Here x_r is a steady state probability that r backup servers are busy under overload condition,

$$x_r = x_0 \frac{\prod_{i=0}^{r-1} (N-i)}{r!} \left(\frac{\tau}{\tilde{\mu}}\right)^r, \quad r = 1, 2, \dots, R, \tag{2}$$

$$x_0 = \left[1 + \sum_{r=1}^R \frac{\prod_{i=0}^{r-1} (N-i)}{r!} \left(\frac{\tau}{\tilde{\mu}}\right)^r \right]^{-1}.$$

Proof. As follows from [14], the QBD process $\xi_t, t \geq 0$, is ergodic if and only if the following inequality is fulfilled:

$$\mathbf{x}Q_1\mathbf{e} < \mathbf{x}Q_{-1}\mathbf{e} \tag{3}$$

where the vector \mathbf{x} is the unique solution to the system of linear algebraic equations

$$\mathbf{x}(Q_{-1} + Q_0 + Q_1) = \mathbf{0}, \quad \mathbf{x}\mathbf{e} = 1, \tag{4}$$

\mathbf{e} is the column vector consisting of 1's, $\mathbf{0}$ is the row vector consisting of 0's.

Using expressions for the matrices Q_{-1}, Q_0, Q_1 , given by Lemma 1, we rewrite system (4) as follows:

$$\begin{aligned} -x_0 N \tau + x_1 \tilde{\mu} &= 0, \\ x_{r-1}(N-r+1)\tau - x_r[(N-r)\tau + r\tilde{\mu}] + x_{r+1}(r+1)\tilde{\mu} &= 0, \\ x_{R-1}(N-R+1)\tau - x_R[(N-R)\tau + R\tilde{\mu}] &= 0, \end{aligned}$$

$$\sum_{r=0}^R x_r = 1.$$

This system defines the steady state probabilities $x_r, r = 0, 1, \dots, R$, of the birth-and-death process with the intensities $(N - r)\tau$ and $r\bar{\mu}$. It is well known that such a system has the unique solution of form (2). Substituting (2) in (3) and taking into account the explicit expressions for Q_{-1}, Q_1 given by Lemma 1, we reduce ergodicity condition (3) to the form (1).

Introduce the notation for the steady state probabilities:

$$p(i, r) = \lim_{t \rightarrow \infty} P\{i_t = i, r_t = r\}, \quad i \geq 0, \quad r = 0, 1, \dots, \min\{i, R\},$$

and for the row-vectors composed of these probabilities:

$$\mathbf{p}_i = (p(i, 0), p(i, 1), \dots, p(i, \min\{i, R\})), \quad i \geq 0.$$

Then the Chapman-Kolmogorov (balance or equilibrium) equations for the steady state probabilities are as follows:

$$\begin{aligned} p_0 \mathcal{H}_0^{(0,0)} + \mathbf{p}_1 \mathcal{H}_0^{(1,0)} &= \mathbf{0}, \\ \mathbf{p}_{i-1} \mathcal{H}_0^{(i-1,i)} + \mathbf{p}_i \mathcal{H}_0^{(i,i)} + \mathbf{p}_{i+1} \mathcal{H}_0^{(i+1,i)} &= \mathbf{0}, \quad i = 1, 2, \dots, N - 1, \\ \mathbf{p}_{N-1} \mathcal{H}_0^{(N-1,N)} + \mathbf{p}_N \mathcal{H}_0^{(N,N)} + \mathbf{p}_{N+1} Q_{-1} &= \mathbf{0}, \\ \mathbf{p}_N H_1^{(N)} + \mathbf{p}_{N+1} Q_0 + \mathbf{p}_{N+2} Q_{-1} &= \mathbf{0}, \\ \mathbf{p}_{i-1} Q_1 + \mathbf{p}_i Q_0 + \mathbf{p}_{i+1} Q_{-1} &= \mathbf{0}, \quad i \geq N + 2. \end{aligned}$$

To calculate the vectors $\mathbf{p}_i, i \geq 0$, we slightly generalize the known algorithm for calculating the stationary distribution of QBD process, see [14], to the case with many boundary levels.

Let $\tilde{Q}_{-1} = (O_{(R+1) \times a} \mid Q_{-1})$, $\mathcal{H}_0 = (\mathcal{H}_0^{(i,j)})_{i,j=0,\overline{N}}$ and $\mathcal{H}_1 = \begin{pmatrix} \mathbf{0}_a^T \\ H_1^{(N)} \end{pmatrix}$

where $a = N + \sum_{i=0}^{N-1} \min\{i, R\}$, $O_{(R+1) \times a}$ is the zero matrix having $(R + 1)$ rows and a columns. Denote also $\boldsymbol{\pi} = (\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_N)$.

Then the algorithm for calculating the stationary distribution is described as follows.

Algorithm

1. Calculate the matrix R as the minimal nonnegative solution of the non-linear matrix equation

$$R^2 Q_0 + R Q_1 + Q_2 = O.$$

2. Calculate the vector \mathbf{p}_{N+1} as the unique solution of the system

$$\begin{aligned} \mathbf{p}_{N+1} [Q_1 + \tilde{Q}_{-1} (-\mathcal{H}_0)^{-1} \mathcal{H}_1 + R Q_0] &= \mathbf{0}, \\ \mathbf{p}_{N+1} [\mathbf{e} + \tilde{Q}_{-1} (-\mathcal{H}_0)^{-1} \mathbf{e} + R(I - R)^{-1} \mathbf{e}] &= 1. \end{aligned}$$

3. Calculate the vector $\boldsymbol{\pi}$ as follows:

$$\boldsymbol{\pi} = \mathbf{p}_{N+1} \tilde{Q}_{-1} (-\mathcal{H}_0)^{-1}.$$

4. Calculate the vectors \mathbf{p}_i , $i = \overline{0, N}$, as follows:

$$\mathbf{p}_i = \boldsymbol{\pi} \text{diag} \left\{ O_{i + \sum_{k=0}^{i-1} \min\{k, R\}}, I_{\min\{i, R\} + 1}, O_{N-i + \sum_{k=i+1}^N \min\{k, R\}} \right\}, \quad i = \overline{0, N}.$$

5. Calculate the vectors \mathbf{p}_{N+i} , $i \geq 2$, by $\mathbf{p}_{N+i} = \mathbf{p}_{N+1} R^{i-1}$, $i \geq 2$.

5 Stationary Performance Measures

1. Average number of customers in the system $L = \sum_{l=0}^{\infty} l \mathbf{p}_l \mathbf{e}$.
2. The probability that the system is idle at an arbitrary moment $p_0 = \mathbf{p}_0 \mathbf{e}$.
3. Steady state probabilities of the number of busy main servers

$$p_n = \mathbf{p}_n \mathbf{e}, \quad n = 0, 1, \dots, N-1, \quad p_N = \mathbf{p}_N \mathbf{e} + \mathbf{p}_{N+1} (I - R)^{-1} \mathbf{e}.$$

4. Average number of busy main servers $N_{busy} = \sum_{n=1}^N n p_n$.
6. Joint probability that there are i customers in the system, r busy main servers working in regime 1 and $\min\{i, N\} - r$ main servers working in regime 0.

$$p_i^{(r)} = \mathbf{p}_i \mathbf{Z}^{(i,r)}, \quad r = 0, 1, \dots, \min\{i, R\}, \quad i \geq 0, \quad (5)$$

where $\mathbf{Z}^{(i,r)} = \begin{pmatrix} \mathbf{0}_r^T \\ 1 \\ \mathbf{0}_{\min\{i, R\}}^T \end{pmatrix}$.

The brief explanation of formula (5) is as follows. Multiplying the vector \mathbf{p}_i by the vector $\mathbf{Z}^{(i,r)}$, we select the part of this vector corresponding to r busy backup servers.

7. Stationary distribution of the number of busy backup servers

$$q_r = \sum_{i=r}^{\infty} p_i(r), \quad r = 0, 1, \dots, R.$$

8. Average number of busy backup servers $N_{busy}^{(backup)} = \sum_{r=1}^R r q_r$.
9. Stationary distribution of the number of busy servers working in regime 1

$$g_n = \sum_{i=n}^{\infty} p_i^{(\min\{i, N\} - n)}, \quad n = 0, 1, \dots, N.$$

10. Average number of busy servers working in regime 0

$$N_{busy}^{(non-support)} = \sum_{n=1}^N n g_n.$$

11. The probability that an arbitrary customer will be lost

$$P_{loss} = 1 - \frac{\sum_{i=1}^{\infty} \mathbf{p}_i \sum_{r=1}^{\min\{i,R\}} [\mathbf{Z}^{(i,r)}(\min\{i, N\} - r)\mu + r\tilde{\mu}]}{\lambda}. \tag{6}$$

In formula (6), the expression $\mathbf{p}_i \mathbf{Z}^{(i,r)}(\min\{i, N\} - r)$ is the probability that i customers stay in the system and r backup servers are working. Multiplying this expression by μ , we obtain the intensity of output flow of customers from the main servers working without support. By analogy, we get the expression $\mathbf{p}_i \mathbf{Z}^{(i,r)} r \tilde{\mu}$ for the intensity of output flow of customers from the main servers working with support. Then the numerator of the fraction in the right-side of (6) is the intensity of output flow from the system while the denominator is the intensity of input flow. So, the fraction (6) is the probability that an arbitrary customer will not be lost. The probability P_{loss} is computed as the complimentary probability.

6 Laplace-Stieltjes Transform of the Sojourn Time Distribution

In this section, we assume that $R = N$, i.e. the number of main servers is equal to the number of backup servers. In this case, the sojourn time of a tagged customer does not depend on the customers arriving to the system after the tagged customer. This fact allow us to derive analytically the Laplace-Stieltjes transform (*LST*) of the sojourn time distribution.

Let $\varphi(u), Re u > 0$, be the *LST* of the sojourn time distribution of an arbitrary customer in the system. This time consists of the waiting time of the customer and its actual service time. The actual sojourn time depends on whether or not the timer expires before the service of the customer by the main server finishes.

Lemma 2. *The Laplace-Stieltjes transform of actual service time distribution of an arbitrary customer is calculated as*

$$\varphi^{(s)}(u) = \frac{1}{u + \mu + \tau} \left(\mu + \frac{\tau \tilde{\mu}}{u + \tilde{\mu}} \right). \tag{7}$$

Proof. We consider the following service scenarios: (i) if the service of a customer by a main server finishes before the timer expires, the actual service time of the customer coincides with the service time by the main server defined by exponential distribution with parameter μ ; (ii) otherwise, the actual service time

consists of time to timer expiration plus the remaining service time which is exponentially distributed with parameter $\tilde{\mu}$.

In case (i) we can derive the following expression for the *LST*: $\int_0^\infty e^{-ut} e^{-(\mu+\tau)t} \mu dt$, in case (ii) the expression for the *LST* is as follows: $\int_0^\infty e^{-ut} \int_0^t e^{-(\mu+\tau)x} \tau dx e^{-\tilde{\mu}(t-x)} \tilde{\mu} dt$. Summing these two expressions, we obtain the desired *LST* in the form

$$\varphi^{(s)}(u) = \int_0^\infty e^{-ut} e^{-(\mu+\tau)t} \mu dt + \int_0^\infty e^{-ut} \int_0^t e^{-(\mu+\tau)x} \tau dx e^{-\tilde{\mu}(t-x)} \tilde{\mu} dt. \quad (8)$$

Calculating the integrals in equation (8), we reduce this equation to the form (7).

Theorem 2. *The Laplace-Stieltjes transform of the sojourn time distribution of an arbitrary customer in the system is calculated by*

$$\varphi(u) = \sum_{i=0}^{N-1} p_i \varphi^{(s)}(u) + \sum_{i=N}^\infty p_i [\varphi^{(s)}(u)]^{i-N+2}. \quad (9)$$

Proof. Formula (9) is derived using the law of total probability. Let an arbitrary customer arrives at the system and finds idle servers, i.e., $i < N$ customers stay in system at the arrival epoch. Then the customer immediately occupies an idle server and starts its service. In this case the sojourn time of the customer is equal to its actual service time which is defined by the *LST* $\varphi^{(s)}(u)$. From what has been said follows that the Laplace-Stieltjes transform under examination is equal to $\varphi^{(s)}(u)$ with probability $\sum_{i=0}^{N-1} p_i$. This explain the first sum over i that occur in the right-hand side of formula (9).

Let now an arbitrary customer arrives into the system and finds all server busy and $i - N$ customers staying in the queue. Such an events occurs with probability p_i , $i \geq N$. In this case the customer goes to the end of the queue and waits for the service until the service of one customer in the service will finish and $i - N$ customers from the queue will be served. Using the above reasonings, memoryless property of exponential distribution and convolution property we derive the *LST* transform of sojourn time of the arriving customer as the product $[\varphi^{(s)}(u)]^{i-N+2}$. This explain the second sum over i in the right-hand side of formula (9).

Corollary 2. *The average sojourn time of an arbitrary customer in the system is calculated using the following formula*

$$\bar{\varphi} = - \sum_{i=0}^{N-1} p_i \frac{\varphi^{(s)}(u)}{du} \Big|_{u=0} - \sum_{i=N}^\infty p_i (i - N + 2) \frac{d\varphi^{(s)}(u)}{du} \Big|_{u=0} \quad (10)$$

where

$$\frac{d\varphi^{(s)}(u)}{du} \Big|_{u=0} = - \frac{1}{(\mu + \tau)} \left(1 + \frac{\tau}{\tilde{\mu}} \right). \quad (11)$$

Proof. To derive formulas (10)–(11), we use the known formula $\bar{\varphi} = -\varphi'(0)$. Differentiating in (9) and using equation (7), after some algebraic transformation we get (10)–(11).

7 Conclusion

In this paper, we investigated a multi-server queueing system that can be used to solve the problem of finding a trade-off between energy saving and quality of service by using backup servers. The service mechanism assumes the connection of a backup server to the service of a customer, if the service of this customer lasts too long. The behavior of the system is described by two-dimensional continuous time Markov chain. We derive the non-trivial but intuitive ergodicity condition, the steady state probabilities and a number of performance characteristics of the system. In case when the numbers of main servers and reserve servers coincide, we derived the sojourn time of an arbitrary customer in the system.

Acknowledgments. This work has been financially supported by the Russian Science Foundation and the Department of Science and Technology (India) via grant No 16-49-02021 (INT/RUS/RSF/16) for the joint research project by the V.A. Trapeznikov Institute of Control Problems of the Russian Academy Sciences and the CMS College Kottayam.

References

1. Klimenok, V., Savko, R.: Tandem system with retrials and impatient customers. *Autom. Remote Control* **76**, 1387–1399 (2015)
2. Kim, C.S., Klimenok, V., Taramin, O.: A tandem retrial queueing system with two Markovian flows and reservation of channels. *Comput. Oper. Res.* **37**, 1238–1246 (2010)
3. Kim, C.S., Klimenok, V., Dudin, A.: Priority tandem queueing system with retrials and reservation of channels as a model of call center. *Comput. Ind. Eng.* **96**, 61–71 (2016)
4. Arnon, S., Barry, J., Karagiannidis, G., Schober, R., Uysal, M.: *Advanced Optical Wireless Communication Systems*. Cambridge University Press, Cambridge (2012)
5. Vishnevsky, V., Kozyrev, D., Semenova, O.: Redundant queueing system with unreliable servers. In: *Proceedings of the 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, Moscow, pp. 383–386 (2014)
6. Vishnevsky, V., Semenova, O., Sharov, S.: Modeling and analysis of a hybrid communication channel based on free-space optical and radio-frequency technologies. *Autom. Remote Control* **72**, 345–352 (2013)
7. Sharov, S., Semenova, O.: Simulation model of wireless channel based on FSO and RF technologies. In: *Proceedings of the International Conference on Distributed Computer and Communication Networks, Theory and Applications (DCCN-2010)*, Moscow, pp. 368–374 (2010)

8. Mitrani, I.: Trading power consumption against performance by reserving blocks of servers. In: Tribastone, M., Gilmore, S. (eds.) *EPEW 2012*. LNCS, vol. 7587, pp. 1–15. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36781-6_1
9. Mitrani, I.: Managing performance and power consumption in a server farm. *Ann. Oper. Res.* **202**, 121–134 (2013)
10. Mitrani, I.: Service center trade-offs between customers impatience and power consumption. *Perform. Eval.* **68**, 1222–1231 (2011)
11. Shwartz, C., Pries, R., Tran-Gia, P.: A queueing analysis of an energy-saving mechanism in data centers. In: *Proceedings of the International Conference on Information Networking*, pp. 70–75 (2012)
12. Kim, C.S., Dudin, A., Dudin, S., Dudina, O.: Hysteresis control by the number of active servers in queueing system with priority service. *Perform. Eval.* **101**, 20–33 (2016)
13. Klimenok, V., Dudin, A., Samouylov, K.: Analysis of the BMAP/PH/ N queueing system with backup servers. *Appl. Math. Model.* **57**, 64–84 (2018)
14. Neuts, M.: *Matrix-Geometric Solutions in Stochastic Models*. The Johns Hopkins University Press, Baltimore (1981)



Multiclass GI/GI/ ∞ Queueing Systems with Random Resource Requirements

Ekaterina Lisovskaya¹(✉), Svetlana Moiseeva¹, and Michele Pagano²

¹ Tomsk State University, Tomsk, Russia
ekaterina_lisovs@mail.ru, smoiseeva@mail.ru

² University of Pisa, Pisa, Italy
m.pagano@iet.unipi.it

Abstract. In the paper we consider a GI/GI/ ∞ queueing system with n types of customers under the assumptions that customers arrive at the queue according to a renewal process and occupy random resource amounts, which are independent of their service times. Since, in general, the analytical solution of the corresponding Kolmogorov differential equations is not available, we focus on the amount of resources occupied by each class of customers under the assumption of infinitely growing arrival rate, and derive its first and second-order asymptotic approximations. In more detail, we show that the n -dimensional probability distribution of the total resource amount is asymptotically n -dimensional Gaussian, and we verify the accuracy of the asymptotics (in terms of Kolmogorov distance) by means of discrete event simulation.

Keywords: Queueing system · Renewal arrival process
Different types of servers · Asymptotic analysis
Infinitely growing arrival rate

1 Introduction

Modern computer networks are characterized by the integration of heterogeneous services (phone calls, text messages, media content, cloud computing) over the same physical infrastructure. The traffic flows generated by the different applications have specific statistical features (in terms of packet size, bit-rate and service requirements) and hence is of primary importance the analysis of queueing systems with several classes of customers [3, 4, 8, 9, 15]. Moreover, due to the heterogeneity of services provided by communication networks [6, 10–14, 16], the features of the required resources should be taken into account.

In traditional multiclass queueing systems the service process is typically characterized in terms of service time distribution. In this paper we assume that customers have different random capacity requirements (depending on their class), so that the proposed model can be useful for analysis and design issues in high-performance computer and communication systems, in which service time and customer volume are independent quantities (see [7] and references therein).

In more detail, the application of the dynamic screening method permit us to analyse heterogeneous resource queuing system with unlimited servers number, non-exponential service time and renewal arriving process are investigated.

The remainder of this paper is structured as follows. Section 2 introduces the mathematical model and the application of the dynamic screening method to the considered multiclass queueing system, while in Sect. 3 the corresponding Kolmogorov equations are presented. Section 4 highlights our main contribution, the derivation of first and second order asymptotics under heavy load conditions (i.e., when the mean interarrival time tends to 0), and their applicability is verified in Sect. 5 by means of discrete event simulation. Finally, Sect. 6 ends the paper with some final remarks.

2 Mathematical Model

Consider a queuing system with an infinite number of servers and n types of customers, characterized by different service times and queuing resource requirements. Arrivals are described by a renewal process with interarrival time distribution $A(z)$ and for each of them class i is selected with probability p_i ($i = 1, \dots, n$), where $\sum_{i=1}^n p_i = 1$. Each arriving customer instantly occupies the first free server, with service time distribution $B_i(\tau)$ and required resource distribution $G_i(y)$, both depending on the type i of the customer. At the end of the service, the customer leaves the system. Resource amount and service times are mutually independent, and do not depend on the epochs of customer arrivals.

Denote by $V_i(t)$ ($i = 1, \dots, n$) the total resource amount occupied by each type of customers at the moment t . The aim of this work is to determine the probabilistic characterization of the n -dimensional process $\{\mathbf{V}(t)\}$. This process is, in general, non Markovian, but it can be investigated by means of the dynamic screening method.

In Fig. 1, $n + 1$ time axes, labeled from 0 to n , are shown: axis 0 indicates the epochs of customers arrivals, while the remaining axes $i = 1, \dots, n$ correspond to the different types of customers.

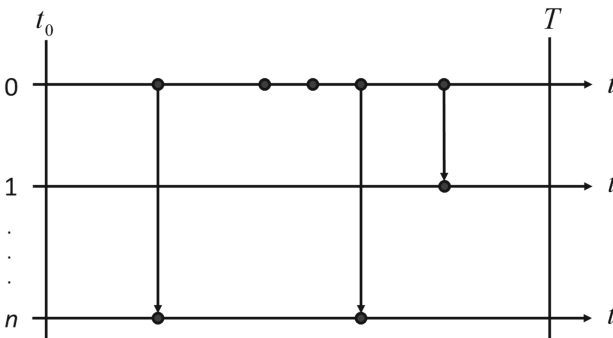


Fig. 1. Dynamic screening of the arrival process

We define a set of n functions (dynamic probabilities) $S_i(t)$, that satisfy the conditions

$$0 \leq S_i(t) \leq 1, \quad \sum_{i=1}^n S_i(t) \leq 1,$$

and assume that a customer, arrived in the system at time t , is screened to axis i with probability $S_i(t)$, and is not screened anywhere with probability

$$S_0(t) = 1 - \sum_{i=1}^n S_i(t).$$

Let the system be empty at time t_0 , and let us choose some arbitrary time T with $T > t_0$. Hence, the probability $S_i(t)$ that a i -type customer, arrived at time t with $t_0 \leq t \leq T$, will be serviced by time T , is given by

$$S_i(t) = 1 - B_i(T - t), \quad (i = 1, \dots, n).$$

Denote by $W_i(t)$ the total resource amount screened on axis i . Then, the extended process $\{\mathbf{V}(t)\}$ satisfies the following property:

$$P\{\mathbf{V}(T) < \mathbf{x}\} = P\{\mathbf{W}(T) < \mathbf{x}\} \tag{1}$$

for all $\mathbf{x} = \{x_1, \dots, x_n\}$, where the inequalities $\mathbf{V}(T) < \mathbf{x}$ and $\mathbf{W}(T) < \mathbf{x}$ mean that $V_1(T) < x_1, \dots, V_n(T) < x_n$ and $W_1(T) < x_1, \dots, W_n(T) < x_n$, respectively. Equality (1) permits us to investigate the process $\{\mathbf{V}(t)\}$ via the analysis of the process $\{\mathbf{W}(t)\}$.

3 Kolmogorov Differential Equations

Let $z(t)$ be the residual time from t to the next arrival (in the renewal input process) and let us denote by

$$P(z, \mathbf{w}, t) = P\{z(t) < z, \mathbf{W}(t) < \mathbf{w}\}$$

the probability distribution of the $n + 1$ -dimensional Markovian process $\{z(t), \mathbf{W}(t)\}$.

By the law of total probability, we get the following system of Kolmogorov differential equations:

$$\begin{aligned} \frac{\partial P(z, \mathbf{w}, t)}{\partial t} &= \frac{\partial P(z, \mathbf{w}, t)}{\partial z} + \frac{\partial P(0, \mathbf{w}, t)}{\partial z} (A(z) - 1) \\ &+ A(z) \sum_{i=1}^n p_i S_i(t) \left[\int_0^{w_i} \frac{\partial P(0, \mathbf{w} - \mathbf{y}_i, t)}{\partial z} dG_i(y) - \frac{\partial P(0, \mathbf{w}, t)}{\partial z} \right], \end{aligned}$$

where $\mathbf{w} = \{w_1, \dots, w_n\}$, $\mathbf{y}_i = \{0, \dots, y, \dots, 0\}$, $z > 0$, $w_i > 0$ ($i = 1, \dots, n$), with the initial condition

$$P(z, \mathbf{w}, t_0) = \begin{cases} R(z), & \mathbf{w} = \mathbf{0}, \\ 0, & \text{otherwise,} \end{cases}$$

where $R(z)$ represents the stationary probability distribution of the values of the random process $z(t)$.

By introducing the partial characteristic function:

$$h(z, \mathbf{v}, t) = \int_0^\infty e^{jv_1 w_1} \dots \int_0^\infty e^{jv_n w_n} P(z, d\mathbf{w}, t) \quad z > 0, w_i > 0,$$

where $j = \sqrt{-1}$ denotes the imaginary unit, we obtain the following equations:

$$\begin{aligned} \frac{\partial h(z, \mathbf{v}, t)}{\partial t} &= \frac{\partial h(z, \mathbf{v}, t)}{\partial z} \\ &+ \frac{\partial h(0, \mathbf{v}, t)}{\partial z} \left[A(z) - 1 + A(z) \sum_{i=1}^n p_i S_i(t) (G_i^*(v_i) - 1) \right], \end{aligned} \quad (2)$$

where

$$G_i^*(v_i) = \int_0^\infty e^{jv_i y} dG_i(y),$$

with the initial condition

$$h(z, \mathbf{v}, t_0) = R(z). \quad (3)$$

4 Asymptotic Analysis

In general, Eq. (2) cannot be solved analytically, but it is possible to find approximate solutions under suitable asymptotic conditions; in this paper we focus on the case of infinitely growing arrival rate.

To this aim, let us write the distribution function of the interarrival times as $A(Nz)$, where N is some parameter that tends to infinity in the asymptotic analysis [1, 2].

Then, Eq. (2) becomes

$$\begin{aligned} \frac{1}{N} \frac{\partial h(z, \mathbf{v}, t)}{\partial t} &= \frac{\partial h(z, \mathbf{v}, t)}{\partial z} \\ &+ \frac{\partial h(0, \mathbf{v}, t)}{\partial z} \left[A(z) - 1 + A(z) \sum_{i=1}^n p_i S_i(t) (G_i^*(v_i) - 1) \right], \end{aligned} \quad (4)$$

with the initial condition (3).

We solve the problem (4)–(3) under the asymptotic condition $N \rightarrow \infty$, and obtain approximate solutions with different levels of accuracy, denoted in the following as “first-order asymptotic” $h(z, \mathbf{v}, t) \approx h^{(1)}(z, \mathbf{v}, t)$ and “second-order asymptotic” $h(z, \mathbf{v}, t) \approx h^{(2)}(z, \mathbf{v}, t)$.

4.1 The First-Order Asymptotic Analysis

As a preliminary result, in this section we present the first-order asymptotic as the following lemma.

Lemma. *The first-order asymptotic characteristic function of the process $\{z(t), \mathbf{W}(t)\}$ is given by*

$$h^{(1)}(z, \mathbf{v}, t) = R(z) \exp \left\{ N\lambda \sum_{i=1}^n jv_i a_1^{(i)} p_i \int_{t_0}^t S_i(\tau) d\tau \right\},$$

where $\lambda = \left(\int_0^\infty (1 - A(x)) dx \right)^{-1}$ and $a_1^{(i)} = \int_0^\infty y dG_i(y)$ is the mean amount of resources required by i -type customers.

Proof. By introducing the following notations

$$\varepsilon = \frac{1}{N}, \mathbf{v} = \varepsilon \mathbf{y}, h(z, \mathbf{v}, t) = f_1(z, \mathbf{y}, t, \varepsilon), \tag{5}$$

in expressions (4) and (3), we get

$$\begin{aligned} \varepsilon \frac{\partial f_1(z, \mathbf{y}, t, \varepsilon)}{\partial t} &= \frac{\partial f_1(z, \mathbf{y}, t, \varepsilon)}{\partial z} \\ &+ \frac{\partial f_1(0, \mathbf{y}, t, \varepsilon)}{\partial z} \left[A(z) - 1 + A(z) \sum_{i=1}^n p_i S_i(t) (G_i^*(\varepsilon y_i) - 1) \right], \end{aligned} \tag{6}$$

with the initial condition

$$f_1(z, \mathbf{y}, t_0, \varepsilon) = R(z). \tag{7}$$

The asymptotic solution of the problem (6)–(7), i.e. the function $f_1(z, \mathbf{y}, t) = \lim_{\varepsilon \rightarrow 0} f_1(z, \mathbf{y}, t, \varepsilon)$, can be obtained in two steps.

Step 1. Let $\varepsilon \rightarrow 0$; then Eq. (6) becomes:

$$\frac{\partial f_1(z, \mathbf{y}, t)}{\partial z} + \frac{\partial f_1(0, \mathbf{y}, t)}{\partial z} (A(z) - 1) = 0.$$

and hence $f_1(z, \mathbf{y}, t)$ can be expressed as

$$f_1(z, \mathbf{y}, t) = R(z) \Phi_1(\mathbf{y}, t), \tag{8}$$

where $\Phi_1(\mathbf{y}, t)$ is some scalar function, satisfying the condition $\Phi_1(\mathbf{y}, t_0) = 1$.

Step 2. Now let $z \rightarrow \infty$ in (6):

$$\varepsilon \frac{\partial f_1(\infty, \mathbf{y}, t, \varepsilon)}{\partial t} = \frac{\partial f_1(0, \mathbf{y}, t, \varepsilon)}{\partial z} \sum_{i=1}^n p_i S_i(t) (G_i^*(\varepsilon y_i) - 1).$$

Then, we substitute here the expression (8), take advantage of the Taylor expansion

$$e^{j\varepsilon s} = 1 + j\varepsilon s + O(\varepsilon^2), \tag{9}$$

divide by ε and perform the limit as $\varepsilon \rightarrow 0$. Since $R'(0) = \lambda$, we get the following differential equation:

$$\frac{\partial \Phi_1(\mathbf{y}, t)}{\partial t} = \Phi_1(\mathbf{y}, t) \lambda \sum_{i=1}^n p_i S_i(t) j y_i a_1^{(i)}, \tag{10}$$

where $a_1^{(i)} = \int_0^\infty y dG_i(y)$.

Taking into account the initial condition, the solution of (10) is

$$\Phi_1(\mathbf{y}, t) = \exp \left\{ \lambda \sum_{i=1}^n j y_i a_1^{(i)} p_i \int_{t_0}^t S_i(\tau) d\tau \right\}.$$

By substituting $\Phi_1(\mathbf{y}, t)$ from (8), we obtain

$$f_1(z, \mathbf{y}, t) = R(z) \exp \left\{ \lambda \sum_{i=1}^n j y_i a_1^{(i)} p_i \int_{t_0}^t S_i(\tau) d\tau \right\}.$$

Therefore, we can write

$$h(z, \mathbf{v}, t) = f_1(z, \mathbf{y}, t, \varepsilon) \approx f_1(z, \mathbf{y}, t) = R(z) \Phi_1(\mathbf{y}, t) = R(z) \exp \left\{ \lambda \sum_{i=1}^n j y_i a_1^{(i)} p_i \int_{t_0}^t S_i(\tau) d\tau \right\} = R(z) \exp \left\{ N\lambda \sum_{i=1}^n j v_i a_1^{(i)} p_i \int_{t_0}^t S_i(\tau) d\tau \right\}.$$

The proof is complete.

4.2 The Second-Order Asymptotic Analysis

Now we are able to formulate the main contribution of this work, which is summarized by the following theorem.

Theorem. *The second-order asymptotic characteristic function of the process $\{z(t), \mathbf{W}(t)\}$ is given by*

$$h^{(2)}(z, \mathbf{v}, t) = R(z) \exp \left\{ N\lambda \sum_{i=1}^n j v_i a_1^{(i)} p_i \int_{t_0}^t S_i(\tau) d\tau \right. \\ \left. + N\lambda \sum_{i=1}^n \frac{(j v_i)^2}{2} a_2^{(i)} p_i \int_{t_0}^t S_i(\tau) d\tau \right\}$$

$$+ \frac{N\kappa}{2} \sum_{i=1}^n \sum_{m=1}^n jv_i a_1^{(i)} p_i jv_m a_1^{(m)} p_m \int_{t_0}^t S_i(\tau) S_m(\tau) d\tau \Bigg\}, \tag{11}$$

where $a_2^{(i)} = \int_0^\infty y^2 dG_i(y)$ and $\kappa = \lambda^3 (\sigma^2 - a^2)$, a and σ^2 being the mean and the variance of the interarrival time, respectively.

Proof. Let $h_2(z, \mathbf{v}, t)$ be a solution of the following equation

$$h(z, \mathbf{v}, t) = h_2(z, \mathbf{v}, t) \exp \left\{ N\lambda \sum_{i=1}^n jv_i a_1^{(i)} p_i \int_{t_0}^t S_i(\tau) d\tau \right\} \tag{12}$$

Substituting this expression into (3) and (4), we get the following equivalent problem:

$$\begin{aligned} & \frac{1}{N} \frac{\partial h_2(z, \mathbf{v}, t)}{\partial t} + \lambda h_2(z, \mathbf{v}, t) \sum_{i=1}^n jv_i a_1^{(i)} p_i S_i(t) = \\ & \frac{\partial h_2(z, \mathbf{v}, t)}{\partial z} + \frac{\partial h_2(0, \mathbf{v}, t)}{\partial z} \left[A(z) - 1 + A(z) \sum_{i=1}^n p_i S_i(t) (G_i^*(v_i) - 1) \right], \end{aligned} \tag{13}$$

with the initial condition

$$h_2(z, \mathbf{v}, t_0) = R(z). \tag{14}$$

By performing the following changes of variable

$$\varepsilon^2 = \frac{1}{N}, \mathbf{v} = \varepsilon \mathbf{y}, h_2(z, \mathbf{v}, t) = f_2(z, \mathbf{y}, t, \varepsilon). \tag{15}$$

in (13) and (14), we get the following problem:

$$\begin{aligned} & \varepsilon^2 \frac{\partial f_2(z, \mathbf{y}, t, \varepsilon)}{\partial t} + f_2(z, \mathbf{y}, t, \varepsilon) \lambda \sum_{i=1}^n j\varepsilon y_i a_1^{(i)} p_i S_i(t) = \frac{\partial f_2(z, \mathbf{y}, t, \varepsilon)}{\partial z} \\ & + \frac{\partial f_2(0, \mathbf{y}, t, \varepsilon)}{\partial z} \left[A(z) - 1 + A(z) \sum_{i=1}^n p_i S_i(t) (G_i^*(\varepsilon y_i) - 1) \right], \end{aligned} \tag{16}$$

with the initial condition

$$f_2(z, \mathbf{y}, t_0, \varepsilon) = R(z). \tag{17}$$

As a generalization of the approach used in the previous subsection, the asymptotic solution of this problem

$$f_2(z, \mathbf{y}, t) = \lim_{\varepsilon \rightarrow 0} f_2(z, \mathbf{y}, t, \varepsilon)$$

can be derived in three steps.

Step 1. Letting $\varepsilon \rightarrow 0$ in (16), we get the following equation:

$$\frac{\partial f_2(z, \mathbf{y}, t)}{\partial z} + \frac{\partial f_2(0, \mathbf{y}, t)}{\partial z} (A(z) - 1) = 0.$$

Hence, we can express $f_2(z, \mathbf{y}, t)$ as

$$f_2(z, \mathbf{y}, t) = R(z) \Phi_2(\mathbf{y}, t), \tag{18}$$

where $\Phi_2(\mathbf{y}, t)$ is some scalar function that satisfies the condition $\Phi_2(\mathbf{y}, t_0) = 1$.

Step 2. The solution $f_2(z, \mathbf{y}, t)$ can be represented in the expansion form

$$f_2(z, \mathbf{y}, t) = \Phi_2(\mathbf{y}, t) \left[R(z) + f(z) \sum_{i=1}^n j\varepsilon y_i a_1^{(i)} p_i S_i(t) \right] + O(\varepsilon^2), \tag{19}$$

where $f(z)$ is a suitable function. By substituting the previous expression and the Taylor-Maclaurin expansion (9) in (16), taking into account that $R'(z) = \lambda(1 - A(z))$, it is easy to verify that

$$f'(0) = \lambda f(\infty) + \frac{\kappa}{2},$$

and $\kappa = \lambda^3(\sigma^2 - a^2)$, where a and σ_2 are the mean and the variance of the interarrival time.

Step 3. Letting $z \rightarrow \infty$ in (16), by the definition of the function $f_2(z, \mathbf{y}, t, \varepsilon)$, we obtain

$$\lim_{z \rightarrow \infty} \frac{\partial f_2(z, \mathbf{y}, t, \varepsilon)}{\partial z} = 0,$$

and, taking into account the expansion

$$e^{j\varepsilon s} = 1 + j\varepsilon s + \frac{(j\varepsilon s)^2}{2} + O(\varepsilon^3),$$

we can write

$$\begin{aligned} &\varepsilon^2 \frac{\partial f_2(\infty, \mathbf{y}, t, \varepsilon)}{\partial t} + f_2(\infty, \mathbf{y}, t, \varepsilon) \lambda \sum_{i=1}^n p_i S_i(t) j\varepsilon y_i a_1^{(i)} \\ &= \frac{\partial f_2(0, \mathbf{y}, t, \varepsilon)}{\partial z} \sum_{i=1}^n p_i S_i(t) \left(j\varepsilon y_i a_1^{(i)} + \frac{(j\varepsilon y_i)^2}{2} a_2^{(i)} \right) + O(\varepsilon^3), \end{aligned}$$

where $a_2^{(i)} = \int_0^\infty y^2 dG_i(y)$.

By substituting here the expansion (19) and taking the limit as $z \rightarrow \infty$, we get

$$\begin{aligned} &\varepsilon^2 \frac{\partial \Phi_2(\mathbf{y}, t)}{\partial t} + \Phi_2(\mathbf{y}, t) \lambda \sum_{i=1}^n j\varepsilon y_i a_1^{(i)} p_i S_i(t) \sum_{m=1}^n j\varepsilon y_m a_1^{(m)} p_m S_m(t) f(\infty) \\ &= \Phi_2(\mathbf{y}, t) \lambda \sum_{i=1}^n p_i S_i(t) \left(j\varepsilon y_i a_1^{(i)} + \frac{(j\varepsilon y_i)^2}{2} a_2^{(i)} \right) \\ &+ \Phi_2(\mathbf{y}, t) f'(0) \sum_{i=1}^n p_i S_i(t) j\varepsilon y_i a_1^{(i)} \sum_{m=1}^n p_m S_m(t) \left(j\varepsilon y_m a_1^{(m)} + \frac{(j\varepsilon y_m)^2}{2} a_2^{(m)} \right) + O(\varepsilon^3). \end{aligned}$$

After simple manipulations, and taking into account that

$$\frac{\kappa}{2} = (f'(0) - f(\infty)),$$

we get the following differential equation for $\Phi_2(\mathbf{y}, t)$:

$$\begin{aligned} \frac{\partial \Phi_2(\mathbf{y}, t)}{\partial t} = & \Phi_2(\mathbf{y}, t) \left[\lambda \sum_{i=1}^n \frac{(jy_i)^2}{2} a_2^{(i)} p_i S_i(t) \right. \\ & \left. + \frac{\kappa}{2} \sum_{i=1}^n \sum_{m=1}^n jy_i a_1^{(i)} p_i jy_m a_1^{(m)} p_m S_i(t) S_m(t) \right], \end{aligned}$$

whose solution (with the given initial condition) can be expressed

$$\begin{aligned} \Phi_2(\mathbf{y}, t) = & \exp \left\{ \lambda \sum_{i=1}^n \frac{(jy_i)^2}{2} a_2^{(i)} p_i \int_{t_0}^t S_i(\tau) d\tau \right. \\ & \left. + \frac{\kappa}{2} \sum_{i=1}^n \sum_{m=1}^n jy_i a_1^{(i)} p_i jy_m a_1^{(m)} p_m \int_{t_0}^t S_i(\tau) S_m(\tau) d\tau \right\} \end{aligned}$$

Substituting this expression into (18) and performing the inverse substitutions of (15) and (12), we get the expression (11) for the asymptotic characteristic function of the process $\{z(t), \mathbf{W}(t)\}$.

The proof is complete.

Corollary. *For $z \rightarrow \infty, t = T$ and $t_0 \rightarrow -\infty$ we get the characteristic function of the process $\{\mathbf{V}(t)\}$ in the steady state regime*

$$\begin{aligned} h(\mathbf{v}) = & \exp \left\{ N\lambda \sum_{i=1}^n jv_i a_1^{(i)} b_i \right. \\ & \left. + N\lambda \sum_{i=1}^n \frac{(jv_i)^2}{2} a_2^{(i)} p_i b_i + \frac{N\kappa}{2} \sum_{i=1}^n \sum_{m=1}^n jv_i a_1^{(i)} jv_m a_1^{(m)} K_{im} \right\}, \end{aligned} \tag{20}$$

where

$$\begin{aligned} b_i = & p_i \int_0^\infty (1 - B_i(\tau)) d\tau, \\ K_{im} = & p_i p_m \int_0^\infty (1 - B_i(\tau))(1 - B_m(\tau)) d\tau. \end{aligned}$$

The structure of function (20) implies that the n -dimensional process $\{\mathbf{V}(t)\}$ is asymptotically Gaussian with mean

$$\mathbf{a} = N\lambda \left[a_1^{(1)} b_1 \ a_1^{(2)} b_2 \ \dots \ a_1^{(n)} b_n \right]$$

and covariance matrix

$$\mathbf{K} = N \left[\lambda \mathbf{K}^{(1)} + \kappa \mathbf{K}^{(2)} \right],$$

where

$$\mathbf{K}^{(1)} = \begin{bmatrix} a_2^{(1)} b_1 & 0 & \dots & 0 \\ 0 & a_2^{(2)} b_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_2^{(n)} b_n \end{bmatrix},$$

$$\mathbf{K}^{(2)} = \begin{bmatrix} a_1^{(1)} a_1^{(1)} K_{11} & a_1^{(1)} a_1^{(2)} K_{12} & \dots & a_1^{(1)} a_1^{(n)} K_{1n} \\ a_1^{(2)} a_1^{(1)} K_{21} & a_1^{(2)} a_1^{(2)} K_{22} & \dots & a_1^{(2)} a_1^{(n)} K_{2n} \\ \dots & \dots & \dots & \dots \\ a_1^{(n)} a_1^{(1)} K_{n1} & a_1^{(n)} a_1^{(2)} K_{n2} & \dots & a_1^{(n)} a_1^{(n)} K_{nn} \end{bmatrix}.$$

5 Simulation Results

The previous results, summarized by (20), are obtained under the asymptotic condition of infinitely growing arrival rate ($N \rightarrow \infty$) and, hence, they can provide suitable approximations only for sufficiently large values of N . To investigate their practical applicability, we have considered several simulation scenarios, varying all the system parameters (i.e., the distributions of the interarrival and service times and of the customer capacity as well as the probabilities p_i). Since all the different simulation sets led to similar results, for sake of brevity, we present just one of them.

In more detail, we assume that the input renewal process is characterized by a uniform distribution of the interarrival time in the interval $[0.5, 1.5]$, corresponding to a fundamental rate of arrivals $\lambda = 1$ customers per time unit. Moreover, each arriving customer may belong to one of $n = 3$ types, according to the following probabilities: $p_1 = 0.5$, $p_2 = 0.3$ and $p_3 = 0.2$. We assume that resource amounts occupied by each customer type have exponential distribution, with parameters 2, 1 and 0.4, respectively. Finally, the service times have gamma distribution with shape and inverse scale parameters equal to $\alpha_1 = \beta_1 = 0.5$, $\alpha_2 = \beta_2 = 1.5$ and $\alpha_3 = \beta_3 = 2.5$, respectively.

Our aim is to show that the Gaussian approximation gets better and better as N goes to infinity, thus providing some indications on *reasonable* lower bounds of N for the applicability of (20). Hence, we carried out different sets of simulation experiments (in each of them 10^{10} arrivals were generated) for increasing values of N and compared the asymptotic distributions with the empiric ones in terms of Kolmogorov distance [5]

$$\Delta = \sup_x |F(x) - G(x)|$$

where $F(x)$ is the cumulative distribution function built on the basis of simulation results, and $G(x)$ is the Gaussian approximation given by (20); the corresponding parameters for the three classes are summarized in Table 1. For sake

Table 1. Parameters of Gaussian approximations

Customers class	Mean	Variance
First	$0.25 N$	$0.229 N$
Second	$0.3 N$	$0.553 N$
Third	$0.5 N$	$2.349 N$

of brevity, we show the results only for the marginal distributions of the total resource amount for each class of customers.

Tables 2, 3 and 4 report the values of the Kolmogorov distance for the three types of customers, highlighting that the goodness of the approximation depends not only on N , but also on the different statistical features of the considered customers class.

Table 2. Kolmogorov distance for the first type of customers

N	1	35	50	75	100	125	200	500	1000
Δ	0.294	0.033	0.027	0.022	0.019	0.017	0.014	0.009	0.006

Table 3. Kolmogorov distance for the second type of customers

N	1	35	50	75	100	125	200	500	1000
Δ	0.377	0.042	0.035	0.028	0.025	0.022	0.018	0.011	0.008

Table 4. Kolmogorov distance for the third type of customers

N	1	35	50	75	100	125	200	500	1000
Δ	0.419	0.053	0.043	0.036	0.031	0.028	0.022	0.014	0.009

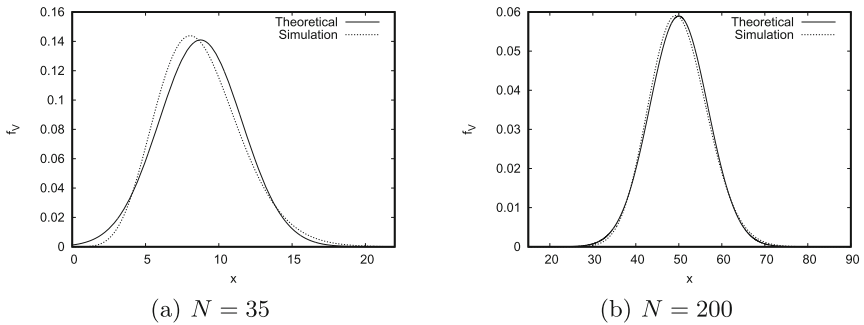


Fig. 2. Distributions of the total resource amount for the first type of customers

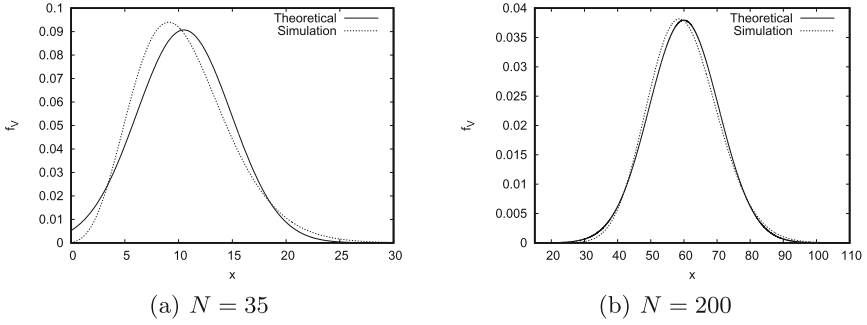


Fig. 3. Distributions of the total resource amount for the second type of customers

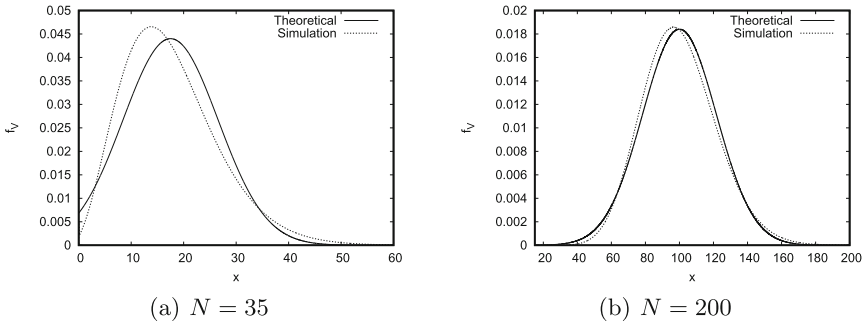


Fig. 4. Distributions of the total resource amount for the third type of customers

As expected, the asymptotic results become more and more accurate when the scale parameter N increases. This conclusion is also confirmed by Figs. 2, 3 and 4, which compare the asymptotic approximations with the empirical histograms for the total resource amount of each type of customers for two different values of N .

6 Conclusions

In this work we considered a GI/GI/ ∞ queue with n types of customers under the assumption that arrivals follow a renewal process and each customer occupies a random resource amount, independent of its service time. At first we determined the corresponding Kolmogorov differential equations, which in the general case cannot be solved analytically. Hence, we derived first and second-order asymptotic approximations in case of infinitely growing arrival rate, and we pointed out that the n -dimensional probability distribution of the total resource amount is asymptotically n -dimensional Gaussian. Finally, by means of discrete-event simulation we verified the goodness of the approximation, and highlighted how the applicability region of the asymptotic approximation (i.e., lower bounds on

the scale parameter N for all the different classes of users) can be determined by considering the Kolmogorov distance as accuracy measure.




References

1. Alexander, M., Anatoly, N.: Asymptotic analysis of the infinite-server queueing system with high-rate semi-markov arrivals. In: 2014 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), pp. 507–513, October 2014
2. Alexander, M., Anatoly, N.: Queueing network with high-rate arrivals. *Eur. J. Oper. Res.* **254**(1), 161–168 (2016)
3. Ali, T., Winfried, G., Javad, T.: Optimal policies of $M(t)/M/c/c$ queues with two different levels of servers. **249**, 10 (2015)
4. Efrosinin, D., Feichtenschlager, M.: Optimal control of $M(t)/M/K$ queues with homogeneous and heterogeneous servers. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) DCCN 2016. CCIS, vol. 678, pp. 132–144. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-51917-3_13
5. Fedorova, E.: The second order asymptotic analysis under heavy load condition for retrial queueing system MMPP/M/1. In: Dudin, A., Nazarov, A., Yakupov, R. (eds.) ITMM 2015. CCIS, vol. 564, pp. 344–357. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25861-4_29
6. Lisovskaya, E., Moiseeva, S., Pagano, M.: On the total customers' capacity in multi-server queues. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) ITMM 2017. CCIS, vol. 800, pp. 56–67. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_5
7. Ekaterina, L., Svetlana, M., Michele, P., Viktoriya, P.: Study of the MMPP/GI/ ∞ queueing system with random customers' capacities. *Inf. Appl.* **11**(4), 109–117 (2017)
8. Pankratova, E., Farkhadov, M., Gelenbe, E.: Research of heterogeneous queueing system $SM-M^{(n)}|\infty$. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) ITMM 2017. CCIS, vol. 800, pp. 122–132. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_10
9. Pankratova, E., Moiseeva, S.: Queueing system $MAP/M/\infty$ with n types of customers. In: Dudin, A., Nazarov, A., Yakupov, R., Gortsev, A. (eds.) ITMM 2014. CCIS, vol. 487, pp. 356–366. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13671-4_41
10. Kumar, V., Gupta, A.K., Kumar, S., Kumar, V.: Mobility management in heterogeneous wireless networks based on IEEE 802.21 framework. In: 2015 International Conference on Advances in Computer Engineering and Applications, pp. 930–935, March 2015
11. Oleg, T., Wojciech, K.: Queueing system with processor sharing and limited memory under control of the AQM mechanism. *Autom. Remote Control* **76**(10), 1784–1796 (2015)
12. Valeriy, N., Konstantin, S.: Analysis of multi-resource loss system with state-dependent arrival and service rates. *Probab. Eng. Inf. Sci.* **31**(4), 413–419 (2017)
13. Valeriy, N., Konstantin, S., Eduard, S., Sergey, A.: Two approaches to analyzing dynamic cellular networks with limited resources. In: 2014 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), pp. 485–488, October 2014

14. Valeriy, N., Konstantin, S., Nataliya, Y., Eduard, S., Sergey, A., Andrey, S.: LTE performance analysis using queuing systems with finite resources and random requirements. In: 2015 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), pp. 100–103, October 2015
15. Wang, R., Jouini, O., Benjaafar, S.: Service systems with finite and heterogeneous customer arrivals. *Manuf. Serv. Oper. Manag.* **16**(3), 365–380 (2014)
16. Borodakiy, V.Y., Samouylov, K.E., Gudkova, I.A., Markova, E.V.: Analyzing mean bit rate of multicast video conference in lte network with adaptive radio admission control scheme*. *J. Math. Sci.* **218**(3), 257–268 (2016)



Cost and Effect of Replication and Quorum in Desktop Grid Computing

Alexander Rumyantsev^{1,2} , Srinivas Chakravarthy³ , Evsey Morozov¹ ,
and Stanislav Remnev²

¹ Institute of Applied Mathematical Research, Karelian Research Centre of RAS,
Petrozavodsk, Russia

ar0@krc.karelia.ru

² Petrozavodsk State University, Petrozavodsk, Russia

³ Departments of Industrial and Manufacturing Engineering and Mathematics,
Kettering University, Flint, MI 48504, USA

Abstract. A Split–Merge multiserver model of a Desktop Grid computing system is studied. Heavy-tailed distributions are used for service times of tasks in a system, including the Pareto distribution, which allows one to obtain some analytical results. The effects of replication and quorum parameters on the key performance measures such as response time and cost of a Desktop Grid system are studied both analytically and through simulation under a variety of scenarios for system configuration and system load. Moment properties of the workload vector, which not only highlight possible heterogeneity but also play a key role in practical applications, are derived.

Keywords: Split–Merge Model · Heavy tails · Pareto · Multiserver
Desktop grid · Replication · Quorum · Moment properties

1 Introduction

Parallel computing is widely used as a tool for solving many practical problems that require a large number of resource-intensive computational experiments as well as processing large amounts of data. Some notable ones can be found in quantum chemistry, molecular biology, hydrodynamics and other branches of science. With explosive use of the Internet coupled with the growth in the number, power, accessibility, performance and decrease in cost of personal computers, the Desktop Grid (DG) computing, a particular and popular case of parallel computing is in high demand [5]. In the DG system, when heterogeneous computing resources like personal computers, laptops, web servers, cluster nodes, as well as wearable devices, are *idle* used as computing resources. This is accomplished through splitting a special set of tasks into loosely coupled (or independent) and relatively small units of work. Further, DG systems are subdivided to Volunteer Computing (VC), in which the resources are donated by the volunteers

and Enterprise DG (EDG), where resources belong to an organization/group of organizations [9].

Fast application turnaround, low delays and low response time are among the key optimization goals for a DG system [5]. In DG systems the response time can be reduced by sending a adequate number of *replicas* (identical copies) of workunits (hereafter we use the word *task* as a synonym of workunit) to compute nodes and waiting for a *quorum* (fixed number of valid results). At the same time, the cost of the calculations (expressed in time and/or energy consumed during problem solving activities) are important, especially for EDG systems, and such costs also depend on the *replication and quorum* parameters. Thus, it is important to study the stochastic models of DG systems which fall in the class of multiserver models.

In this paper we continue the study of the effect of replication and quorum parameters on key performance characteristics of an EDG model discussed in the papers [4,15]. While in [15] the authors study the model under the assumptions of Poisson arrivals and phase type (PH) services, in [4] the authors use a more versatile point process, namely, Markovian arrival process to model the arrivals and employ the well-known matrix-analytic method under the assumption of phase type services. Further, for more general set of assumptions (e.g. Weibull service time distribution) simulation approach is used. Continuing that research along with inspiring open problems stated in [10], in this paper we study in more detail the effect of the tails of service time distribution on the *response time* (also known as latency or sojourn time) of a task and the *cost* (which is proportional to service time of a task).

Heavy-tailed distributions are widely used in modeling of computer and communication systems [3,6,7,14,19]. Key features of heavy-tailed distributions are related to the finiteness of moments of random variables which, under suitable aggregation, may help to model long range dependence and hence capture key performance characteristics of a system [16] that otherwise would have been dramatically hidden. Moreover, modeling heavy-tailed distribution for service times of tasks in multiserver model may lead to heterogeneous moment properties of servers resulting in many servers being busy with unusually long service times of tasks [18]. (This effect may explain uneven behavior of the time required to complete the so-called tail computation of a finite number of workunits in a DG system, the problem pointed out in [5].) In the present paper we study these effects adopting a Split–Merge model of an EDG presented in [4,15]. We use the celebrated Pareto distribution for the service times of tasks in the model allowing us to obtain some analytical results.

The paper is organized as follows. In Sect. 2, we briefly recall the Split–Merge model of an EDG and highlight some known basic properties of order statistics sampled from heavy-tailed Pareto distribution. Moreover, in this section we discuss the stability condition of the multiserver system. In Sect. 3 we discuss in detail the moment properties of the stationary workload of a multiserver system with special focus on the novel results obtained in [18]. We apply these results to the EDG model to obtain the moment properties for stationary workload

and also highlight the possible trade-off between the cost and response time of such a system which rely on replication and quorum. Finally, in Sect. 4 we provide numerical results based on simulating a few scenarios to illustrate the dependence of the response time and the cost on the replication and quorum parameters as well as the system load under a wide set of service time distributions that includes heavy-tailed ones. Some concluding remarks and directions for possible future research are outlined in Sect. 5.

2 Split–Merge Model of a Desktop Grid Computing

In this section we give a brief overview of Split–Merge model of a DG presented in [4, 15]. Consider a single EDG system in which a fixed number m identical servers (known as hosts) process the tasks arriving according to a renewal process, $\{t_i\}_{i \geq 1}$, with intensity λ . Upon arrival, each task is replicated $r \leq m$ times (i.e., each task is made into r identical copies) and waits (if necessary) in a single First-Come-First-Served queue until a required r servers become free. It is assumed that all r copies of a task begin service simultaneously (provided there are r servers available). It is assumed that the service times of these copies are independent and identically distributed (i.i.d) with distribution function F_S . In Sect. 3, we will consider the case when F_S is a Pareto distribution. Once simultaneous services begin for the r replicates of a task, we say a *quorum* is formed when q ($q \leq r$) of the r replicas complete their services. Soon after the quorum is achieved for a task, the (on-going) services of the remaining $r - q$ replicas of that task will be canceled and hence removed from the system. Thus, in this model the service of a task is assumed to be completed when a quorum is achieved for that task.

Let S_1, \dots, S_r be the i.i.d. random variables describing the service times of a generic task. Since the services of r replicates of a task are to be started simultaneously, the time, say, $S_{q:r}$, to achieve the quorum for the task is the q^{th} order statistics of r random variables that are i.i.d. Thus, the system under study is equivalent to a $G/G/\lfloor \frac{m}{r} \rfloor$ queueing system with generic service time $S_{q:r}$ (where $\lfloor x \rfloor$ is the greatest integer less or equal to x). It should be pointed out from the system’s point of view that it is efficient to have m to be multiples of r . The key performance measures of the model are:

Response time $Z(q, r) := E(D) + E(S_{q:r})$, where $E(D)$ is the mean stationary delay of a task,

Cost $C(q, r) := rE(S_{q:r})$ is obtained using the mean service time of a quorum.

In Fig. 1 we display graphically our model in detail.

2.1 Properties of Order Statistics

The two key performance measures listed earlier involve order statistics of i.i.d. random variables and hence in this section we will list some basic properties of order statistics.

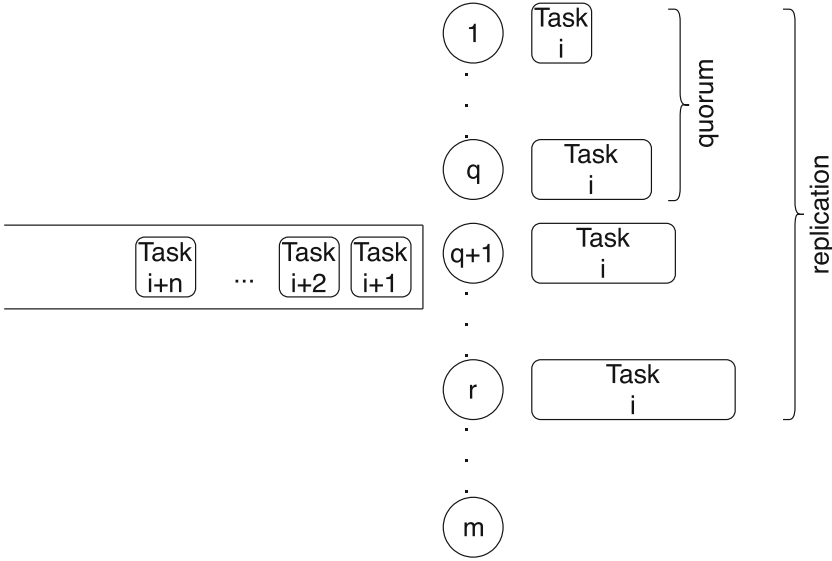


Fig. 1. Split-Merge multiserver model of an EDG

Suppose that X_1, \dots, X_r be r i.i.d. copies of a r.v. X with distribution function (d.f.) F . Let $X_{q:r}$ denote the q^{th} order statistics. That is,

$$X_{1:r} \leq \dots \leq X_{r:r}.$$

Recall (see, for example, [2]) that

$$P(X_{q:r} \leq x) = \sum_{i=q}^r \binom{r}{i} F^i(x) \bar{F}^{r-i}(x),$$

where $\bar{F}(x) := 1 - F(x)$ is the tail distribution. In particular, the minimum, $X_{1:r}$, has the tail distribution

$$P(X_{1:r} > x) = \bar{F}^r(x),$$

while the maximum, $X_{r:r}$, has d.f.

$$P(X_{r:r} \leq x) = F^r(x).$$

Thus, using the terminology of the stochastic ordering, we have the following stochastic inequalities

$$X_{1:r} \leq_{st} X \leq_{st} X_{r:r},$$

which, in particular, imply the corresponding inequalities for the expected values. This property can be used to minimize the time to obtain result in systems with

replication and quorum, such as the EDG. However, the distribution of $X_{q:r}$ for arbitrary values of q and r is not known (except some special cases), and thus obtaining the optimal (in some sense) replication and quorum parameters is difficult in general. However, we recall some important monotonicity properties of order statistics which are hereafter used as accuracy checks in our computations.

For fixed r , the following relation holds [1]:

$$P(X_{q:r} \leq x) = P(X_{q-1:r} \leq x) - \binom{r}{q-1} F^{q-1}(x) \bar{F}^{r-q+1}(x), \quad q > 1. \quad (1)$$

This relation in particular means that $X_{q:r} \geq_{st} X_{q-1:r}$, which is as is to be expected, and hence the cost, $C(q, r)$, monotonically non-decreases with increasing q (for a fixed r). That is,

$$C(q, r) \geq C(q - 1, r), \quad q > 1.$$

Moreover, for some $r, q > 1$, the following monotonicity result holds [1]:

$$P(X_{q:r} \leq x) = P(X_{q-1:r-1} \leq x) - \binom{r-1}{q-1} F^{q-1}(x) \bar{F}^{r-q+1}(x), \quad q > 1. \quad (2)$$

Thus, in particular, $X_{q:r} \geq_{st} X_{q-1:r-1}$ and the cost $C(q, r)$ monotonically non-decreases with simultaneous increase in q and r . That is,

$$C(q, r) \geq C(q - 1, r - 1), \quad q, r > 1.$$

Finally, it follows from [1] that

$$P(X_{q:r} \leq x) = P(X_{q:r-1} \leq x) + \binom{r-1}{q-1} F^q(x) \bar{F}^{r-q}(x), \quad q \geq 1. \quad (3)$$

In particular this means that $X_{q:r} \leq_{st} X_{q:r-1}$; however, this does not induce monotonicity of $C(q, r)$ on r for a given q , forcing one to study based on the type of service time distribution considered.

2.2 Cost of Computations for Pareto Service Time Distribution

Now, to prepare the study of EDG model with Pareto service time distribution, we present the moment properties of Pareto distribution obtained in [13]. This analysis has been further developed in [8]. Consider Pareto r.v. S with density

$$f(x) = \alpha x_0^\alpha x^{-\alpha-1}, \quad x > x_0 \quad (f(x) = 0, x \leq x_0), \quad \alpha > 1.$$

Note that

$$E(S^k) < \infty, \quad k < \alpha.$$

Then the following result holds [8]

$$E(S_{q:r}^k) = x_0^k \frac{r!}{(r-q)!} \frac{\Gamma(r-q+1-k/\alpha)}{\Gamma(r+1-k/\alpha)}, \quad 0 < k < \alpha(r-q+1), \quad (4)$$

where Γ is the gamma function. In particular, (4) means that the r.v. $S_{q:r}$ has *moment index* $\alpha(r - q + 1)$. That is,

$$\sup\{k > 0 : E(S_{q:r}^k) < \infty\} = \alpha(r - q + 1) \geq \alpha,$$

where α is the moment index of S . Moreover, the moment index of $S_{q:r}$ *linearly increases* in $r - q$. The latter may be used to select r and q in such a way to guarantee finiteness of the required moments (say, the variance) of the replicated task even if the moment index α of the original r.v. S is small. However, to guarantee $C(q, r) < \infty$ for given q and r , it is required that $\alpha(r - q + 1) - 1 > 0$. Note also that (4) allows one to obtain the cost $C(q, r)$ analytically. In particular, for $q = 1$ we obtain the cost as

$$C(1, r) = x_0 \frac{r^2 \alpha}{r\alpha - 1}, \quad r > 1/\alpha. \tag{5}$$

It follows that the cost

$$\frac{C(1, r)}{r} \rightarrow x_0, \quad r \rightarrow \infty.$$

The above indicates that the cost grows asymptotically linear in r . After some standard algebraic manipulations, we deduce from (4) the following recurrent relation

$$C(q, r + 1) = \frac{(r + 1)^2(r - q + 1 - 1/\alpha)}{r(r - q + 1)(r + 1 - 1/\alpha)} C(q, r). \tag{6}$$

Straightforward manipulations with r.h.s. of (6) allow one to conclude that the cost $C(q, r)$ *increases* in r , for a given q , if the following inequality (of the second order in r) holds good:

$$\alpha r^2 + (2\alpha - \alpha q - q - 1)r - \alpha q + \alpha - 1 > 0, \tag{7}$$

and is *non-increasing*, otherwise. The discriminant of (7), after some algebra, equals

$$\mathcal{D}(q) = (\alpha q + q - 1)^2 + 4q > 0, \quad q \geq 1, \tag{8}$$

and thus (7) always has two distinct roots. It can be verified that the smallest root of (7) is less than $1/\alpha$. In order to guarantee a finite cost, it is necessary that $r > 1/\alpha$. Hence, Equation (7) holds good for a given q if

$$r \geq r_0(q) := \left\lceil \frac{1 + q + \alpha q - 2\alpha + \sqrt{\mathcal{D}(q)}}{2\alpha} \right\rceil. \tag{9}$$

Thus, $r_0(q)$ gives the *minimum* of cost for a given q . In particular, it can be shown from (9) that $r_0(1) = 1$ for $\alpha > 1.5$.

To illustrate the dependence of the cost function on q and r for Pareto services we display the cost of a 100-server EDG by varying $r = 1, \dots, 10$ and $q = 1, \dots, r$ for a standard (with $x_0 = 1$) Pareto distribution with $\alpha = 1.1$ (note that such a Pareto r.v. has an infinite variance) and $\alpha = 3$, in Fig. 2. The nonlinear

dependence of the cost on replication and quorum parameters is clearly seen and, as expected from (1)–(2), the cost is monotone both with increasing q (for a fixed r), and with simultaneous increase in q and r . Moreover, for a fixed q it can be seen that for $r > r_0(q)$ the cost is increasing with r . The points, $r_0(q)$, are marked with squares in those plots. The obtained monotonicity properties allow one to make the following recommendations with regard to optimal r and q values that minimizes the cost $C(q, r)$:

- since $C(q, r)$ is monotone in q for a fixed r , it is preferable to have q as small as possible by taking into account other considerations such as reliability and properties of the problem under study in the EDG model;
- since the minimum for $C(q, r)$ occurs at $r_0(q)$ and since this point (see (9)) depends on α , it is preferable to select r close to $r_0(q)$. The monotonicity on r for a given q is used to arrive at this recommendation.

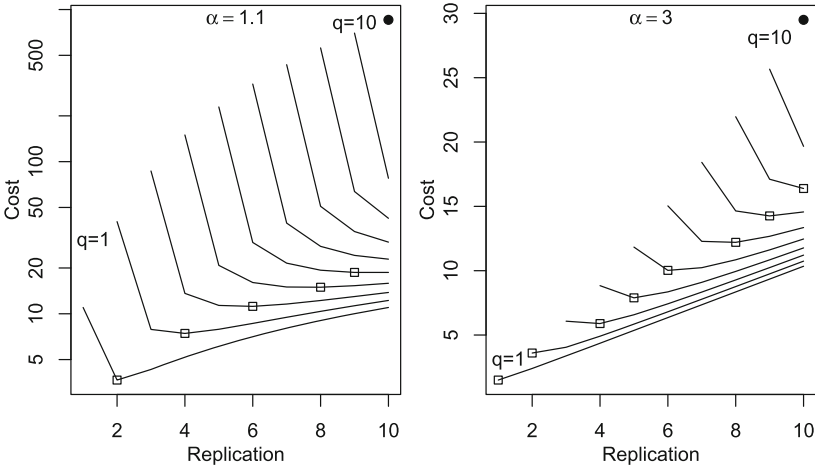


Fig. 2. Cost of computations for a 100-server EDG model depending on replication $r = 1, \dots, 10$ and quorum $q = 1, \dots, r$, with service times having standard Pareto distribution with $\alpha = 1.1$ (left) and $\alpha = 3$ (right). Squares indicate the points $(q, r_0(q))$ such that $C(q, r)$ is increasing in r for $r > r_0(q)$, thus attaining local minimum at $r_0(q)$. Note the logarithmic y axis in the left picture.

2.3 Cost of Computations for Weibull Service Time Distribution

Consider the standard Weibull r.v. S with d.f.

$$F(x) = 1 - e^{-x^\xi}, \quad \xi > 0.$$

It is known that Weibull distribution has a decreasing (increasing) failure rate $f(x)/\bar{F}(x)$ if $\xi < 1$ ($\xi > 1$), and reduces to exponential distribution for $\xi = 1$.

In this case, we have (see for example, [12,20])

$$E(S_{q:r}^k) = q \binom{r}{q} \Gamma(1+k/\xi) \sum_{j=0}^{q-1} (-1)^j \binom{q-1}{j} (r-q+1+j)^{-1-k/\xi}, \quad k > 0. \quad (10)$$

Note that all moments of $S_{q:r}$, as well as original r.v. S , are finite.

Recalling that $C(q, r) = rE(S_{q:r})$, analytical expression for the cost of computations may be obtained. However, it remains unclear if monotonicity properties of $C(q, r)$ for general q and r can be deduced, and we leave this research for a future study. Instead, here we focus on the study of $C(1, r)$. It is easy to verify, from (10), that

$$C(1, r) = \Gamma(1+1/\xi)r^{1-1/\xi}, \quad r \geq 1. \quad (11)$$

Thus, it is easy to see that $C(1, r)$ is increasing (decreasing) in r if $\xi > 1$ ($\xi < 1$). To illustrate the dependence of $C(q, r)$ on q and r , we display the cost of a 100-server EDG by varying $r = 1, \dots, 10$ and $q = 1, \dots, r$ for a standard Weibull distribution with $\xi = 0.5$ and $\xi = 1.5$ in Fig. 3. The nonlinear dependence of the cost on replication and quorum is clearly seen in this case also similar to the Pareto service times. Like in Pareto case, the cost function for Weibull case is also monotone both with increasing q for a fixed r , and with simultaneous increase in q and r . Moreover, for $q = 1$ the cost is monotone in r , as is to be expected.

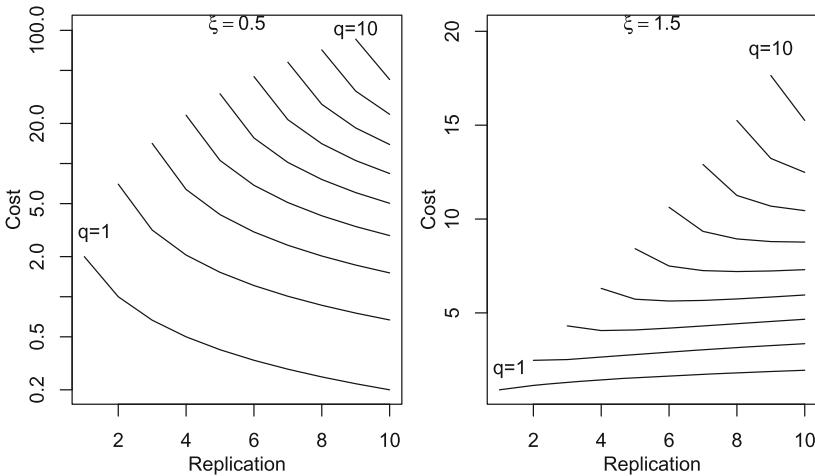


Fig. 3. Cost of computations for a 100-server EDG model depending on replication $r = 1, \dots, 10$ and quorum $q = 1, \dots, r$, with service times having standard Weibull distribution with $\xi = 0.5$ (left) and $\xi = 1.5$ (right). Cost monotonely increases (decreases) for $\xi > 1$ ($\xi < 1$), as expected. Note the logarithmic y axis in the left picture.

3 Moment Properties of EDG Workload

In this section we study the moment indices of workload vector components of EDG model. This is based on the results of [18]. We first briefly summarize the key results from [18] in Sect. 3.1.

3.1 Moment Properties of Multiserver Systems

Consider a classical s -server $G/G/s$ system with general service time distribution and general renewal input flow. Recall the celebrated Kiefer–Wolfowitz recursion for the vector $W_n = (W_{n,1}, \dots, W_{n,s})$ of workload (remaining work) at each server (in ascending order) at n^{th} task arrival epoch t_n is as follows [11]

$$W_{n+1} = R(W_n + eS_n - \mathbf{1}T_n)^+, \tag{12}$$

where $e = (1, 0, \dots, 0)$, $\mathbf{1}$ is the vector of ones, $R(\cdot)$ places vector components in ascending order, and $(x)^+ = \max(0, x)$. We stress that the servers are numbered in the workload ascending order and thus, the number of a server may be changed at each recursion step. Note that $D_n := W_{n,1}$ is the delay of n^{th} task. Denoting S to be a generic service time, and T to be a generic interarrival time, when the stability condition:

$$\rho = E(S)/E(T) < s,$$

holds good, then W_n converges in distribution to $\mathbf{W} := W_\infty$, which is the stationary workload. An important recent result considering the moment properties of vector $\mathbf{W} = (\mathbf{W}_1, \dots, \mathbf{W}_s)$ has been obtained in [18]. This result states that the moment properties of the components of workload vector \mathbf{W} vary with the index (or position) of the components. That is, for index $i \leq \lceil \rho \rceil$ the following result holds good.

$$E\left(S^{1+\beta/(s-\lceil \rho \rceil)}\right) < \infty \quad \Rightarrow \quad E[\mathbf{W}_i]^\beta < \infty. \tag{13}$$

On the contrary, if $i > \lceil \rho \rceil$ (if any), then the moment properties depend on the index of the component:

$$E\left(S^{1+\beta/(s-i+1)}\right) < \infty \quad \Rightarrow \quad E[\mathbf{W}_i]^\beta < \infty. \tag{14}$$

In particular, the following moment property of stationary delay $D = \mathbf{W}_1$ holds [17] good.

$$E\left(S^{1+\beta/(s-\lceil \rho \rceil)}\right) < \infty \quad \Rightarrow \quad E(D^\beta) < \infty. \tag{15}$$

An intuitive explanation of lighter moment conditions (13) resides on the fact that $\lceil \rho \rceil$ servers (note that ρ gives the average number of servers busy when the queue is stable) are in fact enough to *keep the system stable*. On the other hand, it is most probable that servers with indices (or server numbers) $i > \lceil \rho \rceil$ are busy with unevenly large tasks [18]. (We recall that the server numbering in Kiefer–Wolfowitz representation follows the work ascending order.)

3.2 Moment Properties of Workload Vector of EDG Split-Merge Model

Now we consider the EDG model of this paper and apply the results of Sect. 3.1. Recall that our model is equivalent to an $G/G/\lfloor m/r \rfloor$ -type multiserver model with general service times whose distribution function is given by the order statistics $S_{q;r}$. The stability condition here is given by

$$\rho := \lambda E(S_{q;r}) < \lfloor m/r \rfloor, \quad (16)$$

where $E(S_{q;r})$ is as given in (4). Then it follows from (4) and (13) that for a given α, q, r , the moment index β of the workload components \mathbf{W}_i , $i \leq \lceil \rho \rceil$, is of the form:

$$\beta = \alpha(r - q + 1 - 1/\alpha) \left(\lfloor m/r \rfloor - \left\lfloor \lambda x_0 \frac{r!}{(r-q)!} \frac{\Gamma(r-q+1-1/\alpha)}{\Gamma(r+1-1/\alpha)} \right\rfloor \right). \quad (17)$$

and for $i > \lceil \rho \rceil$ (if exists) the moment index is of the form:

$$\beta = \alpha(r - q + 1 - 1/\alpha) (\lfloor m/r \rfloor - i + 1). \quad (18)$$

In particular, the maximal component of the workload vector, $\mathbf{W}_{\lfloor m/r \rfloor}$, has the following moment index which does not depend on the input rate λ :

$$\beta = \alpha(r - q + 1) - 1. \quad (19)$$

Note that when $q = r$, (19) reduces to $\beta = \alpha - 1$, which extends the classical result obtained in [18] for $r = q = 1$ (multiserver system without replication). In Fig. 4 we display the two measures: the maximal input rate, λ (see (16)), to guarantee the stability of the queue, and the moment index, β (see (17)), under various scenarios by varying $r = 1, \dots, 10$ and $q = 1, \dots, r$. When dealing with the moment index, we fix $\lambda = 2$. It is seen that in some cases the stability condition (16) is violated and those scenarios are removed from further consideration. Note that the sources of (non-linear) dependence are the expected service time of a task as well as the (step-wise) dependence of the number of servers $\lfloor m/r \rfloor$ on replication parameter, r . A quick look at Fig. 4 reveals the following observations for the maximal input rate λ that guarantees stability.

- For any replication r , the maximal input rate λ is provided when quorum is set at $q = 1$.
- For a fixed r input rate may be increased by decreasing the quorum.
- For a fixed q the dependence of maximal input rate on r is non-monotone, however, for larger values of q , increasing r leads to an increase in λ .
- Whenever $q = r$ (i.e. quorum is obtained after completion of all r replicas), the maximal input rate monotonically decreases with increasing r .

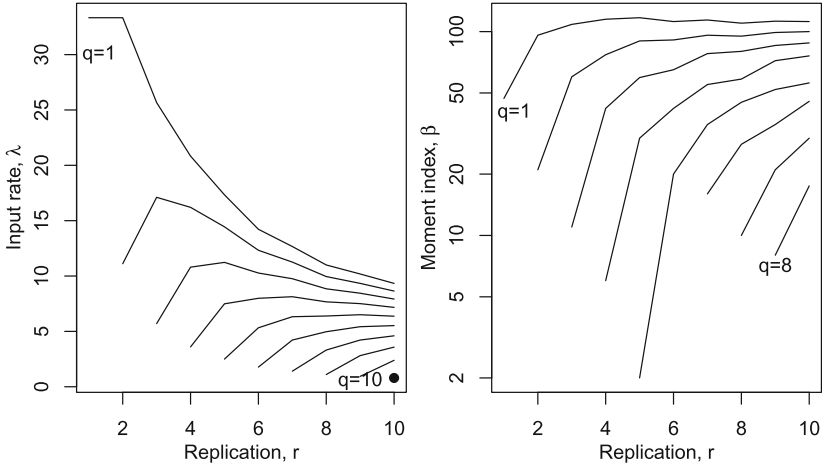


Fig. 4. The maximal input rate λ providing stability (left), and the moment index β of the stationary delay for fixed $\lambda = 2$ (right), for a 100-server EDG model depending on replication $r = 1, \dots, 10$ and quorum $q = 1, \dots, r$, with service times having standard Pareto distribution with $\alpha = 1.5$.

4 Response Time of an EDG Model

In general there is no closed form expression available for waiting time in a multiserver system of $G/G/-$ type. Thus, we need to rely on simulation. However, there are some particular cases that allow explicit expressions. Consider the case when $m = r$ and let the arrivals be modeled by Poisson process. In this case, the system is equivalent to an $M/G/1$ with generic service time $S_{q:r}$. Thus, using the well-known Pollaczek-Khintchine formula, we obtain that the response time $Z(q, r)$ is equal to

$$Z(q, r) = E(S_{q:r}) + \frac{\lambda E(S_{q:r}^2)}{2[1 - \lambda E(S_{q:r})]}. \tag{20}$$

Then the Equations (4) and (10) allow one to deduce explicit expressions for Pareto and Weibull services. The details are omitted for lack of space. Instead, we illustrate the dependence by plotting $C(q, r)$ and $Z(q, r)$ by fixing $r = 10$ and varying q , for both Weibull and Pareto services under various shape parameters. The results are displayed in Fig. 5.

Finally, we illustrate the dependence of $C(q, r)$ and $Z(q, r)$ on $r = 1, \dots, 10$ and $q = 1, \dots, r$, for both Weibull (with $\xi = 1.5$) and Pareto (with $\alpha = 2.5$) services for fixed input rate $\lambda = 5/E(S_{10:10})$ to guarantee a fixed load (half capacity) for most systems. The results are displayed in Fig. 6.

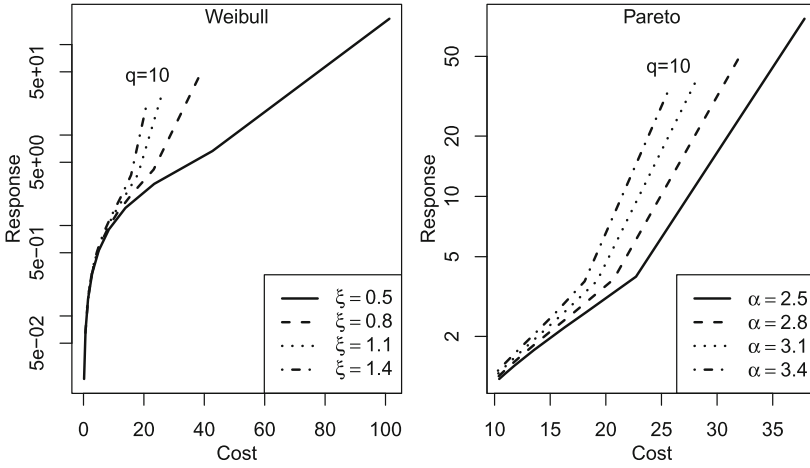


Fig. 5. The cost $C(q, r)$ vs. response time $Z(q, r)$ for Weibull (left) and Pareto (right) distributions with various shape parameters, $q = 1, \dots, 10$ and fixed $m = r = 10$. Note the logarithmic y axis.

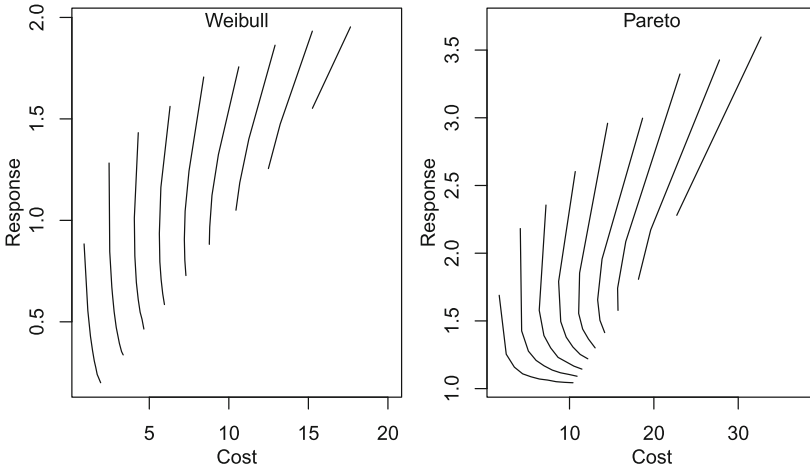


Fig. 6. The cost $C(q, r)$ vs. response time $Z(q, r)$ for Weibull with shape $\xi = 1.5$ (left) and Pareto with shape $\alpha = 2.5$ (right) distributions for $r = 1, \dots, 10$ and $q = 1, \dots, r$ in a 100-server system with load not exceeding 0.5.

5 Conclusion

In this paper, we analyze the effect of replication and quorum in a Split-Merge multiserver model useful in a Desktop Grid computing system. We focus on the heavy-tailed distributions like Pareto and Weibull for service times of tasks and obtain some analytical results using the properties of order statistics. We study effects of replication and quorum parameters on the key performance measures

such as response time and cost of a Desktop Grid system. Moment properties of the workload vector, which play a key role in practical applications, are obtained as well. We illustrate the results through simulation under a variety system configuration and system load.

Acknowledgements. The study was carried out under state order to the Karelian Research Centre of the Russian Academy of Sciences (Institute of Applied Mathematical Research KRC RAS). This research is partially supported by RF President's grant MK-1641.2017.1 and RFBR, projects 16-07-00622, 18-07-00147, 18-07-00156, 18-37-00094.

References

1. Balakrishnan, N., Joshi, P.C.: A note on order statistics from Weibull distribution. *Scand. Actuar. J.* **1981**(2), 121–122 (1981). <https://doi.org/10.1080/03461238.1981.10413737>
2. Balakrishnan, N.: Permanents, order statistics, outliers, and robustness. *Rev. Mat. Complut.* **20**(1), 7–107 (2007). <http://eudml.org/doc/41927>
3. Boxma, O.J., Cohen, J.W.: The M/G/1 queue with heavy-tailed service time distribution. *IEEE J. Sel. Areas Commun.* **16**(5), 749–763 (1998). <https://doi.org/10.1109/49.700910>
4. Chakravarthy, S.R., Rumyantsev, A.: Efficient redundancy techniques in cloud and desktop grid systems using MAP/G/c-type queues. *Open Eng.* **8**(1), 17 (2018). <https://doi.org/10.1515/eng-2018-0004>
5. Ilya, C., Natalia, N., Evgeny, I.: Task scheduling in desktop grids: open problems. *Open Eng.* **7**(1), 343 (2017). <https://doi.org/10.1515/eng-2017-0038>
6. Feitelson, D.G.: *Workload modeling for computer systems performance evaluation*. Cambridge University Press, Cambridge (2015). <https://doi.org/10.1017/CBO9781139939690>
7. Foss, S.G., Korshunov, D., Zachary, S.: *An Introduction to Heavy-Tailed and Subexponential Distributions*. Operations Research and Financial Engineering. Springer, New York (2011). <https://doi.org/10.1007/978-1-4614-7101-1>
8. Huang, J.S.: A note on order statistics from Pareto distribution. *Scand. Actuar. J.* **1975**(3), 187–190 (1975). <https://doi.org/10.1080/03461238.1975.10405095>
9. Ivashko, E.: Enterprise desktop grids. In: *Proceedings of the Second International Conference BOINC-based High Performance Computing: Fundamental Research and Development (BOINC:FAST 2015)*, CEUR Workshop Proceedings, vol. 1502, pp. 16–21 (2015)
10. Joshi, G.: *Efficient redundancy techniques to reduce delay in Cloud systems*. Ph.D. thesis, Massachusetts Institute of Technology (2016). <https://dspace.mit.edu/handle/1721.1/105944>
11. Kiefer, J., Wolfowitz, J.: On the theory of queues with many servers. *Trans. Am. Math. Soc.* **78**, 1–18 (1955). <http://www.jstor.org/stable/1992945>
12. Lieblein, J.: On moments of order statistics from the Weibull distribution. *Ann. Math. Stat.* **26**(2), 330–333 (1955). <https://doi.org/10.1214/aoms/1177728551>
13. Malik, H.J.: Exact moments of order statistics from the Pareto distribution. *Scand. Actuar. J.* **1966**(3–4), 144–157 (1966). <https://doi.org/10.1080/03461238.1966.10404562>

14. Ramaswami, V., Jain, K., Jana, R., Aggarwal, V.: Modeling heavy tails in traffic sources for network performance evaluation. In: Krishnan, G.S.S., Anitha, R., Lekshmi, R.S., Kumar, M.S., Bonato, A., Graña, M. (eds.) Computational Intelligence, Cyber Security and Computational Models. AISC, vol. 246, pp. 23–44. Springer, New Delhi (2014). https://doi.org/10.1007/978-81-322-1680-3_4
15. Rumyantsev, A., Chakravarthy, S.: Split-merge model of workunit replication in distributed computing. In: Proceedings of the Third International Conference BOINC:FAST 2017, CEUR-WS, vol. 1973, pp. 27–34 (2017). <http://ceur-ws.org/Vol-1973/paper03.pdf>
16. Samorodnitsky, G.: Long range dependence. *Found. Trends Stoch. Syst.* **1**(3), 163–257 (2006). <https://doi.org/10.1561/0900000004>
17. Scheller-Wolf, A., Vesilo, R.: Structural interpretation and derivation of necessary and sufficient conditions for delay moments in FIFO multiserver queues. *Queueing Syst.* **54**(3), 221–232 (2006). <https://doi.org/10.1007/s11134-006-0068-1>
18. Scheller-Wolf, A., Vesilo, R.: Sink or swim together: necessary and sufficient conditions for finite moments of workload components in FIFO multiserver queues. *Queueing Syst.* **67**(1), 47–61 (2011). <https://doi.org/10.1007/s11134-010-9198-6>
19. Sigman, K.: Appendix: a primer on heavy-tailed distributions. *Queueing Syst.* **33**(1), 261–275 (1999). <https://doi.org/10.1023/A:1019180230133>
20. Sultan, K.S., Moshref, M.E.: Moments of order statistics from Weibull distribution in the presence of multiple outliers. *Commun. Stat. - Theor. Methods* **43**(10–12), 2214–2226 (2014). <https://doi.org/10.1080/03610926.2013.783072>



Optimal Estimation of the States of Synchronous Generalized Flow of Events of the Second Order Under Its Complete Observability

Luydmila Nezhelskaya and Ekaterina Sidorova^(✉)

National Research Tomsk State University,
Lenina Avenue 36, 634050 Tomsk, Russia
ludne@mail.ru, katusha_sidorova@mail.ru

Abstract. We consider the optimal estimation problem of the states of synchronous generalized flow of events of the second order with two states; it is one of the mathematical models for an incoming stream of claims (events) in digital integral servicing networks and which is related to the class of Markov chains. The observation conditions for this flow are such that each event is accessible to observation. We offer the optimal estimation algorithm for the flow states, where the decision about the flow state is made by criterion of a posteriori probability maximum. The results of the analytical calculations of a posteriori probability and the simulation experiments with numerical results are presented.

Keywords: Synchronous generalized flow of events of the second order
Doubly stochastic flows · States estimation
Accompanying random process · A posteriori probability
Criterion of a posteriori probability maximum

1 Introduction

Recently, when analyzing the situations that arise in economic and logistical spheres, as well as other spheres of human activity related to organization of planning and operation of production and consumption processes, to design of functioning process of automated control systems, to operation of electronic computing systems, terminals, transport networks, to technical and military equipment, it is often to use the mathematical apparatus of queueing theory, as one of the most intensively developing sections of the theory of random processes, the subject of which, in particular, are incoming streams of queueing systems. Random incoming flows of events, as the main elements of queueing systems, are widely used as mathematical models of real information flows of claims (events) in telecommunication systems, global computer networks, satellite communication networks [1, 2].

Due to the rapid evolution of digital integral servicing systems, the use of mathematical models of incoming flows of events in the Poisson flows form has

become difficult due to the inapplicability of such models to describe information flows. Thus, it became necessary to construct new mathematical models of doubly stochastic flows of events [3–6], adequately describing real information flows. For such flows, firstly, the moments of occurrence of events are random, and secondly, the intensity is a fundamentally unobservable random process. Doubly stochastic flows can be divided into two classes. The first class includes flows whose intensity is a continuous random process, the second – flows whose intensity is a piecewise constant random process with a finite number of states. The latter are called MC (Markov chain) flows or MAP (Markovian Arrival Process) flows [7, 8] and are the most used for solving applied problems. A generalization of the MAP-flows of events is given in [9], the connection of MC-flows and MAP-flows is established in [10].

A further research of synchronous generalized flow of events of the second order, begun in [11], is carried out in this paper. For the considering flow, which belongs to the class of MC-flows, analytical and numerical results of the optimal states estimation are given. An optimal estimation algorithm for the flow states is proposed. The decision about the flow state is made according to criterion of a posteriori probability maximum, which is the most complete characteristic of the flow states that can be received from observations of the flow. The criterion also minimizes the total (unconditional) probability of making a wrong decision [12].

A number of statistical experiments were carried out on the simulation model of synchronous generalized flow of events of the second order constructed in [11] and modified in this work to obtain numerical results of the estimation.

2 Problem Statement

We consider a synchronous generalized flow of events of the second order (flow of events), accompanying random process $\lambda(t)$ of which, is an unobservable piecewise constant process with two states S_1 and S_2 . Hereinafter, it is understood the i th state of process $\lambda(t)$ as the state S_i , $i = 1, 2$.

The duration of interval between the flow events at the i th state is determined by random variable $\eta_i = \min(\xi_i^{(1)}, \xi_i^{(2)})$, where random variable $\xi_i^{(1)}$ is distributed according to the law $F_i^{(1)}(t) = 1 - e^{-\lambda_i t}$, random variable $\xi_i^{(2)}$ is distributed according to the law $F_i^{(2)}(t) = 1 - e^{-\alpha_i t}$; $\xi_i^{(1)}$ and $\xi_i^{(2)}$ are independent random variables, $i = 1, 2$. At the moment when a flow event occurs, process $\lambda(t)$ transits from the i th state to the j th either with probability $P_1^{(1)}(\lambda_j|\lambda_i)$, or with probability $P_1^{(2)}(\lambda_j|\lambda_i)$ depending on the value, which random variable η_i has taken, $i, j = 1, 2$, $i \neq j$. At the moment when a flow event occurs, process $\lambda(t)$ stays at the i th state either with probability $P_1^{(1)}(\lambda_i|\lambda_i)$, or with probability $P_1^{(2)}(\lambda_i|\lambda_i)$ depending on the value, which random variable η_i has taken, $i = 1, 2$. Here $P_1^{(1)}(\lambda_j|\lambda_i) + P_1^{(1)}(\lambda_i|\lambda_i) = 1$, $P_1^{(2)}(\lambda_j|\lambda_i) + P_1^{(2)}(\lambda_i|\lambda_i) = 1$, $i, j = 1, 2$, $i \neq j$. Thus, the duration of interval between the flow events at the i th state of process $\lambda(t)$ is a random variable with the distribution function $F_i(t) = 1 - e^{-(\lambda_i + \alpha_i)t}$, $i = 1, 2$.

In the sequel it is assumed that the state S_1 (the first state) of process $\lambda(t)$ takes place, if $\lambda(t) = \lambda_1$, and the state S_2 (the second state) of process $\lambda(t)$ takes place, if $\lambda(t) = \lambda_2$ ($\lambda_1 > \lambda_2 \geq 0$).

Proposition. *For synchronous generalized flow of events of the second order a piecewise constant random process $\lambda(t)$ is a Markov process.*

Indeed, it is not difficult to show that the duration of process $\lambda(t)$ remains at the i th state, $i = 1, 2$, is a random variable with the exponential distribution function of the form $F_i(t) = 1 - e^{-(\lambda_i P_1^{(1)}(\lambda_j|\lambda_i) + \alpha_i P_1^{(2)}(\lambda_j|\lambda_i))t}$, $i = 1, 2, j = 1, 2, i \neq j$. Consequently, process $\lambda(t)$ is a Markovian.

The infinitesimal characteristics matrices for the process $\lambda(t)$ have the form

$$D_0 = \begin{vmatrix} -(\lambda_1 + \alpha_1) & 0 \\ 0 & -(\lambda_2 + \alpha_2) \end{vmatrix},$$

$$D_1 = \begin{vmatrix} \lambda_1 P_1^{(1)}(\lambda_1|\lambda_1) + \alpha_1 P_1^{(2)}(\lambda_1|\lambda_1) & \lambda_1 P_1^{(1)}(\lambda_2|\lambda_1) + \alpha_1 P_1^{(2)}(\lambda_2|\lambda_1) \\ \lambda_2 P_1^{(1)}(\lambda_1|\lambda_2) + \alpha_2 P_1^{(2)}(\lambda_1|\lambda_2) & \lambda_2 P_1^{(1)}(\lambda_2|\lambda_2) + \alpha_2 P_1^{(2)}(\lambda_2|\lambda_2) \end{vmatrix}.$$

Diagonal elements of the matrix D_0 are the intensities of the process $\lambda(t)$ output from its states taken with the opposite sign; off-diagonal elements are the intensities of the transitions from state to state without an event occurrence. Elements of the matrix D_1 are the intensities of the process $\lambda(t)$ transitions from state to state upon a flow event occurs.

An example of the arising situation is shown in Fig. 1, where S_1, S_2 are states of the random process $\lambda(t)$, t_1, t_2, \dots denote the moments when events occur in the considering flow.

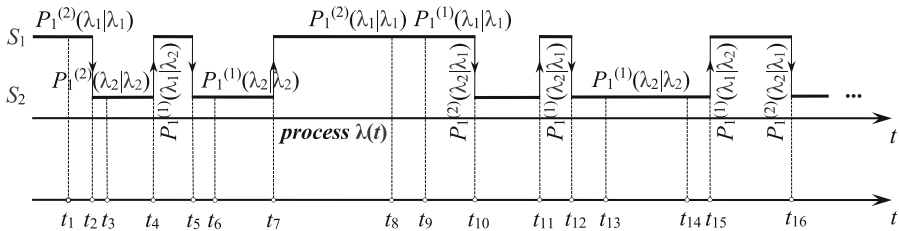


Fig. 1. Formation of synchronous generalized flow of events of the second order

Due to the fact that the process $\lambda(t)$ is fundamentally unobservable (latent Markov process) and only the time moments of occurrence of events t_1, t_2, \dots are observed, it is necessary to estimate the process $\lambda(t)$ (or flow) state at the end time of the period of observations using only these observable instants of time.

We consider the stationary operation mode of the flow of events, that is why at the interval of observations (t_0, t) , where t_0 – is an instant of the beginning the observations, t – is an instant of its ending or the moment of making a decision about the process $\lambda(t)$ state, we neglect transient processes. Thus, without loss of generality, in steady-state conditions we may take $t_0 = 0$.

To make a decision regarding to the state of the unobservable stationary random process $\lambda(t)$ at the time moment t , it is necessary to determine a posteriori probabilities $w(\lambda_i|t) = w(\lambda_i|t_1, \dots, t_m, t)$, $i = 1, 2$ (m is the number of events that occurred during the time period of duration t), that at the moment of making a decision t the process value $\lambda(t) = \lambda_i$, $i = 1, 2$, here obviously $w(\lambda_1|t) + w(\lambda_2|t) = 1$. The decision about the process $\lambda(t)$ state is made by comparing a posteriori probabilities: if $w(\lambda_i|t) \geq w(\lambda_j|t)$, $i, j = 1, 2, i \neq j$, then estimation of the process state is $\hat{\lambda}(t) = \lambda_i$, $i = 1, 2$, otherwise – $\hat{\lambda}(t) = \lambda_j$, $j = 1, 2$.

3 The Optimal Estimation Algorithm for the States of Synchronous Generalized Flow of Events of the Second Order

The moment of making a decision t about the process $\lambda(t)$ state belongs to some interval (t_k, t_{k+1}) , $k = 1, 2, \dots$, between neighboring events of the considering flow. For the interval (t_0, t_1) the moment t is between the instant of the beginning the observations t_0 and the time moment when the first flow event occurs t_1 .

To derive the equations for a posteriori probability $w(\lambda_1|t)$, we use the technique described in [12].

Let a time t changes discretely with step Δt : $t^{(k)} = k\Delta t$, $k = 0, 1, \dots$, at the interval of observations $(0, t)$. We introduce a bivariate random process $(\lambda^{(k)}, r_k)$, where $\lambda^{(k)} = \lambda(k\Delta t)$ is a value of the process $\lambda(t)$ at the time moment $t^{(k)} = k\Delta t$ ($\lambda^{(k)} = \lambda_i$, $i = 1, 2$); $r_k = r[k\Delta t] - r[(k-1)\Delta t]$ is a number of flow events occurred at the interval $((k-1)\Delta t, k\Delta t)$ of duration Δt , $r_k = 0, 1, \dots$. Denote by $\mathbf{r}_m = (r_0, r_1, \dots, r_m)$ the sequence of a number of events occurred during a time period from 0 to $m\Delta t$ at the intervals $((k-1)\Delta t, k\Delta t)$ of duration Δt , $k = \overline{0, m}$. Here r_0 is a number of events occurred at the interval $(-\Delta t, 0)$. It is supposed that $r_0 = 0$ since there was no observation of the flow at that interval. Denote by $\boldsymbol{\lambda}^{(m)} = (\lambda^{(0)}, \lambda^{(1)}, \dots, \lambda^{(m)})$ the sequence of unknown values of the process $\lambda(k\Delta t)$ at the time moment $k\Delta t$, $k = \overline{0, m}$ ($\lambda^{(0)} = \lambda(0) = \lambda_i$, $i = 1, 2$).

Let us introduce $w(\lambda^{(m)}|\mathbf{r}_m)$ – the conditional probability of a value $\lambda^{(m)}$ under condition of observability the implementation \mathbf{r}_m . Similarly, define the probability $w(\lambda^{(m+1)}|\mathbf{r}_{m+1})$. The process $(\lambda^{(k)}, r_k)$ is a Markov process. Then for the Markov random process, as shown in [13], the recurrence relation for the introduced probabilities $w(\lambda^{(m)}|\mathbf{r}_m)$, $w(\lambda^{(m+1)}|\mathbf{r}_{m+1})$ is valid

$$w(\lambda^{(m+1)}|\mathbf{r}_{m+1}) = \frac{\sum_{\lambda^{(m)}=\lambda_1}^{\lambda_2} w(\lambda^{(m)}|\mathbf{r}_m)p(\lambda^{(m+1)}, r_{m+1}|\lambda^{(m)}, r_m)}{\sum_{\lambda^{(m)}=\lambda_1}^{\lambda_2} \sum_{\lambda^{(m+1)}=\lambda_1}^{\lambda_2} w(\lambda^{(m)}|\mathbf{r}_m)p(\lambda^{(m+1)}, r_{m+1}|\lambda^{(m)}, r_m)}, \quad (1)$$

where $p(\lambda^{(m+1)}, r_{m+1} | \lambda^{(m)}, r_m)$ is the probability of the process $(\lambda^{(k)}, r_k)$ transition from state $(\lambda^{(m)}, r_m)$ to state $(\lambda^{(m+1)}, r_{m+1})$ in one step Δt ; $w(\lambda^{(m)} | \mathbf{r}_m) = w(\lambda^{(m)} | t)$, $w(\lambda^{(m+1)} | \mathbf{r}_{m+1}) = w(\lambda^{(m+1)} | t + \Delta t)$.

For synchronous generalized flow of events of the second order the transition probability $p(\lambda^{(m+1)}, r_{m+1} | \lambda^{(m)}, r_m)$, which is included into formula (1), can be written as $p(\lambda^{(m+1)}, r_{m+1} | \lambda^{(m)}, r_m) = p(\lambda^{(m+1)} | \lambda^{(m)})p(r_{m+1} | \lambda^{(m)}, \lambda^{(m+1)})$. Then the recurrence relation (1) takes the form

$$\begin{aligned} & w(\lambda^{(m+1)} | t + \Delta t) = \\ &= \frac{\sum_{\lambda^{(m)}=\lambda_1}^{\lambda_2} w(\lambda^{(m)} | t) p(\lambda^{(m+1)} | \lambda^{(m)}) p(r_{m+1} | \lambda^{(m)}, \lambda^{(m+1)})}{\sum_{\lambda^{(m)}=\lambda_1}^{\lambda_2} \sum_{\lambda^{(m+1)}=\lambda_1}^{\lambda_2} w(\lambda^{(m)} | t) p(\lambda^{(m+1)} | \lambda^{(m)}) p(r_{m+1} | \lambda^{(m)}, \lambda^{(m+1)})}. \end{aligned} \quad (2)$$

Remark 1. Based on the definition of synchronous generalized flow of events of the second order, the variable r_k can take only two values: $r_k = 0$ or $r_k = 1$. The probability that $r_k = 2, 3, \dots$ is $o(\Delta t)$.

Let in (2) $r_{m+1} = 0$, which corresponds to the case of the absence of flow events at the interval $(t, t + \Delta t)$, where $t = m\Delta t$, $t + \Delta t = (m + 1)\Delta t$. Taking into account the matrix \mathbf{D}_0 , we write down the transition probabilities in (2) in the following form

$$\begin{aligned} & p(\lambda^{(m+1)} = \lambda_i | \lambda^{(m)} = \lambda_i) p(r_{m+1} = 0 | \lambda^{(m)} = \lambda_i, \lambda^{(m+1)} = \lambda_i) = \\ &= p(r_{m+1} = 0, \lambda^{(m+1)} = \lambda_i | \lambda^{(m)} = \lambda_i) = 1 - (\lambda_i + \alpha_i) \Delta t + o(\Delta t), \quad i = 1, 2; \quad (3) \\ & p(\lambda^{(m+1)} = \lambda_j | \lambda^{(m)} = \lambda_i) p(r_{m+1} = 0 | \lambda^{(m)} = \lambda_i, \lambda^{(m+1)} = \lambda_j) = \\ &= p(r_{m+1} = 0, \lambda^{(m+1)} = \lambda_j | \lambda^{(m)} = \lambda_i) = 0, \quad i, j = 1, 2, \quad i \neq j. \end{aligned}$$

For definiteness, in (2) we set $\lambda^{(m+1)} = \lambda_1$. Then the following lemma holds.

Lemma 1. *At time intervals (t_0, t_1) and (t_k, t_{k+1}) , $k = 1, 2, \dots$, between neighboring events of the flow a posteriori probability $w(\lambda_1 | t)$ satisfies a Bernoulli differential equation*

$$\frac{dw(\lambda_1 | t)}{dt} = w^2(\lambda_1 | t)(\lambda_1 + \alpha_1 - \lambda_2 - \alpha_2) - w(\lambda_1 | t)(\lambda_1 + \alpha_1 - \lambda_2 - \alpha_2), \quad (4)$$

$$\lambda_1 + \alpha_1 - \lambda_2 - \alpha_2 \neq 0.$$

Proof. Substituting (3) into (2), performing the necessary transformations and proceeding to the limit $\Delta t \rightarrow 0$, we obtain a Bernoulli differential eq. (4). Lemma is proved.

Suppose that $r_{m+1} = 1$ in (2); this corresponds to the case of the observing one event of the flow at the time interval $(t, t + \Delta t)$. In (2) we set $\lambda^{(m+1)} = \lambda_1$. We divide the time interval $(t, t + \Delta t)$ into two adjacent intervals (t, t_k) and $(t_k, t + \Delta t)$ of durations $\Delta t' = t_k - t$ and $\Delta t'' = t + \Delta t - t_k$, respectively (Fig. 2).

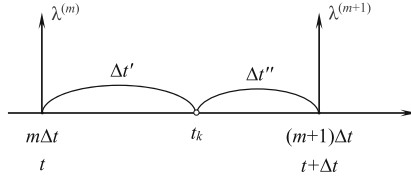


Fig. 2. The dividing the time interval

Then under the made assumptions (2) takes the form

$$\begin{aligned}
 & w(\lambda_1|t_k + \Delta t'') = \\
 &= \frac{\sum_{\lambda^{(m)}=\lambda_1}^{\lambda_2} w(\lambda^{(m)}|t_k-\Delta t')p(\lambda^{(m+1)}=\lambda_1|\lambda^{(m)})p(r_{m+1}=1|\lambda^{(m)},\lambda^{(m+1)}=\lambda_1)}{\sum_{\lambda^{(m)}=\lambda_1}^{\lambda_2} \sum_{\lambda^{(m+1)}=\lambda_1}^{\lambda_2} w(\lambda^{(m)}|t_k-\Delta t')p(\lambda^{(m+1)}|\lambda^{(m)})p(r_{m+1}=1|\lambda^{(m)},\lambda^{(m+1)})}. \tag{5}
 \end{aligned}$$

For synchronous generalized flow of events of the second order Lemma 2 is valid.

Lemma 2. *At the moment t_k when a flow event occurs for a posteriori probability of the first state of the process $\lambda(t)$ the conversion formula takes place*

$$w(\lambda_1|t_k + 0) = \frac{W}{(\lambda_2 + \alpha_2) + w(\lambda_1|t_k - 0)(\lambda_1 + \alpha_1 - \lambda_2 - \alpha_2)}, k = 1, 2, \dots, \tag{6}$$

$$\begin{aligned}
 & W = \lambda_2 P_1^{(1)}(\lambda_1|\lambda_2) + \alpha_2 P_1^{(2)}(\lambda_1|\lambda_2) + \\
 & + w(\lambda_1|t_k - 0)[\lambda_1 P_1^{(1)}(\lambda_1|\lambda_1) + \alpha_1 P_1^{(2)}(\lambda_1|\lambda_1) - \lambda_2 P_1^{(1)}(\lambda_1|\lambda_2) - \alpha_2 P_1^{(2)}(\lambda_1|\lambda_2)],
 \end{aligned}$$

$$w(\lambda_2|t_k - 0) = 1 - w(\lambda_1|t_k - 0).$$

Proof. With considering the matrix \mathbf{D}_1 , we receive expressions for the transition probabilities, which are included into (5), in the form

$$\begin{aligned}
 & p(\lambda^{(m+1)} = \lambda_i|\lambda^{(m)} = \lambda_i)p(r_{m+1} = 1|\lambda^{(m)} = \lambda_i, \lambda^{(m+1)} = \lambda_i) = \\
 & = p(r_{m+1} = 1, \lambda^{(m+1)} = \lambda_i|\lambda^{(m)} = \lambda_i) = \\
 & = (\lambda_i P_1^{(1)}(\lambda_i|\lambda_i) + \alpha_i P_1^{(2)}(\lambda_i|\lambda_i))\Delta t + o(\Delta t), i = 1, 2; \\
 & p(\lambda^{(m+1)} = \lambda_j|\lambda^{(m)} = \lambda_i)p(r_{m+1} = 1|\lambda^{(m)} = \lambda_i, \lambda^{(m+1)} = \lambda_j) = \\
 & = p(r_{m+1} = 1, \lambda^{(m+1)} = \lambda_j|\lambda^{(m)} = \lambda_i) = \\
 & = (\lambda_i P_1^{(1)}(\lambda_j|\lambda_i) + \alpha_i P_1^{(2)}(\lambda_j|\lambda_i))\Delta t + o(\Delta t), i, j = 1, 2, i \neq j. \tag{7}
 \end{aligned}$$

Substituting (7) into (5), taking into account $w(\lambda_2|t_k - \Delta t') = 1 - w(\lambda_1|t_k| - \Delta t')$, and proceeding to the limit $\Delta t \rightarrow 0$ ($\Delta t' \rightarrow 0$ and $\Delta t'' \rightarrow 0$ simultaneously), we obtain the assertion of the lemma. Lemma is proved.

Remark 2. At the time moment t_k , $k = 1, 2, \dots$, of an event occurrence a posteriori probability $w(\lambda_1|t)$ undergoes a discontinuity of the first kind, i.e. a finite jump takes place. The probability $w(\lambda_1|t_k + 0)$ depends on the value of $w(\lambda_1|t_k - 0)$, where $w(\lambda_1|t_k - 0)$ is a value of the probability $w(\lambda_1|t)$ determined by eq. (4) at the time $t = t_k$ when t changes at the interval of time (t_{k-1}, t_k) adjacent to the interval (t_k, t_{k+1}) , $k = 1, 2, \dots$. Thereby, the whole prehistory of the flow observations starting from the time moment $t_0 = 0$ (the beginning of the observations) to the moment t_k (the moment of event occurrence) is concentrated in the value $w(\lambda_1|t_k + 0)$.

Let us denote $\pi_i(t|t^0)$ an a priori probability that $\lambda(t) = \lambda_i$ at the time moment t , $i = 1, 2$, provided that functioning of the flow started at the time t^0 .

Lemma 3. *A priori probabilities $\pi_i(t|t^0)$, $i = 1, 2$ of the process $\lambda(t)$ states for synchronous generalized flow of events of the second order satisfy a system of linear differential equations*

$$\begin{cases} \pi'_1(t|t^0) = -(\lambda_1 + \alpha_1)\pi_1(t|t^0) + \pi_1(t|t^0)(\lambda_1 P_1^{(1)}(\lambda_1|\lambda_1) + \\ \quad + \alpha_1 P_1^{(2)}(\lambda_1|\lambda_1)) + \pi_2(t|t^0)(\lambda_2 P_1^{(1)}(\lambda_1|\lambda_2) + \alpha_2 P_1^{(2)}(\lambda_1|\lambda_2)), \\ \pi'_2(t|t^0) = -(\lambda_2 + \alpha_2)\pi_2(t|t^0) + \pi_2(t|t^0)(\lambda_2 P_1^{(1)}(\lambda_2|\lambda_2) + \\ \quad + \alpha_2 P_1^{(2)}(\lambda_2|\lambda_2)) + \pi_1(t|t^0)(\lambda_1 P_1^{(1)}(\lambda_2|\lambda_1) + \alpha_1 P_1^{(2)}(\lambda_2|\lambda_1)). \end{cases} \quad (8)$$

Proof is carried out using the Δt -method.

Lemma 4. *A priori probabilities of the states of the process $\lambda(t)$ for synchronous generalized flow of events of the second order have the form*

$$\begin{cases} \pi_1(t|t^0) = \frac{b}{a} + (\pi - \frac{b}{a})e^{-a(t-t^0)}, \\ \pi_2(t|t^0) = \frac{a-b}{a} - (\pi - \frac{b}{a})e^{-a(t-t^0)}, \end{cases} \quad (9)$$

$$a = \lambda_1 P_1^{(1)}(\lambda_2|\lambda_1) + \alpha_1 P_1^{(2)}(\lambda_2|\lambda_1) + \lambda_2 P_1^{(1)}(\lambda_1|\lambda_2) + \alpha_2 P_1^{(2)}(\lambda_1|\lambda_2),$$

$$b = \lambda_2 P_1^{(1)}(\lambda_1|\lambda_2) + \alpha_2 P_1^{(2)}(\lambda_1|\lambda_2); \pi = \pi_1(t^0|t^0).$$

Proof. Integrating system (8) and using the initial conditions, according to which at the time moment $t = t^0$ $\pi_1(t^0|t^0) = \pi$, $\pi_2(t^0|t^0) = 1 - \pi$, we arrive at (9). Lemma is proved.

Corollary. *For synchronous generalized flow of events of the second order a priori final probabilities of the states of the process $\lambda(t)$ when $t \rightarrow \infty$ (or $t^0 \rightarrow -\infty$) have the following form*

$$\begin{cases} \pi_1 = \frac{\lambda_2 P_1^{(1)}(\lambda_1|\lambda_2) + \alpha_2 P_1^{(2)}(\lambda_1|\lambda_2)}{\lambda_1 P_1^{(1)}(\lambda_2|\lambda_1) + \alpha_1 P_1^{(2)}(\lambda_2|\lambda_1) + \lambda_2 P_1^{(1)}(\lambda_1|\lambda_2) + \alpha_2 P_1^{(2)}(\lambda_1|\lambda_2)}, \\ \pi_2 = \frac{\lambda_1 P_1^{(1)}(\lambda_2|\lambda_1) + \alpha_1 P_1^{(2)}(\lambda_2|\lambda_1)}{\lambda_1 P_1^{(1)}(\lambda_2|\lambda_1) + \alpha_1 P_1^{(2)}(\lambda_2|\lambda_1) + \lambda_2 P_1^{(1)}(\lambda_1|\lambda_2) + \alpha_2 P_1^{(2)}(\lambda_1|\lambda_2)}. \end{cases} \quad (10)$$

The results of Lemmas 1 and 2 allow us to formulate the theorem.

Theorem. *A posteriori probability $w(\lambda_1|t)$ behavior at the time intervals (t_0, t_1) and (t_k, t_{k+1}) , $k = 1, 2, \dots$, is determined by the explicit formula*

$$w(\lambda_1|t) = \frac{w(\lambda_1|t_k + 0)e^{-(\lambda_1 + \alpha_1 - \lambda_2 - \alpha_2)(t - t_k)}}{1 - w(\lambda_1|t_k + 0) + w(\lambda_1|t_k + 0)e^{-(\lambda_1 + \alpha_1 - \lambda_2 - \alpha_2)(t - t_k)}}, \quad (11)$$

$t_k < t < t_{k+1}$, $k = 0, 1, \dots$; $w(\lambda_1|t_k + 0)$, $k = 1, 2, \dots$, is defined in (6), $w(\lambda_1|t_0 + 0) = \pi_1$, π_1 is given by (10).

Proof is carried out by integrating eq. (4) using the conversion formula (6).

The analytical formulas obtained for $w(\lambda_1|t)$ allow us to formulate the algorithm for a posteriori probability $w(\lambda_1|t)$ calculation, as well as the algorithm for making a decision about the process $\lambda(t)$ state at any time moment t , i.e. the optimal estimation algorithm for the states of synchronous generalized flow of events of the second order:

- (1) at the initial instant $t_0 = 0$, a priori probability of the first state π_1 of the process $\lambda(t)$ is computed according to formula (10) and we put $w(\lambda_1|t_0 + 0) = w(\lambda_1|t_0 = 0) = \pi_1$;
- (2) at any time moment t , $t_0 < t < t_1$, where t_1 is the moment of observation of the first flow event, calculations are made for determination of a posteriori probability $w(\lambda_1|t)$ according to formula (11) for $k = 0$;
- (3) at the moment t_1 the probability $w(\lambda_1|t)$ is calculated according to formula (11) for $k = 0$, i.e. $w(\lambda_1|t_1) = w(\lambda_1|t_1 - 0)$, then, k is incremented, and for $k = 1$, according to formula (6) a posteriori probability is recalculated at the time moment t_1 , $w(\lambda_1|t_1 + 0)$ is the initial value for $w(\lambda_1|t)$ at the next step;
- (4) for $k = 1$, a posteriori probability $w(\lambda_1|t)$ is calculated according to formula (11) at any time moment t , $t_1 < t < t_2$, where t_2 is the moment of observation of the second flow event;
- (5) at the time moment t_2 , for $k = 1$, $w(\lambda_1|t_2) = w(\lambda_1|t_2 - 0)$ is calculated according to formula (11), further, k is increased by one, a posteriori probability is recalculated according to formula (6) for $k = 2$ at the time moment t_2 , $w(\lambda_1|t_2 + 0)$ is the initial value for $w(\lambda_1|t)$ at the next step of the algorithm, etc.

In parallel, during the process of calculating the probability $w(\lambda_1|t)$, a decision about the process $\lambda(t)$ state at any time moment t is made according to criterion of a posteriori probability maximum: if $w(\lambda_1|t) \geq w(\lambda_2|t)$, then estimation of the process state is $\hat{\lambda}(t) = \lambda_1$, otherwise $\hat{\lambda}(t) = \lambda_2$.

4 Results of Numerical Experiments

To obtain numerical results, the algorithm for calculating a posteriori probability $w(\lambda_1|t)$ has developed using formulas (6), (10), (11). The program is

implemented by C# programming language in Microsoft Visual Studio 2013 environment. The first stage of calculations assumes the simulation [14] of synchronous generalized flow of events of the second order (the description of the simulation algorithm is not given here due to the fact that the algorithm does not contain any special difficulties). The second stage is a direct computation of a posteriori probabilities $w(\lambda_1|t)$, $t_0 < t < t_1$; $w(\lambda_1|t_k + 0)$ and $w(\lambda_1|t)$, $t_k < t < t_{k+1}$, $k = 1, 2, \dots$, and determination of estimates $\hat{\lambda}(t)$ (according to method of a posteriori probability maximum) using the obtained sample t_1, t_2, \dots of the moments of occurrence of events in the observed event flow.

For the values of the flow parameters given in Table 1 and modeling time $T = 100$ time units, the calculations were made to find the estimate $\hat{\lambda}(t)$.

Table 1. Initial data

$\lambda_1 = 3$	$P_1^{(1)}(\lambda_1 \lambda_1) = 0,4$	$P_1^{(1)}(\lambda_2 \lambda_1) = 0,6$
$\lambda_2 = 1$	$P_1^{(1)}(\lambda_2 \lambda_2) = 0,5$	$P_1^{(1)}(\lambda_1 \lambda_2) = 0,5$
$\alpha_1 = 0,5$	$P_1^{(2)}(\lambda_1 \lambda_1) = 0,3$	$P_1^{(2)}(\lambda_2 \lambda_1) = 0,7$
$\alpha_2 = 0,8$	$P_1^{(2)}(\lambda_2 \lambda_2) = 0,2$	$P_1^{(2)}(\lambda_1 \lambda_2) = 0,8$

In Fig. 3, as an illustration, a trajectory of the random process $\lambda(t)$ (an actual path of the process) obtained as a result of simulation, is shown, here λ_1, λ_2 are the process $\lambda(t)$ values at the states S_1 and S_2 .

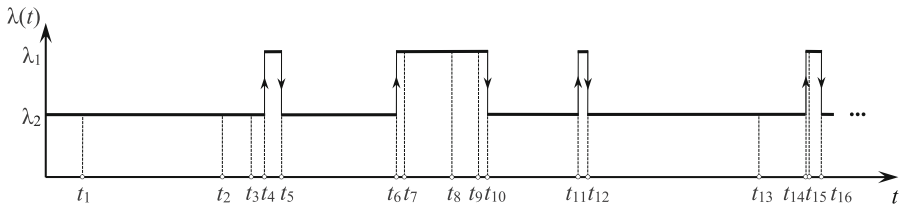


Fig. 3. Implementation of synchronous generalized flow of events of the second order

A trajectory of a posteriori probability $w(\lambda_1|t)$ behavior, which corresponds to the sequence of the moments of events occurrence t_1, t_2, \dots obtained by simulation modeling, is shown in Fig. 4.

Figure 5 shows a trajectory of the estimation $\hat{\lambda}(t)$ of the process $\lambda(t)$, where λ_1, λ_2 are the estimation $\hat{\lambda}(t)$ values. The decision about the process $\lambda(t)$ state was made with the step $\Delta t = 0,001$. Time intervals where the values of the estimation $\hat{\lambda}(t)$ do not coincide with the actual values of the process $\lambda(t)$ are marked on the time axis with hatching (areas of wrong decisions).

To find a frequency of making wrong decisions about the random process $\lambda(t)$ state by the observations of the flow of events, a number of statistical experiments were implemented, consisting of the following stages:

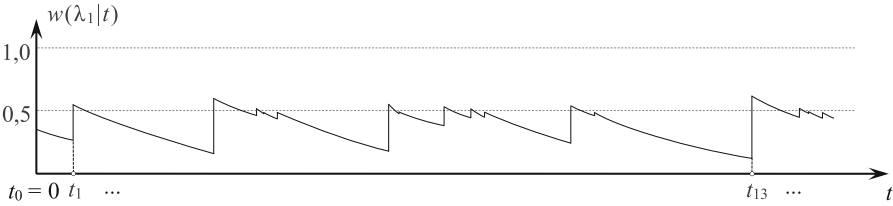


Fig. 4. Trajectory of a posteriori probability $w(\lambda_1|t)$ behavior

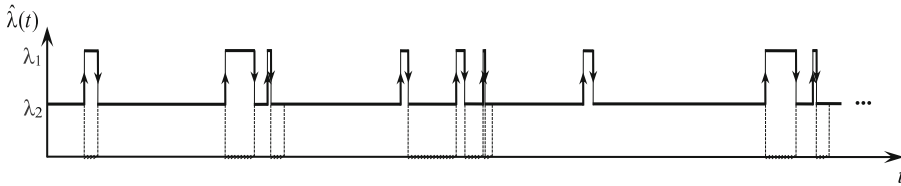


Fig. 5. Trajectory of the estimation $\hat{\lambda}(t)$ behavior

- (1) for a specific set of parameters $(\lambda_1, \lambda_2, \alpha_1, \alpha_2)$ and various values of the probabilities $P_1^{(1)}(\lambda_j|\lambda_i)$ and $P_1^{(2)}(\lambda_j|\lambda_i)$ of the process $\lambda(t)$ transitions from the i th state to the j th, $i, j = 1, 2$, the simulation of the flow is made at the given time interval $[0, T]$ (the separate k th test);
- (2) according to formulas (6), (10), (11), a posteriori probability $w(\lambda_1|t)$ of the first state of the process $\lambda(t)$ is calculated at the interval $[0, T]$;
- (3) at any time moment t at the interval $[0, T]$, the process $\lambda(t)$ value is estimated by criterion of a posteriori probability maximum;
- (4) the total length d_k of the time intervals where the actual values of the process $\lambda(t)$ do not coincide with its estimate $\hat{\lambda}(t)$ is determined (for the separate k th test);
- (5) the fraction of making a wrong decision $\hat{p}_k = \frac{d_k}{T}$ is calculated, where T is a modeling time (a time of the observing the flow);
- (6) steps 1–5 are repeated N times ($k = \overline{1, N}$) to calculate the estimation of unconditional probability of making wrong decision about the random process $\lambda(t)$ states at the considering time interval $[0, T]$.

The result of implementation of the described algorithm is a sample of fractions of making wrong decisions $\hat{p}_1, \hat{p}_2, \dots, \hat{p}_N$ for N experiments. Using this sample, we can find a sample average of unconditional probability of making an error $\hat{P}_{err} = \frac{1}{N} \sum_{k=1}^N \hat{p}_k$, as well as a sample variance $\hat{D}_{err} = \frac{1}{N-1} \sum_{k=1}^N (\hat{p}_k - \hat{P}_{err})^2$.

The first statistical experiment establishes the relation between modeling time T of the flow of events and the estimates \hat{P}_{err} and \hat{D}_{err} with changing the flow parameters λ_1 and α_1 . Tables 4, 5, 6 and 7 show the results obtained for the values of the flow parameters presented in Table 2 and $N = 100$. The first lines of Tables 4, 5, 6 and 7 contain the values of modeling time T ($T = 100, 200, \dots, 800$

time units), the second and the third lines contain the numerical values of the estimates \hat{P}_{err} and \hat{D}_{err} for each T , respectively.

Table 2. Initial data for the first statistical experiment

$\lambda_1 = 5; 7$	$P_1^{(1)}(\lambda_1 \lambda_1) = 0, 6$	$P_1^{(1)}(\lambda_2 \lambda_1) = 0, 4$
$\lambda_2 = 1$	$P_1^{(1)}(\lambda_2 \lambda_2) = 0, 5$	$P_1^{(1)}(\lambda_1 \lambda_2) = 0, 5$
$\alpha_1 = 4; 6$	$P_1^{(2)}(\lambda_1 \lambda_1) = 0, 3$	$P_1^{(2)}(\lambda_2 \lambda_1) = 0, 7$
$\alpha_2 = 1$	$P_1^{(2)}(\lambda_2 \lambda_2) = 0, 8$	$P_1^{(2)}(\lambda_1 \lambda_2) = 0, 2$

Analyzing the numerical results given in Tables 4, 5, 6 and 7, we note that: firstly, the estimate of total (unconditional) probability of making an error \hat{P}_{err} for all variants of calculations is sufficiently stable for $T \geq 100$ time units; secondly, for the fixed value of modeling time T , the estimate \hat{P}_{err} decreases with increasing each of the parameter λ_1 and α_1 , which is natural due to the random process $\lambda(t)$ states are better distinguishable; thirdly, for all variants of calculations, the sample variance \hat{D}_{err} is sufficiently small.

In the second experiment we investigate how the estimates \hat{P}_{err} and \hat{D}_{err} depend on the ratios λ_1/λ_2 and α_1/α_2 . Tables 8 and 9 show the results obtained for the initial data given in Table 3, with $N = 100$ and $T = 100$. The first lines of the Tables show the values of the corresponding ratios λ_1/λ_2 or α_1/α_2 , the second and the third lines contain the values of the \hat{P}_{err} and \hat{D}_{err} , respectively.

Table 3. Initial data for the second statistical experiment

$\lambda_1 = 6$	$P_1^{(1)}(\lambda_1 \lambda_1) = 0, 6$	$P_1^{(1)}(\lambda_2 \lambda_1) = 0, 4$
$\lambda_1/\lambda_2 = 3; \dots; 192$	$P_1^{(1)}(\lambda_2 \lambda_2) = 0, 5$	$P_1^{(1)}(\lambda_1 \lambda_2) = 0, 5$
$\alpha_1 = 6$	$P_1^{(2)}(\lambda_1 \lambda_1) = 0, 3$	$P_1^{(2)}(\lambda_2 \lambda_1) = 0, 7$
$\alpha_1/\alpha_2 = 3; \dots; 192$	$P_1^{(2)}(\lambda_2 \lambda_2) = 0, 8$	$P_1^{(2)}(\lambda_1 \lambda_2) = 0, 2$

The obtained numerical results indicate that better estimation corresponds to greater values of the ratios λ_1/λ_2 and α_1/α_2 . In this case, a frequency of making wrong decisions is reduced due to the distinguishability of the process $\lambda(t)$ states improves. We also note that better quality of the states estimation (in sense of smallness of the estimation of probability of making an error) is provided with an increase the ratio λ_1/λ_2 , than with an increase the ratio α_1/α_2 . This is explained by the set of the probabilities of the process $\lambda(t)$ transitions from the i th state to the j th, $i, j = 1, 2$, specified in Table 3.

It is of interest to consider a particular case of setting values of the probabilities of transition of the process $\lambda(t)$, at which the transitions of the process from the first state to the second, or conversely, take place at each instant of the flow

Table 4. Results of the first statistical experiment ($\lambda_1 = 5, \alpha_1 = 4$)

T	100	200	300	400	500	600	700	800
\hat{P}_{err}	0,1403	0,1427	0,1408	0,1414	0,1413	0,1416	0,1421	0,1431
$\hat{D}_{err} \cdot 10^2$	0,0408	0,0272	0,0116	0,0104	0,0097	0,0079	0,0092	0,0063

Table 5. Results of the first statistical experiment ($\lambda_1 = 7, \alpha_1 = 4$)

T	100	200	300	400	500	600	700	800
\hat{P}_{err}	0,1248	0,1255	0,1277	0,1273	0,1277	0,1286	0,1280	0,1287
$\hat{D}_{err} \cdot 10^2$	0,0368	0,0199	0,0149	0,0091	0,0071	0,0067	0,0062	0,0051

Table 6. Results of the first statistical experiment ($\lambda_1 = 5, \alpha_1 = 6$)

T	100	200	300	400	500	600	700	800
\hat{P}_{err}	0,1177	0,1147	0,1178	0,1159	0,1166	0,1165	0,1162	0,1182
$\hat{D}_{err} \cdot 10^2$	0,0271	0,0146	0,0099	0,0088	0,0063	0,0066	0,0043	0,0047

Table 7. Results of the first statistical experiment ($\lambda_1 = 7, \alpha_1 = 6$)

T	100	200	300	400	500	600	700	800
\hat{P}_{err}	0,1075	0,1097	0,1080	0,1088	0,1092	0,1081	0,1080	0,1089
$\hat{D}_{err} \cdot 10^2$	0,0229	0,0113	0,0098	0,0074	0,0047	0,0042	0,0029	0,0038

Table 8. Results of the second statistical experiment (for $\lambda_1/\lambda_2, \alpha_2 = 2$)

λ_1/λ_2	3	6	12	24	48	96	192
\hat{P}_{err}	0,1952	0,1349	0,1056	0,0869	0,0757	0,0717	0,0697
$\hat{D}_{err} \cdot 10^2$	0,0465	0,0339	0,0327	0,0209	0,0236	0,0189	0,0185

Table 9. Results of the second statistical experiment (for $\alpha_1/\alpha_2, \lambda_2 = 2$)

α_1/α_2	3	6	12	24	48	96	192
\hat{P}_{err}	0,1969	0,1781	0,1667	0,1603	0,1570	0,1536	0,1515
$\hat{D}_{err} \cdot 10^2$	0,0312	0,0424	0,0348	0,0387	0,0335	0,0410	0,0448

event occurrence. In the third experiment, for the appropriate values of the probabilities and the values of the parameters specified in Table 10 (with modeling time $T = 100$ and number of repetitions of the experiment $N = 100$), calculations are made to find the trajectory of the process estimation $\hat{\lambda}(t)$ behavior and the numerical values of the estimates \hat{P}_{err} and \hat{D}_{err} .

Table 10. Initial data for the third statistical experiment

$\lambda_1 = 4$	$P_1^{(1)}(\lambda_1 \lambda_1) = 0$	$P_1^{(1)}(\lambda_2 \lambda_1) = 1$
$\lambda_2 = 0, 75$	$P_1^{(1)}(\lambda_2 \lambda_2) = 0$	$P_1^{(1)}(\lambda_1 \lambda_2) = 1$
$\alpha_1 = 1$	$P_1^{(2)}(\lambda_1 \lambda_1) = 0$	$P_1^{(2)}(\lambda_2 \lambda_1) = 1$
$\alpha_2 = 0, 45$	$P_1^{(2)}(\lambda_2 \lambda_2) = 0$	$P_1^{(2)}(\lambda_1 \lambda_2) = 1$

Corresponding to the set values a trajectory of the random process $\lambda(t)$ behavior obtained in one of the tests using the simulation model, is shown in Fig. 6, where λ_1, λ_2 are values of the process $\lambda(t), t_1, t_2, \dots$ are moments of events occurrence in the considering flow.

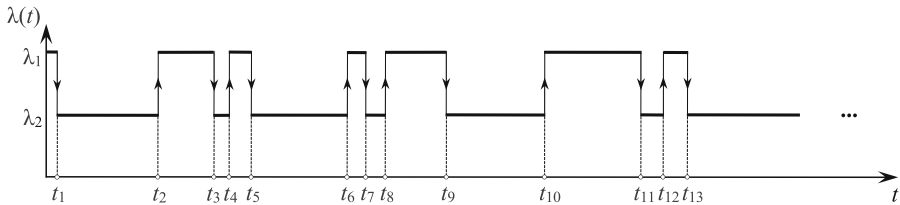


Fig. 6. Implementation of synchronous generalized flow of events of the second order

Figure 7 illustrates a posteriori probability $w(\lambda_1|t)$ behavior corresponding to the obtained sequence of the moments of events occurrence t_1, t_2, \dots

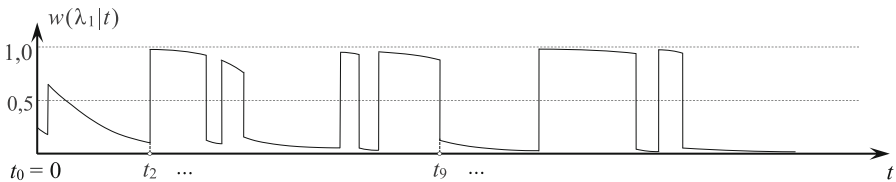


Fig. 7. Trajectory of a posteriori probability $w(\lambda_1|t)$ behavior

A trajectory of the estimation $\hat{\lambda}(t)$ of the process $\lambda(t)$ is shown in Fig. 8, here λ_1, λ_2 are the estimation $\hat{\lambda}(t)$ values. The decision about the first or the second

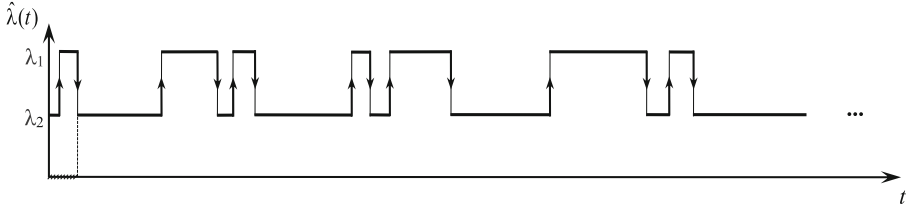


Fig. 8. Trajectory of the estimation $\hat{\lambda}(t)$ behavior

state of the process $\lambda(t)$ was made with the step $\Delta t = 0,001$. Areas of wrong decisions are indicated with hatching.

The fraction of making a wrong decision is $\hat{p}_k = 0,002108$ for this particular test.

In a series of N trials the sample average of unconditional probability of making an error is $\hat{P}_{err} = 0,006016$, and its sample variance is $\hat{D}_{err} = 0,000085$.

For comparison, the values of the estimates \hat{P}_{err} and \hat{D}_{err} , calculated for the same values of the parameters and the probabilities $P_1^{(1)}(\lambda_j|\lambda_i) = 0,5, i, j = 1, 2, l = 1, 2$, are following: $\hat{P}_{err} = 0,202981$ and $\hat{D}_{err} = 0,000971$. Thus, because of the probabilistic mechanism of transition of synchronous generalized flow of events of the second order from state to state at each moment of an event occurrence [11], the constructed optimal estimation algorithm provides the sample average of unconditional probability of making an error for the values of the probabilities of transition given in Table 10, close to zero. The sample variance is sufficiently small (an order of smallness decreases by at least 10 times in comparison with the situation that arises for the same set of parameters, but other values of the probabilities of transition of the process $\lambda(t)$).

5 Conclusion

In this paper, the optimal estimation algorithm for the states of $\hat{\nu}$ synchronous generalized flow of events of the second order is developed, also numerical results of a number of experiments that demonstrate a sufficiently good quality of estimation of the states of the flow (in sense of smallness of the estimation of probability of making an error) using only results of current observations of it during some period of time, were presented.

The formula (11) for calculating a posteriori probability of the first state at the time intervals between the moments of occurrence of the flow events, the conversion formula (6), valid at the time moments of the events occurrence, and the expression (10) for a priori probability of the first state of the process, which are necessary for the estimation of the states of the considering flow, were obtained explicitly, so there was no need to use numerical methods. The very algorithm for optimal estimation provides a minimum of unconditional probability of making an error.

References

1. Dudin, A.N., Klimenok, V.I.: Queueing Systems with Correlated Flows. BSU, Minsk (2000)
2. Basharin, G.P., Gaidamaka, Yu. V., Samouylov, K.E.: Mathematical theory of teletraffic and its application to the analysis of multiservice communication of next generation networks. *Autom. Control Comput. Sci.* **2**(47), 62–69 (2013)
3. Gortsev, A.M., Leonova, M.A., Nezhel'skaya, L.A.: Joint probability density of the duration of intervals of asynchronous generalized flow of events with unextendable dead time. *Tomsk State Univ. J. Control Comput. Sci.* **4**(21), 14–25 (2012)
4. Gortsev, A.M., Kalyagin, A.A., Nezhel'skaya, L.A.: Joint probability density of the duration of intervals of semisynchronous generalized flow of events with unextendable dead time. *Tomsk State Univ. J. Control Comput. Sci.* **2**(27), 19–29 (2014)
5. Gortsev, A.M., Nissenbaum, O.V.: Estimation of dead time period of asynchronous alternative flow of events with unextendable dead time period. *Russ. Phys. J.* **10**, 35–49 (2005)
6. Gortsev, A.M., Leonova, M.A., Nezhel'skaya, L.A.: Comparison of ML- and MM-estimates of dead time period in asynchronous generalized flow of events. *Tomsk State Univ. J. Control Comput. Sci.* **4**(25), 32–42 (2013)
7. Basharin, G.P., Kokotushkin, V.A., Naumov, V.A.: On the equivalent substitutions method for computing fragments of communication networks. *Proc. USSR Acad. Sci. Tech. Cybern.* **6**, 92–99 (1979)
8. Neuts, M.F.: A versatile Markovian point process. *J. Appl. Probab.* **16**, 764–779 (1979)
9. Nezhel'skaya, L.A.: Joint probability density of the intervals duration in modulated MAP event flows and its recurrence conditions. *Tomsk State Univ. J. Control Comput. Sci.* **1**, 57–67 (2015)
10. Gortsev, A.M., Nezhel'skaya, L.A.: On connection of MC flows and MAP flows of events. *Tomsk State Univ. J. Control Comput. Sci.* **1**(14), 13–21 (2011)
11. Nezhel'skaya, L.A., Sidorova, E.F.: Simulation modeling of synchronous generalized flow of events of the second order. *Proc. TSU Ser. Phys. Math.* **299**, 104–109 (2016)
12. Khazen, E.M.: *Methods of Optimal Statistical Decisions and Problems of Optimal Control*. Soviet Radio, Moscow (1968)
13. Gortsev, A.M., Nezhel'skaya, L.A.: The optimal nonlinear filtration of Markovian flow of events with commutation. *Commun. Tech. Ser. Telecommun. Syst.* **7**, 46–54 (1989)
14. Sobol', I.M.: *Numerical Monte Carlo Methods*. Science, Moscow (1973)



Asymptotic Sojourn Time Analysis of Finite-Source M/M/1 Retrial Queuing System with Two-Way Communication

Anatoly Nazarov¹, János Sztrik²(✉), and Anna Kvach¹

¹ National Research Tomsk State University,
36 Lenina ave., Tomsk 634050, Russia
nazarov.tsu@gmail.com, kvach.as@mail.ru

² University of Debrecen, Debrecen, Hungary
sztrik.janos@inf.unideb.hu

Abstract. The aim of the present paper is to investigate a retrial queuing system $M/M/1$ with a finite number of sources and two-way communication. Each source can generate a request after an exponentially distributed time and will not generate another one until the previous call return to the source. If an incoming customer finds the server idle its service starts. Otherwise, if the server is busy an arriving (primary or repeated) customer moves into the orbit and after some exponentially distributed time it retries to enter the server. When the server is idle it generates an outgoing call after an exponentially distributed time with different parameters to the customers in the orbit and to the sources, respectively. The service times of the incoming and outgoing calls are exponentially distributed with different rates. Applying method of asymptotic analysis under the condition of unlimited growing number of sources it is proved that the limiting sojourn/waiting time of the customer in the system follows a generalized exponential distribution with given parameters. In addition, the asymptotic average number of customers in the orbit is obtained.

Keywords: Finite-source queuing system · Retrial queues
Call centers · Two-way communication · Asymptotic analysis
Sojourn time distribution

1 Introduction

Modeling retrial queuing systems with two-way communication has been becoming more and more popular topic of investigations for the last years. The main reason is that in many applications, for example in call centers where the agents could make outgoing calls to advertise, promote and sell packages and services of the center, it is important to increase the utilization the server, see for example [1, 2, 5, 10, 17, 20]. The main feature of two-way communication is that and idle server can generate outgoing calls to the source (primary calls) or

to the orbit (retrial calls). If a primary outgoing call finds the server busy it is lost in infinite source case or returns to the source in finite-source case. If at the arrival of a retrial outgoing request the server is busy it goes back to the orbit and can generate a retrial call. The first results on infinite source queueing systems with two-way communication was published by Falin [9], followed by some recent ones, see for example [3, 4, 6, 7, 13, 14, 16, 18, 19].

Finite-source retrial queueing systems with two-way communication has not been investigated intensively, yet. To the best knowledge of the authors only the paper of Dragieva and Phung-Duc [8] dealt with this problem. They investigated an M/M/1//N retrial model with exponentially distributed retrial times where the primary and retrial outgoing call generation and service times are also exponentially distributed. Recursive formulas for computing the steady-state distribution of the system state were derived as well as expressions for the main performance macro characteristics in terms of the server utilization were obtained. Numerical examples were presented. It is easy to see that by choosing parameters in an appropriate way the previous results for one-way communication and as limiting case for infinite source models two-way communication can be obtained.

In their Conclusion the authors mentioned studying waiting time process among others. Hence it was our main motivation to investigate the distribution of the waiting and response time distribution of primary incoming calls by using asymptotic methods similar to [11, 12, 15]. Assuming that the number of sources N tends to infinity it is proved that the response/waiting time distribution of primary incoming customers in the system/orbit can be approximated by a generalized exponential distribution with given parameters. In addition, the asymptotic average number of customers in the system and in the orbit are obtained. The results are validated by the Little-formula.

The rest of the paper is organized as follows. In Sect. 2 description of the model is given, the corresponding 2-dimensional Markov process is defined. In Sect. 3 the mean normalized number customers in the orbit is obtained. Section 4 deals with the distribution of the response and waiting time of calls. Finally, the paper ends with a Conclusion and some future plans are highlighted.

2 Model Description and Notations

Let us consider a retrial queueing system of type M/M/1//N with two-way communication. The number of sources is N and each of them can generate a primary request with rate λ/N . A source cannot generate a new call until the end of the successful service of this customer. If incoming (primary or retrial) customer finds the server idle, it enters into service immediately, in which the required service time is exponentially distributed random variable with parameter μ_1 . Otherwise, if the server is busy, an arriving (primary or repeated) customer moves into the orbit. The retrial times of requests are assumed to be exponentially distributed with rate σ/N . We suppose that if the server is idle, it generates an outgoing call in an exponentially distributed time with rate α/N for outgoing call to the orbit

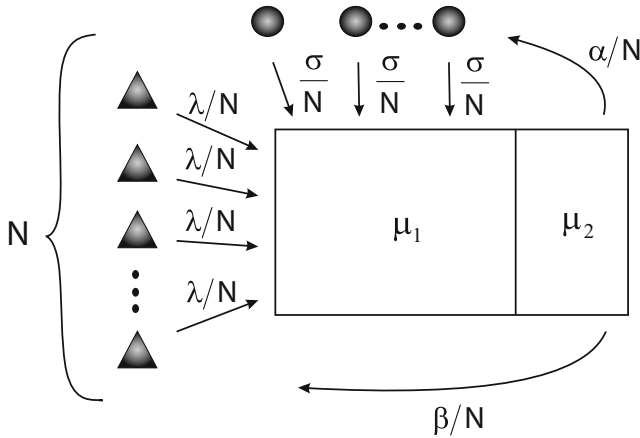


Fig. 1. Retrial queuing system of type M/M/1//N with two-way communication

and with rate β/N for primary outgoing calls. The service times of outgoing calls are assumed to be exponentially distributed random variable with parameter μ_2 . All random variables involved in the model construction are assumed to be independent of each other.

Our main aim is to find the sojourn time distribution of the customers in the system and in the orbit, respectively. The method of asymptotic analysis is used in the condition of an unlimited growing number of sources.

First, we will find the first order asymptotic mean normed number of customers in the orbit, the results of which we will apply later on to study the sojourn time distribution of the customer in the system.

3 First Order Asymptotic for the Number of Customers in the Orbit

Let $Q(t)$ be the number of customers on the orbit at time t , $C(t)$ be the server state at time t , that is

$$C(t) = \begin{cases} 0, & \text{if the server is idle,} \\ 1, & \text{if the server is busy by an incoming call,} \\ 2, & \text{if the server is busy by an outgoing call.} \end{cases}$$

Thus, we will investigate the Markov process $\{C(t), Q(t)\}$.

Let us define the stationary probabilities as follows:

$$P_k(n) = \lim_{t \rightarrow \infty} P\{C(t) = k, Q(t) = n\}.$$

For the stationary probability distribution $P_k(n)$ by using standard methods we can obtain the following system of Kolmogorov equations, namely

$$\begin{aligned}
 & - \left[\lambda + \beta + (\sigma + \alpha - \lambda - \beta) \frac{n}{N} \right] P_0(n) + \mu_1 P_1(n) + \mu_2 P_2(n) = 0, \\
 & - \left[\lambda + \mu_1 - \lambda \frac{n+1}{N} \right] P_1(n) + \lambda \left(1 - \frac{n}{N} \right) [P_0(n) + P_1(n-1)] \\
 & + \sigma \frac{n+1}{N} P_0(n+1) = 0, \\
 & - \left[\lambda + \mu_2 - \lambda \frac{n+1}{N} \right] P_2(n) + \beta \left(1 - \frac{n}{N} \right) P_0(n) + \alpha \frac{n+1}{N} P_0(n+1) \\
 & + \lambda \left(1 - \frac{n}{N} \right) P_2(n-1) = 0.
 \end{aligned} \tag{1}$$

In paper [8] although the notations are different basically this system was solved by using a recursive numerical algorithm. Since our aim is to get the sojourn time distribution of the customers we follow an asymptotic method because to obtain the exact distribution we need rather complicated approach. Let us denote the partial characteristic functions as

$$H_k(u) = \sum_{n=0}^N e^{iun} P_k(n),$$

where $i = \sqrt{-1}$ is imaginary unit, then system (1) will be rewritten in the form

$$\begin{aligned}
 & - (\lambda + \beta) H_0(u) + \mu_1 H_1(u) + \mu_2 H_2(u) + i \frac{[\sigma + \alpha - \lambda - \beta]}{N} H_0'(u) = 0, \\
 & \lambda H_0(u) + \left[\lambda (e^{iu} - 1) \left(1 - \frac{1}{N} \right) - \mu_1 \right] H_1(u) \\
 & + i \frac{(\lambda - \sigma e^{-iu})}{N} H_0'(u) + i \frac{\lambda (e^{iu} - 1)}{N} H_1'(u) = 0, \\
 & \beta H_0(u) + \left(\lambda (e^{iu} - 1) \left(1 - \frac{1}{N} \right) - \mu_2 \right) H_2(u) \\
 & + i \frac{(\beta - \alpha e^{-iu})}{N} H_0'(u) + i \frac{\lambda (e^{iu} - 1)}{N} H_2'(u) = 0.
 \end{aligned} \tag{2}$$

Summarizing equations of system (2) we receive an additional equality of the form

$$\begin{aligned}
 & \lambda (e^{iu} - 1) \left(1 - \frac{1}{N} \right) [H_1(u) + H_2(u)] + i \frac{(\alpha + \sigma)(1 - e^{-iu})}{N} H_0'(u) \\
 & + i \frac{\lambda (e^{iu} - 1)}{N} [H_1'(u) + H_2'(u)] = 0.
 \end{aligned} \tag{3}$$

We will solve system (2) and (3) by the method of asymptotic analysis in the condition of an infinitely increasing number of sources $N \rightarrow \infty$.

Theorem 1. *Let Q be the number of customers in the orbit then*

$$\lim_{N \rightarrow \infty} E \exp \left\{ iw \frac{Q}{N} \right\} = \exp \{ iw \kappa \}, \tag{4}$$

where value of parameter κ is the positive solution of the equation

$$\lambda(1 - \kappa) [R_1(\kappa) + R_2(\kappa)] - (\alpha + \sigma)R_0(\kappa)\kappa = 0. \tag{5}$$

Here the stationary distributions of probabilities $R_k(\kappa)$ of the service state k depends on κ and can be obtained as follows

$$\begin{aligned} R_0(\kappa) &= \left\{ 1 + \frac{1}{\mu_1} [\lambda(1 - \kappa) + \sigma\kappa] + \frac{1}{\mu_2} [\beta(1 - \kappa) + \alpha\kappa] \right\}^{-1}, \\ R_1(\kappa) &= \frac{1}{\mu_1} [\lambda(1 - \kappa) + \sigma\kappa] R_0(\kappa), \\ R_2(\kappa) &= \frac{1}{\mu_2} [\beta(1 - \kappa) + \alpha\kappa] R_0(\kappa). \end{aligned} \tag{6}$$

Proof. Designating $\frac{1}{N} = \varepsilon$, in systems (2–3) let us introduce the following replacements

$$u = \varepsilon w, \quad H_k(u) = F_k(w, \varepsilon), \tag{7}$$

then systems (2–3) can be rewritten as

$$\begin{aligned} & -(\lambda + \beta) F_0(w, \varepsilon) + \mu_1 F_1(w, \varepsilon) + \mu_2 F_2(w, \varepsilon) \\ & + i [\sigma + \alpha - \lambda - \beta] \frac{\partial F_0(w, \varepsilon)}{\partial w} = 0, \\ & \lambda F_0(w, \varepsilon) + [\lambda (e^{i\varepsilon w} - 1) (1 - \varepsilon) - \mu_1] F_1(w, \varepsilon) \\ & + i (\lambda - \sigma e^{-i\varepsilon w}) \frac{\partial F_0(w, \varepsilon)}{\partial w} + i \lambda (e^{i\varepsilon w} - 1) \frac{\partial F_1(w, \varepsilon)}{\partial w} = 0, \\ & \beta F_0(w, \varepsilon) + [\lambda (e^{i\varepsilon w} - 1) (1 - \varepsilon) - \mu_2] F_2(w, \varepsilon) \\ & + i (\beta - \alpha e^{-i\varepsilon w}) \frac{\partial F_0(w, \varepsilon)}{\partial w} + i \lambda (e^{i\varepsilon w} - 1) \frac{\partial F_2(w, \varepsilon)}{\partial w} = 0, \\ & \lambda (1 - \varepsilon) [F_1(w, \varepsilon) + F_2(w, \varepsilon)] + i (\alpha + \sigma) e^{-i\varepsilon w} \frac{\partial F_0(w, \varepsilon)}{\partial w} \\ & + i \lambda \left[\frac{\partial F_1(w, \varepsilon)}{\partial w} + \frac{\partial F_2(w, \varepsilon)}{\partial w} \right] = 0. \end{aligned} \tag{8}$$

Denoting $\lim_{\varepsilon \rightarrow 0} F_k(w, \varepsilon) = F_k(w)$, let us execute this limiting transition in system (8) and as result we will obtain

$$\begin{aligned}
 & -(\lambda + \beta) F_0(w) + \mu_1 F_1(w) + \mu_2 F_2(w) + i[\sigma + \alpha - \lambda - \beta] F_0'(w) = 0, \\
 & \lambda F_0(w) - \mu_1 F_1(w) + i(\lambda - \sigma) F_0'(w) = 0, \\
 & \beta F_0(w) - \mu_2 F_2(w) + i(\beta - \alpha) F_0'(w) = 0, \\
 & \lambda [F_1(w) + F_2(w)] + i(\alpha + \sigma) F_0'(w) + i\lambda [F_1'(w) + F_2'(w)] = 0.
 \end{aligned} \tag{9}$$

We show that the solution of the system (9) can be written in the following form

$$F_k(w) = R_k \Phi(w), \tag{10}$$

where R_k the limiting probability distributions of the service state k under conditions $N \rightarrow \infty$ and $\Phi(w)$ is the limiting characteristic function of the normalized number of customers in the orbit. Substituting solution (10) in (9) we obtain

$$\begin{aligned}
 & -(\lambda + \beta) R_0 + \mu_1 R_1 + \mu_2 R_2 + i[\sigma + \alpha - \lambda - \beta] R_0 \frac{\Phi'(w)}{\Phi(w)} = 0, \\
 & \lambda R_0 - \mu_1 R_1 + i(\lambda - \sigma) R_0 \frac{\Phi'(w)}{\Phi(w)} = 0, \\
 & \beta R_0 - \mu_2 R_2 + i(\beta - \alpha) R_0 \frac{\Phi'(w)}{\Phi(w)} = 0, \\
 & \lambda [R_1 + R_2] + i\{(\alpha + \sigma) R_0 + \lambda (R_1 + R_2)\} \frac{\Phi'(w)}{\Phi(w)} = 0.
 \end{aligned} \tag{11}$$

From the form of equations of system (11) it follows that quantity $\frac{\Phi'(w)}{\Phi(w)}$ does not depend on w , then we can conclude that function $\Phi(w)$ has a form

$$\Phi(w) = \exp(iw\kappa), \tag{12}$$

coinciding with equality (4).

Now, to find κ let us substitute the explicit form of the function $\Phi(w)$ in the equations of system (11) and we obtain the following system

$$\begin{aligned}
 & -(\lambda + \beta) R_0 + \mu_1 R_1 + \mu_2 R_2 - [\sigma + \alpha - \lambda - \beta] R_0 \kappa = 0, \\
 & \lambda R_0 - \mu_1 R_1 - (\lambda - \sigma) R_0 \kappa = 0, \\
 & \beta R_0 - \mu_2 R_2 - (\beta - \alpha) R_0 \kappa = 0, \\
 & \lambda(1 - \kappa) [R_1 + R_2] - (\alpha + \sigma) R_0 \kappa = 0.
 \end{aligned} \tag{13}$$

The quantity κ is the solution of the fourth equation of system (13), which coincides with (5) showing a probabilistic interpretation that the mean arrival rate to the orbit is equal to the mean departure rate from the orbit. From the second and third equations of system (13) taking into account the normalization condition, it is not difficult to obtain expressions for the quantities R_0 , R_1 and R_2 . Let us note that since R_k is the solution of system (13) whose coefficients depend on κ , then $R_k = R_k(\kappa)$ and their form coincides with (6).

The theorem is proved. □

4 Sojourn Time Distribution of the Customer in the System

Let T be the total sojourn time of the tagged customer in the system and $T(t)$ is the time length from moment t until the end of the service of the tagged customer. The total sojourn time T is simply expressed through the residual sojourn time $T(t)$.

Let $S(t)$ describe the server state at time t as follows

$$S(t) = \begin{cases} 0, & \text{server is free,} \\ 1, & \text{server is busy by incoming (not tagged) customer,} \\ 2, & \text{server is busy by outgoing (not tagged) customer,} \\ 3, & \text{server is busy by incoming tagged customer,} \\ 4, & \text{server is busy by outgoing tagged customer.} \end{cases}$$

We will define the conditional characteristic functions in the form

$$G_k(u, n, t) = E \left\{ e^{iuT(t)} | S(t) = k, Q(t) = n \right\}.$$

Assuming that the system is operating in steady-state, it is not difficult to write the following system of Kolmogorov equations

$$\begin{aligned} & \left[iu - (\lambda + \beta) \frac{N-n}{N} - (\sigma + \alpha) \frac{n}{N} \right] G_0(u, n) + \lambda \frac{N-n}{N} G_1(u, n) \\ & + \sigma \frac{n-1}{N} G_1(u, n-1) + \beta \frac{N-n}{N} G_2(u, n) + \alpha \frac{n-1}{N} G_2(u, n-1) \\ & + \frac{\sigma}{N} G_3(u, n-1) + \frac{\alpha}{N} G_4(u, n-1) = 0, \\ & \left[iu - \lambda \frac{N-n-1}{N} - \mu_1 \right] G_1(u, n) + \lambda \frac{N-n-1}{N} G_1(u, n+1) \\ & + \mu_1 G_0(u, n) = 0, \\ & \left[iu - \lambda \frac{N-n-1}{N} - \mu_2 \right] G_2(u, n) + \lambda \frac{N-n-1}{N} G_2(u, n+1) \\ & + \mu_2 G_0(u, n) = 0, \\ & \left[iu - \lambda \frac{N-n-1}{N} - \mu_1 \right] G_3(u, n) + \lambda \frac{N-n-1}{N} G_3(u, n+1) + \mu_1 = 0, \\ & \left[iu - \lambda \frac{N-n-1}{N} - \mu_2 \right] G_4(u, n) + \lambda \frac{N-n-1}{N} G_4(u, n+1) + \mu_2 = 0. \end{aligned} \tag{14}$$

The method of asymptotic analysis, see [11] is applied to prove the following theorem

Theorem 2. *Let T be the total sojourn time of the customer in the system then*

$$\lim_{N \rightarrow \infty} \mathbf{E} \exp \left\{ iw \frac{T}{N} \right\} = R_0 + (1 - R_0) \frac{(\alpha + \sigma) R_0}{(\alpha + \sigma) R_0 - iw}. \quad (15)$$

Proof. Let us denote $\frac{1}{N} = \varepsilon$. Executing the following substitutions in system (14)

$$u = \varepsilon w, \quad \varepsilon n = x, \quad G_k(u, n) = g_k(w, x, \varepsilon), \quad (16)$$

we obtain this system in the following form

$$\begin{aligned} & [i\varepsilon w - (\lambda + \beta)(1 - x) - (\sigma + \alpha)x] g_0(w, x, \varepsilon) + \lambda(1 - x)g_1(w, x, \varepsilon) \\ & + \sigma(x - \varepsilon)g_1(w, x - \varepsilon, \varepsilon) + \beta(1 - x)g_2(w, x, \varepsilon) \\ & + \alpha(x - \varepsilon)g_2(w, x - \varepsilon, \varepsilon) + \sigma\varepsilon g_3(w, x - \varepsilon, \varepsilon) + \alpha\varepsilon g_4(w, x - \varepsilon, \varepsilon) = 0, \\ & [i\varepsilon w - \lambda(1 - x - \varepsilon) - \mu_1] g_1(w, x, \varepsilon) + \lambda(1 - x - \varepsilon)g_1(w, x + \varepsilon, \varepsilon) \\ & + \mu_1 g_0(w, x, \varepsilon) = 0, \\ & [i\varepsilon w - \lambda(1 - x - \varepsilon) - \mu_2] g_2(w, x, \varepsilon) + \lambda(1 - x - \varepsilon)g_2(w, x + \varepsilon, \varepsilon) \\ & + \mu_2 g_0(w, x, \varepsilon) = 0, \\ & [i\varepsilon w - \lambda(1 - x - \varepsilon) - \mu_1] g_3(w, x, \varepsilon) + \lambda(1 - x - \varepsilon)g_3(w, x + \varepsilon, \varepsilon) \\ & + \mu_1 = 0, \\ & [i\varepsilon w - \lambda(1 - x - \varepsilon) - \mu_2] g_4(w, x, \varepsilon) + \lambda(1 - x - \varepsilon)g_4(w, x + \varepsilon, \varepsilon) \\ & + \mu_2 = 0. \end{aligned} \quad (17)$$

Denoting $\lim_{\varepsilon \rightarrow 0} g_k(w, x, \varepsilon) = g_k(w, x)$, we carry out limiting transition under condition $\varepsilon \rightarrow 0$ in the system (17) and the we get

$$\begin{aligned} & - [(\lambda + \beta)(1 - x) + (\sigma + \alpha)x] g_0(w, x) + [\lambda(1 - x) + \sigma x] g_1(w, x) \\ & + [\beta(1 - x) + \alpha x] g_2(w, x) = 0, \\ & \mu_1 [g_0(w, x) - g_1(w, x)] = 0, \\ & \mu_2 [g_0(w, x) - g_2(w, x)] = 0, \\ & \mu_1 [1 - g_3(w, x)] = 0, \\ & \mu_2 [1 - g_4(w, x)] = 0. \end{aligned} \quad (18)$$

From the obtained system it follows that functions $g_3(w, x)$ and $g_4(w, x)$ are equal to unity, and functions $g_0(w, x)$, $g_1(w, x)$ and $g_2(w, x)$ are coincide.

Designating by $g(w, x)$ their common value we can write

$$g(w, x) = g_0(w, x) = g_1(w, x) = g_2(w, x).$$

Thus, the solution of the system (17) can be written in the form of decomposition

$$g_k(w, x, \varepsilon) = g(w, x) + \varepsilon f_k(w, x) + O(\varepsilon^2), \quad k = \overline{0, 2},$$

which we substitute into the first three equations of the system (17) and as a result we obtain

$$\begin{aligned} & [i\varepsilon w - (\lambda + \beta)(1 - x) - (\sigma + \alpha)x]g(w, x) + \sigma\varepsilon g_3(w, x) + \alpha\varepsilon g_4(w, x) \\ & - \varepsilon [(\lambda + \beta)(1 - x) + (\sigma + \alpha)x]f_0(w, x) + \varepsilon [\lambda(1 - x) + \sigma x]f_1(w, x) \\ & + [\lambda(1 - x) + \sigma(x - \varepsilon)]g(w, x) + [\beta(1 - x) + \alpha(x - \varepsilon)]g(w, x) \\ & + \varepsilon [\beta(1 - x) + \alpha x]f_2(w, x) - \varepsilon(\sigma + \alpha)x \frac{\partial g(w, x)}{\partial x} = O(\varepsilon^2), \\ & [i\varepsilon w - \mu_1]g(w, x) - \varepsilon\mu_1 f_1(w, x) + \mu_1 g(w, x) + \varepsilon\mu_1 f_0(w, x) \\ & + \varepsilon\lambda(1 - x) \frac{\partial g(w, x)}{\partial x} = O(\varepsilon^2), \\ & [i\varepsilon w - \mu_2]g(w, x) - \varepsilon\mu_2 f_2(w, x) + \mu_2 g(w, x) + \varepsilon\mu_2 f_0(w, x) \\ & + \varepsilon\lambda(1 - x) \frac{\partial g(w, x)}{\partial x} = O(\varepsilon^2). \end{aligned} \tag{19}$$

After performing simple transformations in system (19) and taking into account that $g_3(w, x) = g_4(w, x) = 1$ we get

$$\begin{aligned} & - [(\lambda + \beta)(1 - x) + (\sigma + \alpha)x]f_0(w, x) + [\lambda(1 - x) + \sigma x]f_1(w, x) \\ & + [\beta(1 - x) + \alpha x]f_2(w, x) = [\sigma + \alpha - iw]g(w, x) - (\sigma + \alpha) \\ & + (\sigma + \alpha)x \frac{\partial g(w, x)}{\partial x} = 0, \\ & \mu_1 [f_0(w, x) - f_1(w, x)] = -iwg(w, x) - \lambda(1 - x) \frac{\partial g(w, x)}{\partial x} = 0, \\ & \mu_2 [f_0(w, x) - f_2(w, x)] = -iwg(w, x) - \lambda(1 - x) \frac{\partial g(w, x)}{\partial x} = 0. \end{aligned} \tag{20}$$

Multiplying the first equation of system (20) by R_0 , the second equation by R_1 , the third by R_2 and then adding them we receive the following equality

$$\begin{aligned}
 & [-(\lambda + \beta)R_0 - (\sigma + \alpha - \lambda - \beta)R_0x + \mu_1R_1 + \mu_2R_2] f_0(w, x) \\
 & + [\lambda R_0 - (\lambda - \sigma)R_0x - \mu_1R_1] f_1(w, x) \\
 & + [\beta R_0 - (\beta - \alpha)R_0x - \mu_2R_2] f_2(w, x) = [(\sigma + \alpha)R_0 - iw] g(w, x) \\
 & - (\sigma + \alpha)R_0 + \{(\sigma + \alpha)R_0x - \lambda(1 - x)(R_1 + R_2)\} \frac{\partial g(w, x)}{\partial x} = 0.
 \end{aligned} \tag{21}$$

From Theorem 1 it follows that $\frac{n}{N} = \kappa$. By virtue of the substitutions (16) carried out earlier, namely $x = n\varepsilon$ we can conclude that $x = \kappa$.

Let us perform the substitution $x = \kappa$ in the equations of system (21) and taking into account system (13), we can find that the multipliers for functions $f_0(w, x)$, $f_1(w, x)$, $f_2(w, x)$ and $\frac{\partial g(w, x)}{\partial x}$ are equal to zero.

As a result equality (21) can be rewritten in the form

$$[(\sigma + \alpha)R_0 - iw] g(w) = (\sigma + \alpha)R_0,$$

from which it obviously follows that

$$g(w) = \frac{(\sigma + \alpha)R_0}{(\sigma + \alpha)R_0 - iw}. \tag{22}$$

Thus we obtain the characteristic function of the probability distribution of the normalized residual sojourn time $T(t)/N$. The using the law of total probability we can receive the characteristic function of the probability distribution of the normalized total sojourn time T/N of the customers in the system in the following form

$$\begin{aligned}
 \lim_{N \rightarrow \infty} \mathbf{E} \exp \left\{ iw \frac{T}{N} \right\} &= R_0g_3(w) + R_1g_1(w) + R_2g_2(w) \\
 &= R_0 + (R_1 + R_2)g(w) = R_0 + (1 - R_0) \frac{(\sigma + \alpha)R_0}{(\sigma + \alpha)R_0 - iw},
 \end{aligned} \tag{23}$$

which coincides with (15).

The theorem is proved. □

In addition let us perform in Eq. (23) reverse replacement $w = \frac{u}{\varepsilon} = uN$ and denoting by $q = 1 - R_0$ and $\gamma = (\sigma + \alpha)R_0/N$ we receive

$$\mathbf{E} \exp \{iuT\} \approx R_0 + (1 - R_0) \frac{(\sigma + \alpha)R_0/N}{(\sigma + \alpha)R_0/N - iu} = 1 - q + q \frac{\gamma}{\gamma - iu},$$

which is the prelimit value, that is for fixed N .

Thus we obtain that the sojourn time of the customer in the system follows a generalized exponential distribution with parameters q and γ .

Consequently the mean response time can be approximated by $\frac{(1-R_0)}{(\sigma+\alpha)R_0/N}$.

Since the service time of a primary incoming customer is bounded then the limiting distribution of the normalized response time and the waiting time coincide. Similarly, the limiting distribution of the normalized number of customers in the system and in the orbit are the same.

Hence the mean arrival rate to the system is $\lambda(1-\kappa)$.

We will use the Little-formula to check our results, namely we have

$$\lambda(1-\kappa)\frac{(1-R_0)}{(\sigma+\alpha)R_0} = \kappa$$

which is equivalent to Eq. 5 from which κ was determined.

5 Conclusion and Future Work

In this paper a finite-source retrial queuing system of type $M/M/1//N$ with two-way communication was considered. The research has been performed by the method of asymptotic analysis under the condition of unlimited growing number of sources. As the result of the analysis it was shown that the limiting sojourn/waiting time of the customer in the system has a generalized exponential distribution with given parameters. The authors plan to continue their research, among others modeling finite-source retrial queuing systems with two-way communication for the case of generally distributed service times.

Acknowledgments. The work/publication of A.A. Nazarov is supported by grant RFBR (Russian Foundation for Basic Research), the Agreement number 18.01.00277.

References

1. Aguir, S., Karaesmen, F., Akşin, O.Z., Chauvet, F.: The impact of retrials on call center performance. *OR Spectr.* **26**(3), 353–376 (2004)
2. Aksin, Z., Armony, M., Mehrotra, V.: The modern call center: a multi-disciplinary perspective on operations management research. *Prod. Oper. Manag.* **16**(6), 665–688 (2007)
3. Artalejo, J.R., Phung-Duc, T.: Markovian retrial queues with two way communication. *J. Ind. Manag. Optim.* **8**(4), 781–806 (2012)
4. Artalejo, J., Phung-Duc, T.: Single server retrial queues with two way communication. *Appl. Math. Model.* **37**(4), 1811–1822 (2013)
5. Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., Zhao, L.: Statistical analysis of a telephone call center: a queueing-science perspective. *J. Am. Stat. Assoc.* **100**(469), 36–50 (2005)
6. Dimitriou, I.: A retrial queue to model a two-relay cooperative wireless system with simultaneous packet reception. In: Wittevrongel, S., Phung-Duc, T. (eds.) *ASMTA 2016*. LNCS, vol. 9845, pp. 123–139. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43904-4_9

7. Dragieva, V., Phung-Duc, T.: Two-way communication M/M/1 retrial queue with server-orbit interaction. In: Proceedings of the 11th International Conference on Queueing Theory and Network Applications, p. 11. ACM (2016)
8. Dragieva, V., Phung-Duc, T.: Two-way communication M/M/1//N retrial queue. In: Thomas, N., Forshaw, M. (eds.) ASMTA 2017. LNCS, vol. 10378, pp. 81–94. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61428-1_6
9. Falin, G.: Model of coupled switching in presence of recurrent calls. Eng. Cybern. **17**(1), 53–59 (1979)
10. Gans, N., Koole, G., Mandelbaum, A.: Telephone call centers: tutorial, review, and research prospects. Manuf. Serv. Oper. Manag. **5**(2), 79–141 (2003)
11. Kvach, A., Nazarov, A.: Sojourn time analysis of finite source Markov retrial queueing system with collision. In: Dudin, A., Nazarov, A., Yakupov, R. (eds.) ITMM 2015. CCIS, vol. 564, pp. 64–72. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25861-4_6
12. Nazarov, A., Sztrik, J., Kvach, A., Bérczes, T.: Asymptotic analysis of finite-source M/M/1 retrial queueing system with collisions and server subject to breakdowns and repairs. Annals of Operations Research (2017). (accepted for publication)
13. Nazarov, A., Phung-Duc, T., Paul, S.: Heavy outgoing call asymptotics for *MMPP/M/1/1* retrial queue with two-way communication. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) ITMM 2017. CCIS, vol. 800, pp. 28–41. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_3
14. Nazarov, A.A., Paul, S., Gudkova, I., et al.: Asymptotic analysis of Markovian retrial queue with two-way communication under low rate of retrials condition. In: Proceedings 31st European Conference on Modelling and Simulation (2017)
15. Nazarov, A., Moiseeva, S.P.: Methods of asymptotic analysis in queueing theory. NTL Publishing House of Tomsk University, Tomsk (2006). (in Russian)
16. Phung-Duc, T., Rogiest, W.: Two way communication retrial queues with balanced call blending. In: Al-Begain, K., Fiems, D., Vincent, J.-M. (eds.) ASMTA 2012. LNCS, vol. 7314, pp. 16–31. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-30782-9_2
17. Pustova, S.: Investigation of call centers as retrial queueing systems. Cybern. Syst. Anal. **46**(3), 494–499 (2010)
18. Sakurai, H., Phung-Duc, T.: Two-way communication retrial queues with multiple types of outgoing calls. Top **23**(2), 466–492 (2015)
19. Sakurai, H., Phung-Duc, T.: Scaling limits for single server retrial queues with two-way communication. Ann. Oper. Res. **247**(1), 229–256 (2016)
20. Wolf, T.: System and method for improving call center communications, 30 Nov 2017. US Patent App. 15/604,068



An Analysis Method of Queueing Networks with a Degradable Structure and Non-zero Repair Times of Systems

Igor E. Tananko^(✉) and Nadezhda P. Fokina

National Research Saratov State University, Saratov, Russia
TanankoIE@info.sgu.ru, fokinanp.sgu@gmail.com

Abstract. An open exponential network with a single class of demands and unreliable queueing systems is considered. System faults occur sequentially with exponential times per fault. When a fault occurs at a system, all the demands there are destroyed immediately. A key performance measure (quality of service measure) of the network is its response time, which is a stochastic process. When a value of the response time exceeds the preset threshold value activation of renewal mechanism of the failed systems takes place. Recovery time has an exponential distribution which does not depend on the number of failed systems. A method, which allows to obtain performance characteristics, was developed. Finally, some numerical examples and a section of conclusions commenting the main research contributions of this paper are presented.

Keywords: Queueing networks · Markov chains
Unreliable queueing systems
Degradable structure of queueing network

1 Introduction

Queueing networks with unreliable systems [1, 4, 7, 8, 20] are used as models for the analysis and control of unreliable discrete stochastic systems with network structure.

Article [22] considers generalized Jackson networks with single-server stations, where nodes may have an infinite supply of work. Assume there are simultaneous breakdown of servers.

An approximate decomposition method for the analysis of production lines with multiple stations is presented in [15]. The machines are unreliable and have exponential operation, failure, and repair processes. Intermediate buffers between different stations are all with finite capacities. Similar models are studied also in [14].

Article [23] presents limiting distributions for multichannel systems and open queueing networks with unreliable elements: nodes, paths between nodes, and channels at nodes.

Different applications of unreliable queueing networks are presented in [3, 6, 11, 13, 16, 21].

Queueing models for satellite networks are studied in [6]. In [3], discuss an extension of Markov reliability models to include parametric sensitivity analysis.

Paper [2] studies the stationary dynamics of a processing system comprised of several parallel queues and a single server of constant rate. The connectivity of the server to each queue is randomly modulated, taking values 1 (connected) or 0 (severed). At any given time, only the currently connected queues may receive service.

Article [16] considers optimal load balancing in a distributed computing environment with several homogeneous unreliable processors that have limited state information. Processors may fail, with arbitrary failure and repair processes that are also independent of the state of the system. Optimal rate allocation in unreliable production networks are considered in [12]. In [9], discuss dynamic routing in heterogeneous unreliable queueing networks.

In [5], the problem of routing packets through several unreliable outgoing links to the same destination is investigated. Paper [18] considers information transfer processes in computer network with centralized control and unreliable outgoing links. To analyze the performance of multimedia service systems with unreliable resources and to estimate the required capacity of the systems, we used article [19] which is devoted to capacity planning model with using open queueing network. Article [24] offers the approximate analysis and application of an unreliable closed queueing network to model the performance of flexible manufacturing systems. The results of this article can be used to analyze flexible production systems with unreliable elements.

In this article, we consider an open exponential queueing network with a single class of demands and unreliable finite-server systems. Operable periods of systems are exponentially distributed random variables. From the moment of system failure the route matrix of queueing network changes so that demands do not arrive to this system. We assume that as soon as the average response time of network exceeds the limit value because of consecutive failure of systems all failed devices of systems begin to restore. All systems are restored during some time with an exponential distribution. Random process of breakdowns and repairs of systems is presented as Markov chain with continuous time. In addition, we use the absorbing Markov chain for time determination between the moments of repairs of systems. Stationary characteristics of unreliable systems and also average time between the moments of network repairs are obtained.

2 Queueing Network with Unreliable Systems

Consider open queueing network Γ with L systems and a single class of demands. System S_i is a finite-server queueing system with κ_i servers, service-time distribution at each server of S_i is exponential with rate μ_i , $i = 1, \dots, L$.

The customers arrive from the external source S_0 according to a Poisson process with arrival rate λ_0 . The network topology is defined by the adjacency

matrix $W = (w_{ij}), i, j = 0, \dots, L$, which corresponds to network of directed graph. Transitions of demands between systems are defined by the routing matrix $\Theta = (\theta_{ij}), i, j = 0, \dots, L$.

We assume that queueing systems fail independently of each other. Times between system failures $S_i, i = 1, \dots, L$, have exponential distributions with parameters γ_i . When a fault occurs at a system, all the demands there are destroyed immediately. At the same time the adjacency matrix W and, therefore, the routing matrix Θ , changes so that demands do not arrive to this system.

Let $b = (b_i), i = 1, \dots, L$, be the structure vector of queueing network Γ , where $b_i = 1$, if systems S_i is failed, or $b_i = 0$ otherwise. Let $\Theta(b)$ be the routing matrix of network Γ provided that the structure of network is defined by vector b .

Let structure vector b convert to a vector \tilde{b} , and routing matrix $\Theta(b)$ – to matrix $\Theta(\tilde{b})$ at the time of system S_m failure, $m \in \{1, \dots, L\}$. Elements of matrix $\Theta(\tilde{b})$ are given by

$$\begin{aligned} \theta_{mk}(\tilde{b}) &= 0, \quad k = 0, \dots, L, k \neq m, \\ \theta_{im}(\tilde{b}) &= 0, \quad i = 0, \dots, L, i \neq m, \\ \theta_{mm}(\tilde{b}) &= 1, \\ \theta_{ik}(\tilde{b}) &= \theta_{ik}(b)/(1 - \theta_{im}(b)), \quad i, k = 0, \dots, L, i, k \neq m. \end{aligned}$$

The state of queueing network Γ is vector $n = (n_i), i = 1, \dots, L$, where n_i denotes the number of demands in system S_i , then $n_m = 0$ at the time of failure of system S_m .

We will consider an expected value of response time $\tau_0(b)$. The value is the main performance measure defining quality of network Γ functioning with structure b and $\hat{\tau}_0$, the threshold value of response of network Γ .

Now we describe the process of network evolution with non-stationary structure. Systems fail independently of each other, causing changes in structure b of network Γ . Times between failures of systems have an exponential distribution. If $\tau_0(b) < \hat{\tau}_0$, then the broken systems are not restored. If $\tau_0(b) \geq \hat{\tau}_0$, then it is supposed that all disabled systems begin to restore, also including systems which will fail from the moment of the beginning of process of repair. The recovery time of the group of failed systems is an exponential random variable with parameter δ . It is supposed that recovery time of the group of systems does not depend on the group sizes.

It is required to find stationary performance measures of queueing network Γ .

3 Analysis Method of Queueing Network

The evolution of network Γ is represented as a set of realization of subnets $\Gamma(b)$. Each realization of a subnet is unambiguously defined by structure b and route matrix $\Theta(b)$. Other parameters of realization of subnets coincide.

We will designate the subnet $\Gamma(b)$ with the *connected configuration* if the route matrix of a set of serviceable (efficient) systems is irreducible.

Let B be the set of all possible structures b , $B = D \cup F \cup G$, where D , is the subset of structures b , forming subnets of $\Gamma(b)$ with a connected configuration, for which

$$\tau_0(b) < \hat{\tau}_0, \tag{1}$$

F is the subset of structures b , forming subnets of $\Gamma(b)$ with a connected configuration, for which there is a steady-state condition and

$$\tau_0(b) \geq \hat{\tau}_0, \tag{2}$$

G is the subset of structures b , forming subnets of $\Gamma(b)$ with violation of connectivity and without the steady-state condition, for this subset we assume $\tau_0(b) = \infty$.

Evolution of network Γ can be considered as proceeding in two parallel ways: the process of breakdowns of systems with their further restoration when the condition is satisfied (2) and the process of service and transitions of demands for networks enclosed in it between the working systems.

Changing of network structure leads to arising of a transition process. We assume that its duration is significantly less than the duration of functioning of network before the next change of structure. Therefore we will neglect transition process and to consider that the network $\Gamma(b)$ instantly changes into the steady-state condition. Stationary performance measures of networks $\Gamma(b)$, at fixed $b \in D$, can be received by known methods [17].

Renumber states of a set D so that $b^{(1)}$ be the structure of network at which all systems are efficient, numbering of other structures is random. Further, if it is clear from the context what state it is the numbering of states will go down. Let $\tilde{D} \subset D$, be the subset of structures $b^{(i)} \in D$, from which transition to a subset is possible. Let us determine $b^{(i)} \in \tilde{D}$ and $C \subset B \setminus D$, through $Z_C = \{b \in C : \|b - b^{(i)}\| = 1\}$. Then rate $\alpha(b^{(i)}, C)$ of transition from $b^{(i)} \in D$ in subsets $C = F$ and $C = G$ is equal to

$$\alpha(b^{(i)}, C) = \begin{cases} \sum_{\substack{b \in C \\ \|b^{(i)} - b\| = 1}} \sum_{k=1}^L \gamma_k (b_k^{(i)} - b_k), & \text{if } Z_C \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases}$$

Rate $\alpha(b^{(i)}, b^{(j)})$ of transitions from $b^{(i)} \in D$ in $b^{(j)} \in D$

$$\alpha(b^{(i)}, b^{(j)}) = \begin{cases} \sum_{k=1}^L \gamma_k (b_k^{(i)} - b_k^{(j)}), & \text{if } \|b^{(i)} - b^{(j)}\| = 1, \\ 0, & \text{otherwise,} \end{cases}$$

and rate $\alpha(b^{(i)})$ of exiting the state

$$\alpha(b^{(i)}) = \sum_{k=1}^L \gamma_k b_k^{(i)}.$$

The stochastic process of systems failures with the their subsequent group restoration subject to the condition (2) can be described by continuous time Markov chain M with the state space $\{1, \dots, d, d + 1, d + 2\}$, where $d = |D|$, the state $d + 1$ corresponds to a set F , $d + 2$ – to a set G .

Denote by $A = (a_{ij})$, $i, j \in \{1, \dots, d + 2\}$, the infinitesimal operator of the Markov chain, where

$$\begin{aligned}
 a_{ij} &= \alpha(b^{(i)}, b^{(j)}), \quad b^{(i)}, b^{(j)} \in D, \quad i \neq j, \quad i, j = 1, \dots, d, \\
 a_{i,d+1} &= \begin{cases} \alpha(b^{(i)}, F), & Z_F \neq \emptyset, \\ 0, & \text{otherwise,} \end{cases} \\
 a_{i,d+2} &= \begin{cases} \alpha(b^{(i)}, G), & Z_G \neq \emptyset, \\ 0, & \text{otherwise,} \end{cases} \\
 a_{ii} &= -\alpha(b^{(i)}), \quad i = 1, \dots, d.
 \end{aligned}$$

For $i = d + 1, d + 2$ we can write

$$\begin{aligned}
 a_{ij} &= \begin{cases} \delta, & j = 1, \\ 0, & j \neq 1, \end{cases} \\
 a_{ii} &= -\delta.
 \end{aligned}$$

The stationary distribution $\pi = (\pi_i)$, $i \in \{1, \dots, d + 2\}$ of the Markov chain, is the solution of the equation $\pi A = 0$ with the normalization condition $\sum \pi_i = 1$.

Then for queueing network Γ with connected configuration and with stationary distribution of states conditional stationary characteristics of systems are defined by

$$\chi_k = \frac{1}{\pi_{d+2}} \left(\sum_{b^{(i)} \in D} \chi_k(b^{(i)}) \pi_i + \pi_{d+1} \sum_{b \in F} \chi_k(b) \right), \quad k = 1, \dots, L,$$

where χ_k is the conditional integrated characteristic of system S_k , $\chi_k(b)$ – is integrated characteristic of system S_k of queueing network $\Gamma(b)$ with structure b , and conditional expectation of response time of network Γ

$$\tau_0 = \frac{1}{\pi_{d+2}} \left(\sum_{b^{(i)} \in D} \tau_0(b^{(i)}) \pi_i + \pi_{d+1} \sum_{b \in F} \tau_0(b) \right).$$

Time interval from the moment of the end of repair and until the beginning of the following renewal process of the network corresponds to the duration of stay of this network in the set of structures D . To calculate the expected duration of network stay in a set of structures D we will consider the absorbing Markov chain of M^* with a set of states $\{1, \dots, d + 1\}$, where $d + 1$ is the absorbing state corresponding to a set of structures $B \setminus D$. Duration of stay in a set of

structures D corresponds to time before absorption of a chain M^* in state $d + 1$. Infinitesimal operator A^* of a chain M^* has the following form:

$$A^* = \begin{bmatrix} S & \mathbf{S}^1 \\ \mathbf{0} & 0 \end{bmatrix},$$

where \mathbf{S}^1 is the vector of absorption rate with size $d \times 1$, $S = (a_{ij}), i, j = 1, \dots, d$, is the subgenerator, matrix of transition rates in a set of irretrievable states with a size $d \times d$, $\mathbf{0}$ is the zero vector with a size $1 \times d$. Initial distribution of a chain M^* is $\beta = (\beta_i), i = 1, \dots, d$, where $\beta_1 = 1$, and $\beta_i = 0, i = 2, \dots, d$.

It is known [10] that time before absorption of a chain M^* has phase distribution with the expected value

$$g = -\beta S^{-1} \mathbf{1},$$

where $\mathbf{1}$ is the identity vector column.

4 Numerical Results

We will consider an open exponential network with $L = 9$ queueing systems and a single class demands. The parameters of unreliable systems are given in Table 1, the rate of repair for the failed systems $\delta = 0.5$.

Table 1. Parameters of queueing systems

Number of system, i	1	2	3	4	5	6	7	8	9
κ_i	2	1	1	2	1	1	1	2	1
μ_i	1.00	1.90	1.80	1.00	1.70	2.00	1.90	1.00	1.80
γ_i	0.01	0.02	0.07	0.03	0.05	0.02	0.005	0.01	0.06

Route matrix of the network: $\theta_{02} = 0.3, \theta_{03} = 0.4, \theta_{04} = 0.3, \theta_{14} = 0.7, \theta_{15} = 0.3, \theta_{25} = 0.8, \theta_{26} = 0.2, \theta_{36} = 1, \theta_{47} = 1, \theta_{57} = 0.1, \theta_{58} = 0.8, \theta_{59} = 0.1, \theta_{69} = 1, \theta_{70} = 1, \theta_{80} = 1, \theta_{90} = 1$.

The network research with absolutely reliable systems and unreliable systems in dependence on rate of the entering flow is conducted.

In the first case the threshold value for unreliable network was defined as $\hat{\tau}_0 = 10$.

The composition of a set D does not change in the value range λ_0 , as the threshold value $\hat{\tau}_0 = 10$ exceeds τ_0 a network with unreliable systems in case of the maximum entering flow of requirements in a network.

The function τ_0 for network with unreliable systems, shown in Fig. 1, represents the smooth ascending curve.

The network research with unreliable systems in case of threshold value $\hat{\tau}_0 = 5$ and $\hat{\tau}_0 = 3$ depending on λ_0 are conducted. From results of the experiments provided in Figs. 2 and 3 it is clear that τ_0 piecewise continuous function

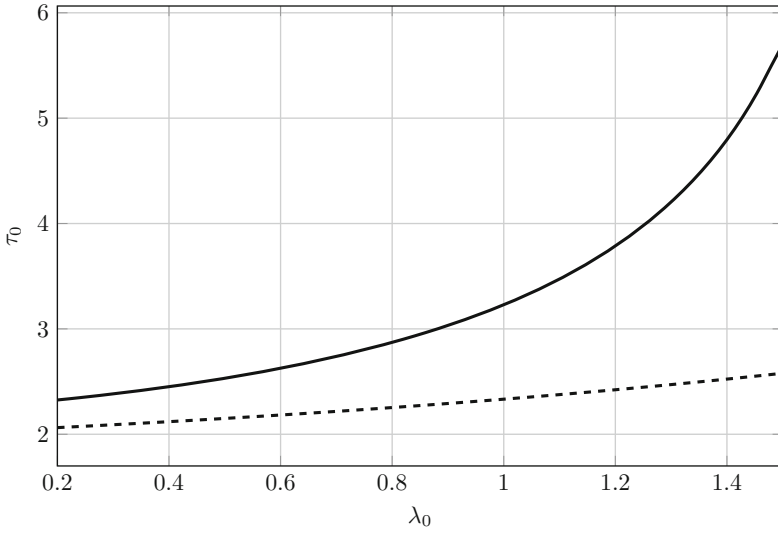


Fig. 1. Dependence of τ_0 on λ_0 at $\hat{\tau}_0 = 10$ for reliable network (*dashed line*) and unreliable network (*solid line*)

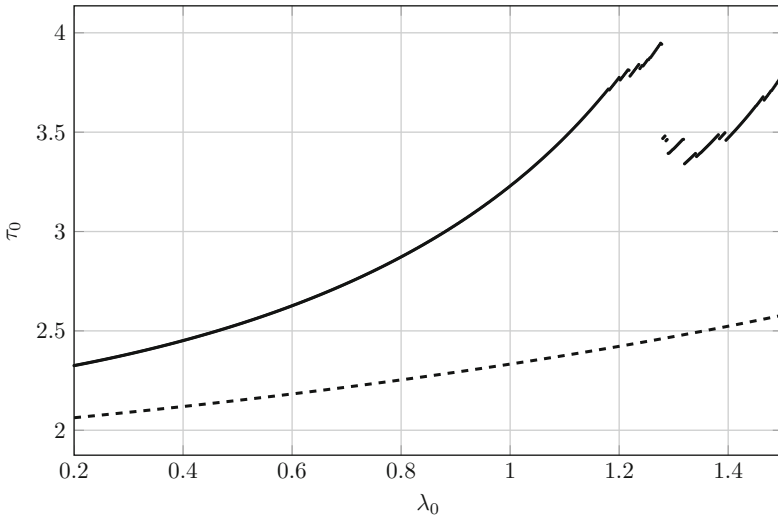


Fig. 2. Dependence of τ_0 on λ_0 at $\hat{\tau}_0 = 5$ for reliable network (*dashed line*) and unreliable network (*solid line*)

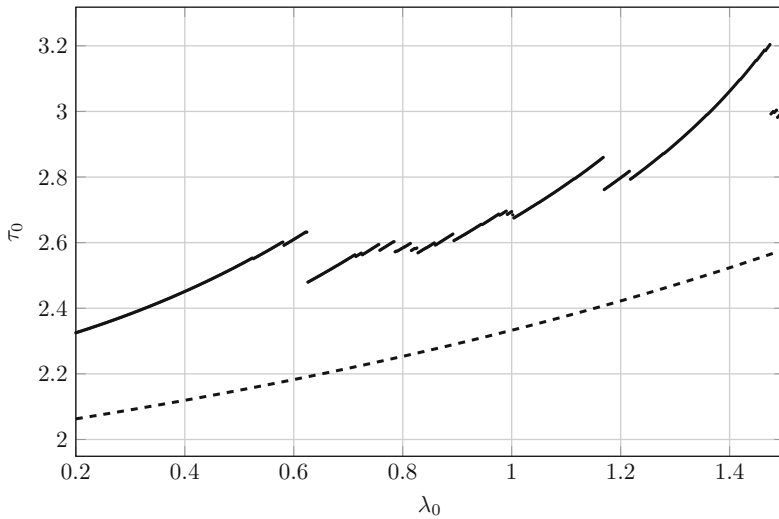


Fig. 3. Dependence of τ_0 on λ_0 for $\hat{\tau}_0 = 3$ for reliable network (*dashed line*) and unreliable network (*solid line*)

(Fig. 4). In discontinuity points the reduction of number of states from set D is made because of states in which τ_0 becomes greater, than $\hat{\tau}_0$.

Figure 5 illustrates dependence of the expected value g on $\hat{\tau}_0$. Figure 6 illustrates dependence of the expected value g on λ_0 .

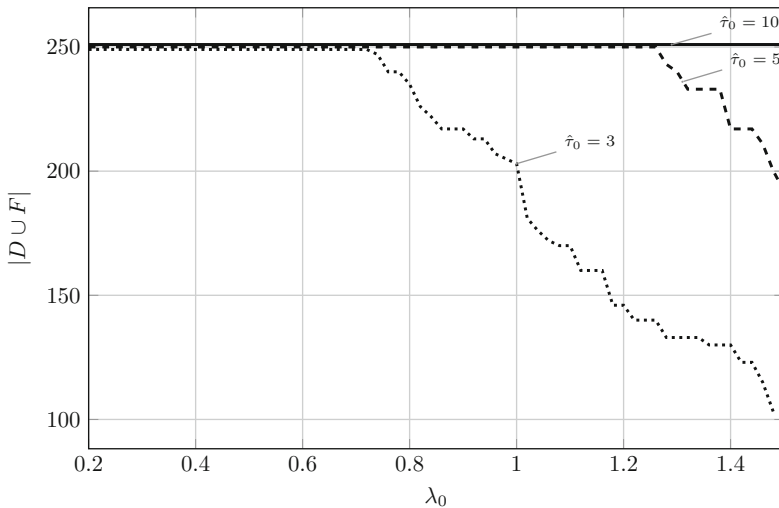


Fig. 4. Dependence of the cardinality of the set $D \cup F$ on λ_0 for unreliable network

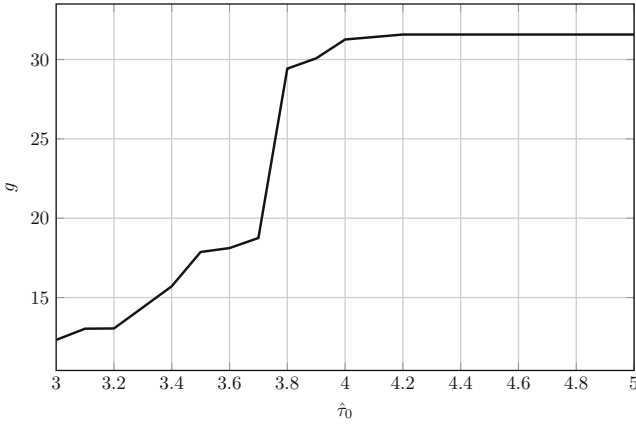


Fig. 5. Dependence of g on $\hat{\tau}_0$

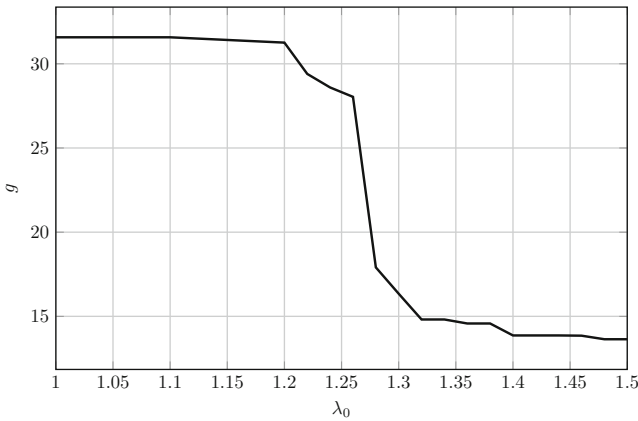


Fig. 6. Dependence of g on λ_0

5 Conclusion

The open queueing network with consistently disconnected multiserver systems is considered in the paper.

System faults occurs sequentially with exponential times per fault. When a fault occurs at a system the structure vector is changed. For each structure of the network defined the expectation of response time. Systems are not restored until the expectation of response time for the network exceeds the set threshold value. The recovery time of all failed systems is an exponential random variable. It is necessary that during recovery time any other systems do not fail.

The method is developed for the analysis of the queueing network. Average performance measures of systems and expectation of duration between the recovery moments of systems are obtained. The method can be used for the analysis

and maintenance of information transfer networks with unreliable elements. In these networks the repair of failed elements begins after the response time of networks reaches critical level.

References

1. Akyildiz, I.F., Liu, W.: Performance optimization of distributed-system models with unreliable servers. *IEEE Trans. Reliab.* **39**(2), 236–243 (1990)
2. Bambos, N., Michailidis, G.: Queueing networks of random link topology: stationary dynamics of maximal throughput schedules. *Queueing Syst.* **50**, 5–52 (2005)
3. Blake, J.T., Reibman, A.L., Trivedi, K.S.: Sensitivity analysis of reliability and performability measures for multiprocessor systems. In: *ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pp. 177–186. ACM, New York (1988)
4. Chao, X.: A queueing network model with catastrophes and product form solution. *Oper. Res. Lett.* **18**, 75–79 (1995)
5. Economides, A.A., Silvester, J.A.: Optimal routing in a network with unreliable links. In: *IEEE INFOCOM 1988*, pp. 288–297 (1988)
6. Ekici, E., Akyildiz, I.F., Bender, M.D.: A distributed routing algorithm for data-gram traffic in LEO satellite networks. *IEEE/ACM Trans. Netw.* **9**(2), 137–147 (2001)
7. Fokina, N.P., Tananko, I.E.: The method of routing in queueing networks with variable topology. *Izv. Saratov Univ. (N.S.), Ser. Math. Mech. Inform.* **13**(2), 82–88 (2013). (in Russian)
8. Gershwin, S.B., Burman, M.H.: A Decomposition Method for Analyzing Inhomogeneous Assembly/Disassembly Systems. *Annal. Oper. Res.* **93**, 91–115 (2000)
9. Glazebrook, K.D., Kirkbride, C.: Dynamic routing to heterogeneous collections of unreliable servers. *Queueing Syst.* **55**, 9–25 (2007)
10. He, Q.-M.: *Fundamentals of Matrix-Analytic Methods*. Springer, New York (2014). <https://doi.org/10.1007/978-1-4614-7330-5>
11. Jia, L., Rajaraman, R., Scheideler, C.: On local algorithms for topology control and routing in ad hoc networks. In: *Fifteenth Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA 2003)*, pp. 220–229. ACM, New York (2003)
12. Kouikoglou, V.S.: Optimal rate allocation in unreliable, assembly/disassembly production networks with blocking. *IEEE Trans. Robot. Autom.* **16**(4), 429–434 (2000)
13. Kumar, A., Karnik, A.: Performance analysis of wireless ad-hoc networks. In: Ilyas, M., Dorf, R.C. (eds.) *The Handbook of Ad Hoc Wireless Networks*, pp. 83–101. CRC Press Inc., Boca Raton (2003)
14. Levantesi, R., Matta, A., Tolio, T.: Performance evaluation of continuous production lines with machines having different processing times and multiple failure modes. *Perform. Eval.* **51**, 247–268 (2003)
15. Liu, J., Wu, A.: Analytical analysis of production lines with parallel machines. In: *30th Chinese Control Conference*, pp. 5480–5483 (2011)
16. Liu, Z., Righter, R.: Optimal load balancing on distributed homogeneous unreliable processors. *Oper. Res.* **46**(4), 563–573 (1998)
17. Mitrophanov, Yu.I.: *Analysis of Queueing Networks*. Nauchnaya kniga, Saratov (2005). (in Russian)
18. Moose Jr., R.L.: Modeling networks with dynamic topologies. *ORSA J. Comput.* **1**(4), 223–231 (1989)

19. Park, K., Kim, S.: A capacity planning model of unreliable multimedia service systems. *J. Syst. Softw.* **63**, 69–76 (2002)
20. Sauer, C., Daduna, H.: BCMP Networks with Unreliable Servers. Preprint No. 2003-01, Schwerpunkt, Mathematische Statistik und Stochastische Prozesse. Universitat Hamburg (2003)
21. Sharony, J.: An architecture for mobile radio networks with dynamically changing topology using virtual subnets. *Mob. Netw. Appl.* **1**(1), 75–86 (1996)
22. Sommer, J., Berkhout, J., Daduna, H., Heidergott, B.: Analysis of Jackson networks with infinite supply and unreliable nodes. *Queueing Syst.* **87**(1–2), 181–207 (2017)
23. Tsitsiashvili, G.S., Osipova, M.A.: Limiting distributions in queueing networks with unreliable elements. *Prob. Inf. Trans.* **44**(4), 385–394 (2008)
24. Vinod, B., Altiok, T.: Approximating unreliable queueing networks under the assumption of exponentiality. *J. Opl. Res. Soc.* **37**(3), 309–316 (1986)



An Infinite-Server Queueing $MMAP_k|G_k|\infty$ Model in Semi-Markov Random Environment Subject to Catastrophes

K. Kerobyan¹, R. Covington², R. Kerobyan³, and K. Enakoutsa^{1(✉)}

¹ Department of Mathematics, California State University, Northridge,
18111 Nordroff Street, Northridge, CA 91330, USA
{khanik.kerobyan,koffi.enakoutsa}@csun.edu

² Department of Computer Science, California State University, Northridge,
18111 Nordroff Street, Northridge, CA 91330, USA

³ Department of Computer Science, University of California San Diego,
9500 Gilman Drive, La Jolla, CA 92093, USA

Abstract. In the present paper the infinite-server $MMAP_k|G_k|\infty$ queueing model with random resource vector of customers, marked MAP arrival and semi-Markov (SM) arrival of catastrophes is considered. The joint generating functions (PGF) of transient and stationary distributions of number of busy servers and numbers of different types served customers, as well as Laplace transformations (LT) of joint distributions of total accumulated resources in the model at moment and total accumulated resources of served customers during time $[0, t)$ interval are found. The basic differential and renewal equations for transient and stationary PGF of queue sizes of customers are found.

Keywords: Marked MAP · Infinite-server queue model
Catastrophe · Resource vector

1 Introduction

From the time of Erlang's pioneering research [1], queueing models with many servers, in particular, the infinite-server models, have been widely used for modelling and performance evaluation of wired and wireless computer and telecommunication networks [2]. As shown by large number of measurements the traffic of modern computer networks has self-similar nature and can be characterized by the heterogeneousness, the non-stationarity, the burstiness, and the correlations [3]. Network traffics in queue are generally described by traffic models based on finite Markovian Processes: Markov Arrival Process (MAP), Batch MAP (BMAP), Marked MAP (MMAP) and their generalizations [3,4]. The MMAP and MAP arrivals properties and their applications are presented in [4–6] and are not duplicated here.

The $BMAP|G|_\infty$ model studied by [7,8]. By using semi-Markov processes (SMP), matrix analytic methods the PGF of the number of busy servers and its moments are considered. The models with phase type arrival $PH|G|_\infty$ and Markov modulated arrival $MMPP|G|_\infty$ are considered in [9,10]. The queueing model $M_k|M_k|_\infty$ with correlated k heterogeneous customers in a batch and exponential service time is studied by [11]. The joint PGF of the number of type k customers being served in the system is derived explicitly by solving partial differential equations. The generalization of this model for general service time $M_k|G_k|_\infty$ Poisson arrival of customers and $BM_k|G_k|_\infty$ with k correlated heterogeneous customers in a batch is considered in [12,13]. In steady state, the joint PGF of queue length of customers by using collective mark method CMM and conditional expectations is derived in [13]. The model $MAP_k|G_k|_\infty$ with the structured batch arrival of k types of customers is considered in [14]. In steady state, the differential equations for PGF of queue length and its moments are obtained. The first and second order asymptotes of queue length for the models $MAP|G|_\infty$, $MMPP|G|_\infty$, $G|G|_\infty$ based on supplementary variable method are studied in [15].

To evaluate the impact of network environment on networks performance metrics the infinite server models in the random environment (RE) are applied. The queue size distribution of the model $M|G|_\infty$ in semi-Markov (SM) environment is studied in [16–19]. By the method of supplementary variable first and second order asymptotes of queue size distribution are obtained [16]. The stochastic decomposition formula for queue length distribution is obtained in [17]. The queue $M|G|_\infty$ in random environment with clearing mechanism is studied in [17,18]. The environmental clearing process is modeled by an m -state irreducible SMP. The transient and steady-state queue length distributions by using renewal arguments are obtained. The $MMAP_k|G_k|_\infty$ queue in SM environment and catastrophes is studied in [20]. The PGFs of joint distributions of queue size and number of served customers by using renewal arguments and differential equations are found. The infinite-server queue $MMAP_k(t)|G_k|_\infty$ with Poisson stream catastrophes and nonhomogeneous marked MAP arrival of customers is studied in [21]. The PGF of queue length of different types of customers is obtained. The model $M|M|_\infty$ with disasters in steady-state is studied in [22,23].

In many applications of queueing models such as computer and communication networks and systems, production systems, transportation systems, economics, finances and insurance systems the customers characterize by vector of requesting resources which components can be deterministic or random quantities. For instance one component can describe number of servers necessary to serve the customer, second amount of time necessary to serve the customer, next the volume of space to save the customer and so on. Also the customers must have some features necessary to be accepted by system. Some components of resource vectors can be discrete (number of servers, amount of parts) and others can be continuous (space-volume to save, amount of finances, power). Despite importance of resource models of customers in queueing theory, there are very few works devoted to research such kind of models, see for example [20,24–29].

The main methods to study the infinite-server queues are: supplementary variables method [30], the method which based on conditional expectations [30,31], and collective marks method (CMM) method [30,32,33]. The last method is also called “supplementary event” [30] or “catastrophes” method [32,33] and has been used successfully for queue models with priorities [33]. CMM have been used for infinite-server models in [13] for $BM_k|G_k|\infty$ queue with Poisson arrival of batches and in [34] for $M|SM|\infty$ queue. In [34] the method is mentioned but does not used to obtain some results.

In this paper we consider some generalizations of [17,20,24] results for infinite-server $MMAP_k|G_k|\infty$ queue in random environment and catastrophes. The joint PGF of transient and stationary distributions of number of busy servers and numbers of served customers, as well as the LT of joint distributions of total accumulated resource in the model and total accumulated resource vectors of served customers during time interval are found. The renewal equations for transient and stationary PGF of queue sizes of different types of customers are found as well. All results are obtained using CMM and renewal process methods.

2 Model Description

We consider an infinite server $MMAP_k|G_k|\infty$ queue model in random environment (RE) with K types of customers and catastrophes. The RE operates according to stationary, irreducible semi-Markov process (SMP) $\xi(t)$, $t \leq 0$ with finite state space $S = \{1, 2, \dots, k\}$. The SMP is given by the vector of initial distribution $p^0 = \{p_0^i, i \in S\}$ and SM matrix $Q(t) = \|Q_{ij}(t)\|$, $t \geq 0, 1, j \in S$. Customers arrive according to homogeneous marked MAP which is given by the sequence of characteristic matrices $\{D_0, D_h, h \in C^0\}$. Here C^0 is a finite or counting set of arriving batches $\mathbf{h} = (h_1, h_2, \dots, h_k)$, $\mathbf{h} \in C^0$, and h_r is a number of type r customers, $0 \leq h, 1 \leq r \leq K$ in a batch. D_0 is a non-singular matrix with negative diagonal elements and D_h are non-negative matrices, $\mathbf{h} \in C^0$. The phase process (PP) $J(t)$ of MMAP is an irreducible Markov process (MP) with generating matrix (GM) D and finite set of states E . D is a matrix of $m \times n$ size.

$$D = D_0 + \sum_{\mathbf{h} \in C^0} D_{\mathbf{h}}, \quad D_e = 0, \quad \pi D = 0, \quad \pi e = 1,$$

where e is the column vector with all components one, π is the vector of stationary distribution $\pi = (\pi_1, \dots, \pi_m)$ of PP $J(t)$.

The service of arriving customers begins immediately. Let the random variable (r.v.) γ_r be a service time of type r customers, and $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_k)$ is a vector of service times. The components of γ are independent, identically distributed (i.i.d.) r.v.s which depend on type of the customer and state of environmental SMP. R.v. γ_r has $B_r(t) = P(\gamma_r < t)$ general distribution and finite mean value $\bar{\gamma}_r$, $1 \leq r \leq K$. Each arriving customer is characterized by k -dimensional random volume (resource) vector $\xi_r = (\xi_{1r}, \dots, \xi_{kr})$ and each departing customer is characterized by random vector $\sigma_r = (\sigma_{1r}, \dots, \sigma_{kr})$ with non-negative components $1 \leq r \leq K$. Let $C_r(\mathbf{x}) = \mathbf{P}(\xi_{1r} \leq \mathbf{x}_1, \dots, \xi_{kr} \leq \mathbf{x}_k)$

and $G_r(\mathbf{x}) = \mathbf{P}(\sigma_{\mathbf{1r}} \leq \mathbf{x}_1, \dots, \sigma_{\mathbf{kr}} \leq \mathbf{x}_k)$ be the joint distributions of resource vectors ξ_r and σ_r , where $\mathbf{x} = (\xi_1, \dots, \xi_k)$. We assume that the service time vector γ and the resource vectors $\xi_r = (\xi_{1r}, \dots, \xi_{kr})$, $\sigma_r = (\sigma_{1r}, \dots, \sigma_{kr})$ are mutually independent.

When SMP $\xi(t)$, $t \leq 0$ jumps from state i to the state r all customers in the model are instantly flashed out and the model jumps into empty state. Let consider the related with MMAP counting processes $N(t), N_s(t), M(t)$: $N(t) = (N_1(t), N_2(t), \dots, N_k(t))$, $N_s(t) = (N_{1s}(t), N_{2s}(t), \dots, N_{ks}(t))$, $M(t) = (M_1(t), M_2(t), \dots, M_k(t))$, where $N_r(t)$ and $M_r(t)$ are the number of type r customers arriving and serving in time interval $[0, t)$, and $N_{rs}(t)$ is the number of customers being in service at moment t . Let $\beta(t) = (\beta_1(t), \dots, \beta_k(t))$ and $\alpha(t) = (\alpha_1(t), \dots, \alpha_k(t))$ be the vectors of total resource served during interval $[0, t)$ and accumulated in the model at moment t . The components of $\beta(t)$, $\alpha(t)$, $N_s(t)$ and $M_s(t)$ vectors are defined as:

$$\beta_r(t) = \sum_{i=1}^{M_r(t)} \sigma_{ri}, \quad \alpha_r(t) = \sum_{i=1}^{N_r(t)} \zeta_{ri},$$

$$N_{rs}(t) = \sum_{h \in C^0} N_{hr}^s(t), \quad M_r(t) = \sum_{h \in C^0} M_{hr}(t), \quad r = 1, 2, \dots, K.$$

Suppose that at initial time $t = 0$ the model is empty, $N(0) = 0$, $M(0) = 0$, $\alpha(0) = 0$, and $\beta(0) = 0$.

3 The Counting Process

Let consider the counting process (CP) $\{N(t), J(t), t \leq 0\}$ with the matrix $P(n, t)$ of transition probabilities: $P(n, t) = \|p_{ij}(n, t)\|$, $p_{ij}(n, t) = P(N, t) = n$, $J(t) = j | J(0) = i$, $1 \leq i, j \leq m$, where $n = (n_1, \dots, n_k)$: n_i are non-negative integers. Let define the following generating functions (GF) $D(z)$, $P(z, t)$,

$$D(z) = D_0 + \sum_{h \in C^0} z^h D_h, \quad |z_r| \leq 1, \quad 0 \leq r \leq K, \quad |z| \leq 1, \quad 0 \leq r \leq K,$$

$$P(z, t) = \sum_{n \geq 0} z^n P(n, t),$$

where $z = (z_1, z_2, \dots, z_k)$ and $z^h = (z_1^{h_1}, z_2^{h_2}, \dots, z_k^{h_k})$.

Theorem 1. The PGF of counting process $\{N(t), J(t), t \geq 0\} P(z, t)$ satisfies the basic differential equation

$$\frac{\partial}{\partial t} P(z, t) = D(z)P(z, t), \quad |z| \leq 1, \tag{1}$$

with initial conditions $P(z, 0) = 1$. The solution of differential equation Eq. (1) is given by

$$P(z, t) = e^{D(z,t)t}. \tag{2}$$

Proof. The transition probabilities $\{P(n, t), n \leq 0\}$ of CP $N(t)$ satisfy the following Kolmogorov backward differential equations

$$\frac{d}{dt}P(n, t) = P(n, t)D_0 + \sum_{h \leq n, h \in C^0} P(n - h, t)D_h, \quad n \geq 0.$$

with initial condition $P_i(n, 0) = 0, n > 0, P_i(0, 0) = 1, i = 1, 2, \dots, k$. Pre-multiplying each equation by corresponding z^n after summation we get differential equations for PDF $P(z, t)$. The solution of this equation in matrix exponential form is given by Eq. (2).

By Theorem 1, we can find moments of CP $\{N(t), J(t), t \geq 0\}$. For example, the mean $E[N_h(t)]$ and variance $V[N_h(t)]$ of CP $N(t)$ can be found explicitly

$$E[N_h(t)] = \lambda_h t + \theta(e^{Dt} - I)(D - e\pi)^{-1}D_h e, \quad t \geq 0,$$

where θ is a initial distribution of PP $J(t), \lambda_h = \pi D_h e$ is the stationary arrival rate of type h batches.

$$\begin{aligned} Var[N_h(t)] &= [\lambda_h - 2\lambda_h^2 - 2\pi D_h (D - e\pi)^{-1}D_h e]t \\ &+ 2\pi D_h (D - e\pi)^{-1}(e^{Dt} - I)(D - e\pi)^{-1}D_h e, \quad t \geq 0. \end{aligned}$$

If as an initial distribution use vector π then for stationary arrival rate of all batches of customers we get $E[N_h(t)] = \lambda_h t$. The stationary arrival rate of all types of customers is given by $\lambda = \sum_h D_h e$. The stationary arrival rates of type r customers and all customers regardless of type arriving in $[0, t)$ are given by

$$E[N_r(t)] = \lambda_r t, \quad E[N(t)] = \lambda t = \sum_{r=1}^K \lambda_r t,$$

where $\lambda_r = \sum_{n=1}^{\infty} n \sum_{h \in C^0, h_r=n} \pi D_h e$.

4 Thinning MMAP

Let consider the following Bernoulli thinning process of MMAP. Each type r customer which arrives at moment t can join to main stream by probability $p_r(t)$ and can be ignored by probability $1 - p_r(t)$. It can be shown (see e.g. [5, 34]), that the main stream is an MMAP process with characteristic matrices $D_{0,T}(t), D_{h,T}(t)$.

Lemma. The thinned process is a MMAP which counting process has matrix GF $D_T(z, t)$ and PGF $P_T(z, t)$, which are defined as follow

$$D_T(z, t) = \sum_{h_1 \geq 0} \sum_{\substack{h_2 \geq 0 \\ h_1 + h_2 + \dots + h_k \geq 1}} \dots \sum_{h_r \geq 0} D_h \prod_{r=1}^k [1 - p_r(t) + z_r p_r(t)]^{h_r}, \quad (3)$$

$$P_T(z, t) = e^{\int_0^t D_T(z, t) dx}.$$

The proof of the lemma is based on Theorem 1, MMAP properties and Bernoulli thinning operation properties [5, 34]. Let just interpret the GF $D_T(z, t)$ by CMM. Let each type r customer of thinned stream marks *red* with the probability z_r and marks *blue* with probability $1 - z_r$. Then the left side of Eq. (3) is the total rate of not *blue* customers at moment t . Right side means the batch $h = (h_1, h_2, \dots, h_k)$ arrives with rate D_h at moment t . In this batch each type r customer is *red* with probability $1 - p_r(t) + z_r p_r(t)$ and all h_r type r customers in this batch are *red* with probability $[1 - p_r(t) + z_r p_r(t)]^{h_r}$. The probability that in arriving batch h are not any *blue* customers is $\prod_{r=1}^k [1 - p_r(t) + z_r p_r(t)]^{h_r}$. Then the product $D_h \prod_{r=1}^k [1 - p_r(t) + z_r p_r(t)]^{h_r}$ is a rate of *red* batch of size h arriving at time t . Taking the sum over all possible batches arriving at moment t we obtain $D_T(z, t)$ the rate of not *blue* customers arriving at moment t .

The subject of our interest is the joint distribution

$$P(n, m, x, y, t) = P(N_s(t) = n, M(t) = m, \alpha(t) \leq x, \beta(t) \leq y).$$

5 Model Analysis

Let suppose that the environmental SMP $\xi(t)$, $t \geq 0$ is in state $i \in S$ and consider the dynamic of the model during time interval $[u, t)$. Each type r customer arriving at moment u will be in service at moment t by probability $1 - B_{ri}(t - u)$ and will finish its service before moment t by probability $B_{ri}(t - u)$. Let $A_{jk}^i(n, m, x, y, u, t)$ be the joint probability that n customers are in service at moment t , and m customers are already served in $[u, t)$, total accumulated resources in the model at moment t is $\alpha(t) \leq x$ and the total accumulated served resource during interval $[u, t)$ is $\beta(t) \leq y$, PP $J(0)$ is in phase $i \in E$ under condition that at initial moment $t = 0$ the model was empty, and PP $J(0)$ was in phase $k \in E$: $A_{jk}^i(n, m, x, y, u, t) = P(N_i^s(u, t) = n, M_i(u, t) = m, \alpha(t) \leq x, \beta(t) \leq y, J(u) = j | N_i^0 = 0, M_i(0), J(0) = k)$. Let be $\tilde{A}^i(z_1, z_2, s_1, s_2, u, t) = \|\tilde{A}_{jk}^i(z_1, z_2, s_1, s_2, u, t)\|$ the matrix which elements are Laplace–Stieltjes transformation (LST) and z transformation of $A_{jk}^i(z_1, z_2, s_1, s_2, u, t)$; $\tilde{F}_{ri}^i(s_1)$ and $\tilde{G}_{ri}^i(s_1)$ be the LST of $F_{ri}^i(s_1)$ and $G_{ri}^i(s_1)$. For homogeneous model we have $\tilde{A}^i(z_1, z_2, s_1, s_2, t) = \tilde{A}^i(z_1, z_2, s_1, s_2, u, t)$, e.g. see [7].

$$\tilde{A}^i(z_1, z_2, s_1, s_2, t) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} z_1^n z_2^m \int_0^{\infty} \int_0^{\infty} e^{-s_1 x - s_2 y} A^i(n, m, dx, dy, t),$$

$$|z_1| \leq 1, |z_2| \leq 1,$$

$$\tilde{F}_{ri}^i(s_1) = \int_0^\infty e^{-s_1 x} dF_{ri}(x), \tilde{G}_{ri}^i(s_2) = \int_0^\infty e^{-s_2 y} dG_{ri}(y).$$

For the PGF $A^i(z_1, z_2, s_1, s_2, t)$ of the model $MMA Pr|Gr|$ can be proved the following result (e.g. see [20, 35, 36]).

Theorem 2. The PGF $\tilde{A}^i(z_1, z_2, s_1, s_2, t)$ satisfy the following basic differential and integral equations

$$\begin{aligned} \tilde{A}^i(z_1, z_2, s_1, s_2, t) &= e^{D_0(i)t} + \\ &\int_0^t e^{D_0(i)u} \tilde{S}_i(z_1, z_2, s_1, s_2, u) \tilde{A}^i(z_1, z_2, s_1, s_2, t-u) du, \end{aligned} \quad (4)$$

$$\frac{\partial}{\partial t} \tilde{A}^i(z_1, z_2, s_1, s_2, t) = [D_0(i) + \tilde{S}_i(z_1, z_2, s_1, s_2, t)] \tilde{A}^i(z_1, z_2, s_1, s_2, t), \quad i \in S, \quad (5)$$

with initial conditions $\tilde{A}^i(z_1, z_2, s_1, s_2, t) = \mathbf{I}$.

Here

$$\tilde{S}_i(z_1, z_2, s_1, s_2, t) = \sum_{h=1}^{\infty} D_{hi} \prod_{r=1}^K [z_{r2} \tilde{G}_{ri}(s_2) B_{ri}(t) + z_{r1} \tilde{F}_{ri}(s_1) (1 - B_{ri}(t))]^{h_r}.$$

The proof of Eq. (4) can be done by using the method of collective marks [32] or renewal arguments [35–37]. For the proof of Eq. (4) we can use Eqs. (1) and (3). Let consider the proof by the CMM. First of all let note that each type r customer which arrives at moment u will be served up to moment t with probability $B_r(t-u)$ or will be in the model at moment t with probability $1 - B_r(t-u)$. We mark each served type r customer *red* or *blue* with probabilities $z_{2r} \tilde{G}_r(s_2)$ and $1 - z_{2r} \tilde{G}_r(s_2)$ resp. Alike, we mark each serving in the model type r customer “*red*” or “*blue*” with probabilities $z_{1r} \tilde{F}_r(s_1)$ and $1 - z_{1r} \tilde{F}_{1r}(s_1)$ resp. Then on the left side of Eq. (4) $\tilde{A}^i(z_1, z_2, s_1, s_2, t)$ is the probability that at moment t in the model there are not *blue* customers. This event is possible if either the model is free and during interval of time $[0, t)$ there are not any arrivals of customers (with probability $e^{D_0(i)u}$) or at moment u a batch h arrives (with probability $e^{D_0(i)u} D_h(i) du$), all customers in the batch are *red* (with probability $\prod_{r=1}^K [z_{r2} \tilde{G}_{ri}(s_2) B_{ri}(u) + z_{r1} \tilde{F}_{ri}(s_1) (1 - B_{ri}(u))]^{h_r}$), and in interval $t-u$ in the model are not any *blue* customers (with probability $\tilde{A}^i(z_1, z_2, s_1, s_2, t-u)$). After using the total probabilities rule we derive the Eq. (5).

Theorem 3. The solution of Eqs. (4) and 5 is given by

$$A^i(z_1, z_2, s_1, s_2, t) = \exp \left\{ \int_0^t [D_0(i) + \tilde{S}_i(z_1, z_2, s_1, s_2, u)] du \right\}, \quad (6)$$

$$|z_1| \leq 1, \quad |z_2| \leq 1.$$

The proof is based on thinning lemma.

When $z_2 = 1, s_2 = 0$ from Theorem 3 we obtain the LST of PGF joint distribution number of busy servers and total accumulated resources in the model at moment t : $\tilde{P}(z_1, s_1, t, i) = \tilde{A}^i(z_1, 1, s_1, 0, t)$,

$$\tilde{P}(z_1, s_1, t, i) = \exp \left\{ \int_0^t \left[D_0(i) + \tilde{S}_i(z_1, s_1, u) \right] du \right\},$$

where $\tilde{S}^i(z_1, s_1, t) = \sum_{h=1}^{\infty} D_{hi} \prod_{r=1}^K \left[B_{ri}(t) + z_{ri} \tilde{F}_{ri}(s_1)(1 - B_{ri}(t)) \right]^{h_r}$. When $z_1 = 1, s_1 = 0$, from (6) we obtain the LST of PGF joint distribution number of served customers and total served resources during interval $[0, t)$:

$$\tilde{W}^i(z_2 = 1, s_2, t) = e^{\int_0^t [D_0(i) + \tilde{S}^i(z_2, s_2, u)] du},$$

where $\tilde{S}^i(z_2, s_2, t) = \sum_{h=1}^{\infty} D_{hi} \prod_{r=1}^K \left[z_{r2} \tilde{G}_{ri}(s_2) B_{ri}(t) + (1 - B_{ri}(t)) \right]^{h_r}$.

If we suppose that at time $t = 0$, there are $h_0 = (h_{01}, h_{02}, \dots, h_{0k})$ initial customers in the model then for $\tilde{A}^i(z_1, z_2, s_1, s_2, t)$ we get

$$\tilde{A}^i(z_1, z_2, s_1, s_2, t) = \prod_{r=1}^K \left[z_{r2} \tilde{G}_{ri}(s_2) B_{ri}(t) + (1 - B_{ri}(t)) \right]^{h_{0r}} \cdot e^{\int_0^t [D_0(i) + \tilde{S}_i(z_1, z_2, s_1, s_2, u)] du}.$$

Using Theorem 3 we can find particular cases of PGFs for $PH|G|\infty$ and $BMAP|G|\infty$ models [7,9].

Let consider an infinite-server queue $BG_k|G_k|\infty$ with general distributed interarrival time of batches and K types of customers. Let $A(t)$ be the DF of interarrival epochs between batches, and $B_r(t)$ is a service time DF of type r customers. In each inter-arrival epoch by $a(n) = a(n_1, \dots, n_K)$ probability generates a batch n with n_1 customers of type 1, ..., n_K customers of type K , where $\sum_{n_1} \sum_{n_2} \dots \sum_{n_k} a(n_1, n_2, \dots, n_k) = 1, a(0, 0, \dots, 0) = 0$. Then for PGF $\tilde{A}(z_1, z_2, s_1, s_2, t)$ of the model $BG_k|G_k|\infty$ we can prove the result below.

Theorem 4. The PGF $\tilde{A}(z_1, z_2, s_1, s_2, t)$ satisfies the following basic integral equations

$$A(z_1, z_2, s_1, s_2, t) = 1 - A(t) + \int_0^t \tilde{S}(z_1, z_2, s_1, s_2, u) \tilde{A}(z_1, z_2, s_1, s_2, t - u) dA(u), \tag{7}$$

where

$$\tilde{S}(z_1, z_2, s_1, s_2, u) = \sum_{\substack{n=0 \\ n_1+n_2+\dots+n_k=1}}^{\infty} a(n) \prod_{r=1}^K \left[z_{r2} \tilde{G}_r(s_2) B_r(t) \right]$$

$$+z_{r1}\tilde{F}_r(s_1)(1 - B_r(t))\Big]^{n_r}.$$

Proof. Let mark the customers as in case of Theorem 2. Then $\tilde{A}(z_1, z_2, s_1, s_2, t)$ in left side of (7) is the probability of event “no blue customers was served in the model during interval of time $[0, t)$ no blue customer is serving in the model at moment t ”. This event can be realized in two mutually independent ways: either “the first customer arrives into empty model after time t ” (with probability $1 - A(t)$) or “first batch of customers arrives at moment $u, u < t$ ” (with probability $dA(u)$), “this batch includes customers of different types and all customers in the batch are red” (with probability $\prod_{r=1}^K [z_{r2}\tilde{G}_r(s_2)B_r(t) + z_{r1}\tilde{F}_r(s_1)(1 - B_r(t))]$) ^{n_r} .

Then applying the total probabilities rule we get the result.

Remark. Let consider the service policy when we do not distinguish the customers of different batches and all customers in the batch serve together as one customer. In this case the MMAP transforms into equivalent MAP with characteristic matrices $D_0 = C, D_1 = \sum_{h \in C^0} D_h$. Let the r.v. γ_0 be a service time of batches of customers with general distribution $B_0(t)$ and mean value $\bar{\gamma}_{01}$. Denote by $N(t)$ number of batches which arrive in interval $[0, t)$ and by $J(t)$ the PP of MAP with generator matrix $D = D_0 + D_1$. $P(n, t), n \geq 0$ are $m \times m$ size matrices with elements $P_r(n, t, i) = P\{N^s(t) = n, J(t) = i | N^s(0) = 0, J(0) = r\}$, where $P_r(n, t, i)$ is a conditional probability of having n batches in service at time t and PP is in state i given that at initial moment model was empty and PP was in state. If $R(z, t)$ be the PGF of $\{P(n, t), n \geq 0\}$, then it satisfies the following differential equation

$$\frac{\partial}{\partial t}P(z, t) = P(z, t) [D_0 + D_1(B_0(x) + z(1 - B_0(x)))], \quad |z| \leq 1, \tag{8}$$

with initial condition $P(z, 0) = I$.

The solution of Eq. (8) in matrix exponential form is:

$$P(z, t) = e^{\int_0^t [D_0 + D_1(B_0(x) + z(1 - B_0(x)))] dx}.$$

Remark. Let consider the service policy when we do not distinguish the customers of the same batch, i.e. all they have the same serve time. The arrival process is a MMAP with characteristic matrices $D_0, D(n) = \sum_{h_1+h_2+\dots+h_K=n} D_h$, where matrix $D(n)$ correspond to arrival of the batch with customers. In this case MMAP transforms to ordinary BMAP [36,37] with characteristic matrices $\{D_0, D(n), n > 0\}$. The service time of customers has general distribution $B(t)$ and mean value $\bar{\gamma}_1$. Thus for this model PGF we get

$$P(z, t) = e^{\int_0^t \left[D_0 + \sum_{n=1}^{\infty} D(n)(B(x) + z(1 - B(x)))^n \right] dx} = e^{\int_0^t [D_0 + D(B(x) + z(1 - B(x)))] dx},$$

where $D(z) = \sum_{n=1}^{\infty} D(n)(B(x) + z(1 - B(x)))^n$ is a rate GM of BMAP.

6 $MMAP_k|G_k|_\infty$ Model with Catastrophes

Let consider the general homogeneous Markovian model under influence of SMP generated catastrophes. As in [17], after every transition of environmental SMP the model jumps into the special state, let say 0-state, and then works from that state. When the SMP is in i state all parameters of the model are related to that state: DF of inter-arrival time of customers, DF and rates of service time of customers, their resource vectors. Let $\bar{P}(n, t, i)$ and $P(n, t, i)$ defined the probabilities of having in the model $n = (n_1, n_2, \dots, n_k)$ customers at moment t when environmental SMP is in state i , for the models without catastrophes and with catastrophes, resp. The following theorem gives the connection between these two models probabilities,

Theorem 5. The probabilities $P(n, t, i)$ satisfy the following integral equations

$$P(n, t, i) = (1 - F_i(t))\bar{P}(n, t, i) + \sum_{j \in S} \int_0^t P(n, t - u, j) dQ_{ij}(u), \quad i \in S.$$

The solution of the previous equations can be found

$$P(n, t, i) = (1 - F_i(t))\bar{P}(n, t, i) + \sum_{j \in S} \int_0^t (1 - F_i(t - u))\bar{P}(n, t - u, j) dH_{ij}(u), \quad i \in S,$$

where $F(t) = \{F_i(t), i \in S\}$ is a sojourn time distribution vector of SMP: $F_i(t) = \sum_{j \in S} Q_{ij}(t)$, $i \in S$. $H(t) = \|H_{ij}(t)\|$ is a renewal matrix of SMP which components satisfy the following equations

$$H_{ij}(t) = 1 - F_i(t) + \sum_{k \in S} \int_0^t H_{kj}(t - u) dQ_{ik}(u). \quad H_i(t) = \sum_{j \in S} H_{ij}(t), \quad i, j \in S.$$

The proof can be done by using standard renewal arguments (see for example [38]). Let $\tilde{P}(z, t, i)$ and $\tilde{\tilde{P}}(z, t, i)$ be the PGFs of $P(z, t, i)$ and $\bar{P}(z, t, i)$, respectively.

Theorem 6. The PGF satisfy the following integral equations

$$\tilde{P}(z, t, i) = (1 - F_i(t))\tilde{\tilde{P}}(z, t, i) + \sum_{j \in S} \int_0^t \tilde{P}(z, t - u, j) dQ_{ij}(u), \quad i \in S,$$

which solution is

$$P(z, t, i) = (1 - F_i(t))\tilde{\tilde{P}}(z, t, i) + \sum_{j \in S} \int_0^t (1 - F_j(t - u))\tilde{\tilde{P}}(z, t - u, j) dH_{ij}(u), \quad i \in S.$$

The proof can be done by using standard renewal arguments or by CMM (e.g. see [38]).

Theorem 7. The limiting distributions of $P(n)$ and $\tilde{P}(z)$ are given by

$$\begin{aligned}
 \hat{P}(n) &= \lim_{t \rightarrow \infty} P(n, t, i) = \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \int_0^\infty (1 - F_j(u)) \bar{P}(n, u, j) du, \quad i \in S, \\
 \tilde{P}(z) &= \lim_{t \rightarrow \infty} \tilde{P}(z, t, i) = \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \int_0^\infty (1 - F_j(u)) \tilde{\bar{P}}(z, u, j) du, \quad i \in S, \quad (9)
 \end{aligned}$$

where $\bar{\eta}_i = \int_0^\infty (1 - F_i(u)) du$, $q_i = \frac{\bar{\eta}_i \rho_i}{\sum_{r \in S} \bar{\eta}_r \rho_r}$, $\sum_{r \in S} q_r = 1$, $\rho_i = \sum_{r \in S} p_{ri} \rho_r$, $\sum_{r \in S} \rho_r = 1$, $p_{ri} = Q_{ri}(\infty)$, $r, i \in S$.

Let $\hat{f}(s)$ denote the Laplace Transformation of a function $f(x)$, $\hat{f}(s) = \int_0^\infty e^{-su} f(u) du$. The transient and stationary moments $L_r(t)$ and L_r of $P(n, t, i)$ can be found as

$$\begin{aligned}
 L_{1r}(t) &= \left. \frac{\partial \tilde{P}(z, t, i)}{\partial z_r} \right|_{z_1 = \dots = z_k = 1}, \quad L_{2r}(t) = \left. \frac{\partial^2 \tilde{P}(z, t, i)}{\partial z_r^2} \right|_{z_1 = \dots = z_k = 1}, \\
 L_{1r} &= \lim_{t \rightarrow \infty} L_{1r}(t), \quad L_{2r} = \lim_{t \rightarrow \infty} L_{2r}(t).
 \end{aligned}$$

Corollary. When $F_i(t) = 1 - e^{-v_i}$, $i \in S$ then for LT of $P(n, t, i)$, PGF $\tilde{P}(z, t, i)$ and their limiting values we get

$$\begin{aligned}
 \hat{P}(n, s, i) &= \hat{\tilde{P}}(n, s + v_i, i) + \frac{1}{s} \sum_{j \in S} \hat{\tilde{P}}(n, s + v_j, j) \hat{H}_{ij}(s), \quad i \in S, \\
 \hat{\tilde{P}}(z, s, i) &= \hat{\tilde{\tilde{P}}}(z, s + v_i, i) + \frac{1}{s} \sum_{j \in S} \hat{\tilde{\tilde{P}}}(z, s + v_j, j) \hat{H}_{ij}(s), \quad i \in S, \\
 \tilde{P}(z) &= \sum_{i \in S} q_i v_i \hat{\tilde{P}}(z, v_i, i), \quad P(n) = \sum_{i \in S} q_i v_i \hat{\tilde{P}}(n, v_i, i).
 \end{aligned}$$

Let consider the infinite-server models $MMA P_k|G_k|\infty$ with catastrophes. In this model the environmental SMP can be considered as a catastrophes process. After every transition of this SMP all customers in the model are flushed out instantly, the model jumps into empty state, and then continues its work from this state. If SMP is in state i at moment t all parameters of the model arrival process and service time distributions of customers, are function of i . Let $L(n, x, t, i)$ is a matrix, $L(n, x, t, i) = \| \| L_{jr}(n, x, t, i) \| \|$, where $L_{jr}(n, x, t, i)$ is the probability of event “in the model there are $n = (n_1, n_2, \dots, n_K)$ customers at moment t , PP $J(t)$ is in the phase j , total accumulated resource is x and SMP $\xi(t)$ is in state i under condition that at initial moment $t = 0$ the model was empty, and PP $J(0)$ was in phase $r \in S$,

$$L_{jr}(n, x, t, i) = P(N_i^s(t) = n, \alpha(t) \leq x, J(t) = j, \xi(t) = i | N_i^s(0) = 0, J(0) = r).$$

Let $L(n, s, t, i)$ be the LST of $L(n, x, t, i)$.

Theorem 8. The probabilities $L(n, s, t, i)$ be the LST of $L(n, s, t)$ satisfy the following renewal equations

$$L(n, s, t, i) = (1 - F_i(t))\bar{P}(n, s, t, i) + \sum_{j \in S} \int_0^t L(n, s, t - u, j) dQ_{ij}(u),$$

which solutions are given by

$$L(n, s, t, i) = (1 - F_i(t))\bar{P}(n, s, t, i) + \sum_{j \in S} \int_0^t (1 - F_j(t - u))\bar{P}(n, s, t - u, j) dH_{ij}(u),$$

$$L(n, s, t) = \sum_{i \in S} p_i^0 (1 - F_i(t))\bar{P}(n, s, t, i) + \sum_{j \in S} \int_0^t (1 - F_j(t - u))\bar{P}(n, s, t - u, j) dH_j(u),$$

where $H_j(t) = \sum_{k \in S} p_k^0 H_{kj}(t)$. The theorem can be proved by using properties of renewal arguments [38].

If SMP $\xi(t)$, $t \geq 0$ is an irreducible, ergodic process then for $\tilde{L}(n, s, t)$ when t is tending toward $+\infty$ by means key renewal theorem from Eq. (9) we derive following asymptotic result

$$L(n, s) = \lim_{t \rightarrow \infty} L(n, s, t) = \sum_{i \in S} \frac{q_i}{\bar{\eta}_i} \int_0^\infty (1 - F_i(u))\bar{P}(n, s, u, i) du. \tag{10}$$

Let $\tilde{L}(z, s, t)$ be PGF of $L(n, s, t)$, $\tilde{L}(z, s, t) = \sum_{n=0}^\infty z^n L(n, s, t)$, and $\tilde{L}(z, s) = \lim_{t \rightarrow \infty} \tilde{L}(z, s, t)$. Thus for $\tilde{L}(z, s, t)$ and $\tilde{L}(z, s)$ we derive

$$\tilde{L}(z, s, t) = \sum_{i \in S} p_i^0 (1 - F_i(t))\tilde{\bar{P}}(z, s, t, i) + \sum_{j \in S} \int_0^t (1 - F_j(t - u))\tilde{\bar{P}}(z, s, t - u, j) dH_j(u),$$

$$\tilde{L}(z, s) = \sum_{i \in S} \frac{q_i}{\bar{\eta}_i} \int_0^\infty (1 - F_i(u))\tilde{\bar{P}}(z, s, u, i) du.$$

By substitution the expression for $\tilde{\bar{P}}(z, s, t, i)$ into Eq. (10) the expression for PGF of joint distribution of number of busy servers and total accumulated resources in the model at moment t given that at $t = 0$ SMP was in state i for the model with catastrophes and its limiting value we get

Theorem 9. The PGFs $\tilde{L}(z, s, t, i)$, $\tilde{L}(z, s, t)$ and its limiting value $\tilde{L}(z, s)$ are given by

$$\tilde{L}(z, s, t, i) = e^{\int_0^t [D_0(i) + \tilde{S}_i(z, s, u)] du} (1 - F_i(t))$$

$$+ \sum_{r \in S} \int_0^t e^{-\int_0^{t-u} [D_0(r) + \tilde{S}_r(z, s, x)] dx} dQ_{ir}(u), \quad i \in S,$$

which solutions are

$$\begin{aligned} \tilde{L}(z, s, t, i) &= e^{\int_0^t [D_0(i) + \tilde{S}_i(z, s, u)] du} (1 - F_i(t)) \\ &+ \sum_{r \in S} \int_0^t (1 - F_r(t - u)) e^{\int_0^{t-u} [D_0(r) + \tilde{S}_r(z, s, x)] dx} dH_{ir}(u), \quad i \in S, \\ \tilde{L}(z, s, t) &= \sum_{i \in S} p_i^0 e^{\int_0^t [D_0(i) + \tilde{S}_i(z, s, u)] du} (1 - F_i(t)) \\ &+ \sum_{r \in S} \int_0^t (1 - F_r(t - u)) e^{\int_0^{t-u} [D_0(r) + \tilde{S}_r(z, s, x)] dx} dH_r(u), \\ \tilde{L}(z, s) &= \lim_{t \rightarrow \infty} \tilde{L}(z, s, t, r) = \sum_{i \in S} \frac{q_i}{\tilde{\eta}_i} \int_0^\infty (1 - F_i(x)) e^{\int_0^x [D_0(i) + \tilde{S}_i(z, s, u)] du} dx, \quad r \in S. \end{aligned}$$

Corollary. When $F_i(t) = 1 - e^{-v_i t}$, $i \in S$ then for LT of PGF $\tilde{L}(z, s, t, i)$ and their limiting values we get

$$\begin{aligned} \tilde{L}(z, s, s, i) &= \tilde{P}(z, s, s + v_i, i) + \frac{1}{s} \sum_{r \in S} \tilde{P}(z, s, s + v_r, r) \hat{H}_{ir}(s), \quad i \in S, \\ \tilde{L}(z, s) &= \sum_{r \in S} \tilde{P}(z, s, v_r, r) q_r v_r. \end{aligned} \tag{11}$$

If SMP is a simple renewal process with exponential distributed renewal time $F(t) = 1 - e^{-vt}$ then by Eq. (11) we get the results for homogeneous model [20]. For example, for $\tilde{L}(z, s)$ and $\tilde{L}(z, s, s)$ we obtain

$$\hat{\tilde{L}}(z, s, s) = \left(1 + \frac{v}{s}\right) \hat{\tilde{P}}(z, s, s + v), \quad \tilde{L}(z, s) = v \hat{\tilde{P}}(z, s, v).$$

7 Performance Evaluation of the Model

Let $\omega_{1r}(t)$, ω_{1r} denote the transient and steady state mean of queue length of type r customers. Then for $\omega_{1r}(t)$ and ω_{1r} we get

$$\omega_{1r}(t) = \lim_{s_1 \rightarrow 0} \omega_{1r}(s_1, t), \quad \omega_{1r}(s_1, t) = \left. \frac{\partial \tilde{L}(z_1, s_1, t)}{\partial z_{1r}} \right|_{z_{1r}=1, z_{11}=z_{12}=\dots=z_{1k}=1},$$

$$\omega_{1r} = \lim_{s_1 \rightarrow 0} \bar{\omega}_{1r}(s_1), \quad \bar{\omega}_{1r}(s_1) = \left. \frac{\partial \tilde{P}(z_1, s_1)}{\partial z_{1r}} \right|_{z_{1r}=1, z_{11}=z_{12}=\dots=z_{1k}=1},$$

$$\omega_{1r}(s_1, t) = \sum_{i \in S} p_i^0 (1 - F_i(t)) \bar{\omega}_{1r}(s_1, t, i)$$

$$+ \sum_{j \in S} \int_0^t (1 - F_j(t-u)) \bar{\omega}_{1r}(s_1, t-u, j) dH_j(u),$$

$$\omega_{1r}(s_1) = \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \int_0^\infty (1 - F_j(u)) \bar{\omega}_{1r}(s_1, u, j) du.$$

Where $\bar{\omega}_{1r}(s_1, t, i)$ is a transient mean queue length of r type customers of the model without catastrophes when the SMP is in state i .

Let $\delta_r(t)$, δ_r , $r = 1, 2, \dots, K$, be the transient and steady-state mean values of accumulated type r resources in the model and δ be a total accumulated resources in the model.

$$\delta_r(t) = \pi \delta_r(t) e, \quad \delta_r(t) = \lim_{s_1 \rightarrow 0} \left. \frac{\partial \tilde{L}(z_1, s_1, t)}{\partial s_{1r}} \right|_{z_{11}=z_{12}=\dots=z_{1k}=1},$$

$$\delta_r(t) = \sum_{i \in S} p_i^0 (1 - F_i(t)) \bar{\delta}_r(t, i) + \sum_{j \in S} \int_0^t (1 - F_j(t-u)) \bar{\delta}_r(t-u, j) dH_j(u),$$

$$\delta_r = \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \lambda_{jr} \bar{c}_{1r}(j) \int_0^\infty \int_0^u (1 - F_j(u))(1 - B_{jr}(x)) dx du,$$

where $\delta_r = \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \lambda_{jr} \bar{c}_{1r}(j) \int_0^\infty \int_0^u (1 - F_j(u))(1 - B_{jr}(x)) dx du$, $\bar{c}_{1r}(j)$ is a mean value of DF $C_{jr}(x)$.

$$\delta = \sum_{r=1}^K \delta_r = \sum_{r=1}^K \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \lambda_{jr} \bar{c}_{1r}(j) \int_0^\infty \int_0^u (1 - F_j(u))(1 - B_{jr}(x)) dx du.$$

If L_{losr} denote the steady state mean number of destroyed type r customers, then

$$L_{losr} = \lim_{s_1 \rightarrow 0} \pi L_{losr}(s_1) e,$$

where $L_{losr}(s_1) = \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \int_0^\infty \bar{\omega}_{1r}(s_1, u, j) dF_j(u)$.

$$L_{losr} = \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \lambda_{jr} \int_0^\infty \int_0^u (1 - B_{jr}(x)) dx dF_j(u).$$

If L_{los} is the steady-state total mean number of destroyed customers of all types, then

$$L_{los} = \sum_{r=1}^K L_{losr} = \sum_{r=1}^K \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \lambda_{jr} \int_0^\infty \int_0^u (1 - B_{jr}(x)) dx dF_j(u).$$

Let L_{qr} and L_q be the steady state mean number of type r and all types customers in the model. Then

$$\begin{aligned} L_{qr} &= \pi \tilde{\omega}_{1r} e = \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \int_0^\infty (1 - F_j(u)) \pi \tilde{\omega}_{1r}(u, j) e du \\ &= \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \lambda_{jr} \int_0^\infty \int_0^u (1 - F_j(u))(1 - B_{jr}(x)) dx du, \\ L_q &= \sum_{r=1}^K L_{qr} = \sum_{r=1}^K \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \lambda_{jr} \int_0^\infty \int_0^u (1 - F_j(u))(1 - B_{jr}(x)) dx du. \end{aligned}$$

Suppose that MMAP is defined by following matrices $D_0(i) = -\alpha_i I$, $D_h(i) = \alpha_i p(h_1, \dots, h_K) I$; $i \in S$, $h \in C^0$, where I is an identity matrix. Then for $\tilde{P}(n, t, i)$, $L(n, t)$, $L(n)$, and L_{losr} we obtain

$$\begin{aligned} \tilde{D}_i(z, u, t) &= \sum_{n=0}^\infty p(n_1, n_2, \dots, n_K) \prod_{r=1}^K [z_{ri}(1 - G_{ri}(t - u)) + G_{ri}(t - u)]^{n_r}, \\ \tilde{P}(z, t, i) &= e^{-\alpha_i \int_0^t \{1 - \tilde{D}_i(z, u, t)\} du}, \\ \tilde{L}(z, t) &= \sum_{j \in S} p_j^0 (1 - F_j(t)) e^{-\alpha_j \int_0^t \{1 - \tilde{D}_j(z, u, t)\} du} \\ &+ \sum_{j \in S} \int_0^t (1 - F_j(t - u)) e^{-\alpha_j \int_0^{t-u} \{1 - \tilde{D}_j(z, u, t-u)\} du} dH_j(u), \\ \tilde{L}(z) &= \sum_{i \in S} \frac{q_i}{\bar{\eta}_i} \int_0^\infty (1 - F_i(u)) e^{-\alpha_i \int_0^u \{1 - \tilde{D}_i(z, u, t)\} du} du, \\ L_{los} &= \sum_{r=1}^K \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \lambda_{jr} \int_0^\infty \int_0^u (1 - B_{jr}(x)) dx dF_j(u), \end{aligned} \tag{12}$$

where $\lambda_{ir} = \alpha_i \sum_{n_1=0}^\infty \dots \sum_{n_K=0}^\infty n_r p(n_1, n_2, \dots, n_K)$.

Equation (12) is a known result for $Mr|Gr|\infty$ model (see for example [13] formula (4)).

8 Conclusion

We consider infinite-server $MMAP_k|G_k|\infty$ queue in SM random environment, marked MAP arrival of customers, random resource vector of each type of customers and catastrophes. The joint PGF of transient and stationary distributions of number of busy servers and numbers of served customers, as well as the LT of joint distributions of total accumulated resource in the model and total accumulated resource vectors of served customers during time interval are found. The renewal equations for transient and stationary PGF of queue sizes and resource vectors of different type of customers are found for $MMAP_k|G_k|\infty$ and $BG_k|G_k|\infty$ queue models. For homogeneous Markov model in SM random environment and catastrophes the transient and limiting distributions renewal equations and their solutions are found. All results are obtained using CMM and renewal process methods. The obtained results may be applied for computer system and network performance metrics evaluation, as well as for design of optimal strategies of resource management of a wide class of subsystems of New Generation Networks, whereas the queue $MMAP_k|G_k|\infty$ may be used as a model of these subsystems.

References

1. Erlang, A.K.: The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik* **20**, 33–41 (1909)
2. Schwartz, M.: *Telecommunication Networks: Protocols, Modeling and Analysis*. Addison-Wesley, Boston (1987)
3. Paxson, V., Floyd, S.: Wide-area traffic: The failure of Poisson modeling. In: *Proceedings of the ACM*, pp. 257–268 (1994)
4. Artalejo, J.R., Gomez-Corral, A., He, Q.M.: Markovian arrivals in stochastic modelling: a survey and some new results. *SORT* **34**(2), 101–144 (2010)
5. He, Q.: *Fundamentals of Matrix-Analytic Methods*. Springer Science and Business Media, NY (2014)
6. He, Q., Neuts, M.: Markov chains with marked transitions. *Stoch Process. Appl.* **74**, 37–52 (1998)
7. Breuer, L.: *From Markov Jump Processes to Spatial Queues*. Springer Science, NY (2003)
8. Baum, D., Kalashnikov, V.: Spatial nowaiting stations with moving customers queue. *IT* **46**, 231–247 (2004)
9. Ramaswami, V., Neuts, M.F.: Some explicit formulas and computational methods for infinite-server ueues with phase-type arrival. *J. Appl. Prob.* **17**, 498–514 (1980)
10. Blom, J., Mandjes, M., Thorsdottir, H.: Time-scaling limits for Markov-modulated infinite server queues. *Stoch. Models.* **29**, 112–137 (2013)
11. Choi, B.: The $M^k|M^k|\infty$ queue with heterogeneous customers in a batch. *J. Appl. Prob.* **29**, 477–481 (1992)
12. Falin, G.: The $M^k|G^k|\infty$ batch arrival queue with heterogeneous dependent demands. *J. Appl. Probab.* **31**(3), 841–846 (1994)
13. Tong, D.C.: On the $BM/G/\infty$ queue with heterogeneous customers in a batch. *J. Appl. Prob.* **31**(1), 280–286 (1994)

14. Masuyama, H.: Studies on algorithmic analysis of queues with batch Markovian arrival streams. Ph.D. thesis, Kyoto University (2003)
15. Moiseev, A., Nazarov, A.: Infinite-server queueing systems and networks. Publ. NTL, Tomsk (2015)
16. Nazarov, A., Baymeeva, G.: The $M|G|\infty$ Queue in random environment. In: Proceedings of 13th International Conference, ITMM, pp. 312–324 (2014)
17. D’Auria, B.: Stochastic Decomposition of the $M|G|$ -infinite Queue in a Random Environment. Eurandom, Eindhoven (2005). 2005045
18. Purdue, P., Linton, D.: An infinite-server queue subject to an extraneous phase process and related models. J. Appl. Prob. **18**, 236–244 (1981)
19. Linton, D., Purdue, P.: An $M|G|\infty$ queue with m customer types subject to periodic clearing. Opsearch **16**, 80–88 (1979)
20. Kerobyan, K., Enakoutsas, K., Kerobyan, R.: An infinite-server queueing model $M\overline{MAP}_k|G_k|\infty$ with semi-Markov random environment and subject to catastrophes. In: Proceedings of the 8th International Conference, CSIT (2018)
21. Kerobyan, Kh., Kerobyan, R.: Transient analysis of infinite-server queue $M\overline{MAP}_k(t)|G_k|\infty$ with marked MAP arrival and disasters. In: Proceedings of the 7th International Working Conference, HET-NETs 2013, Ilkley, UK, 11–13 November 2013 (2013)
22. Böhm, W.: A Note on queueing systems exposed to disasters. Research report series. Dep. of Stat. and Math, 79, WU Vienna Uni. of Econom. and Business, Vienna (2008)
23. Economou, A., Fakinos, D.: Alternative approaches for the transient analysis of Markov chains with catastrophes. J. Stat. Theory Practice **2**(2), 183–197 (2008)
24. Tikhonenko, O.M.: Distribution of the total message flow in group arrival queueing system. Avtomatika i Telemekhanika **11**, 111–120 (1987)
25. Tikhonenko, O.M.: Generalized erlang problem for service systems with finite total capacity. Probl. Inf. Transm. **41**(3), 243–253 (2005)
26. Tikhonenko, O.M.: Models for determination of total capacity in computer and communicating systems. Catholic University in Ružomberok Scientific Issues, Mathematica II, Ružomberok, pp. 73–80 (2008)
27. Lisovskaya, E., Moiseeva, S., Pagano, M., Potatueva, V.: Study of the $M\overline{MPP}|GI|\infty$ queueing system with random customers’ capacities. Informatika i ee prilozheniya **11**(4), 109–117 (2017)
28. Lisovskaya, E., Moiseeva, S.: Asymptotical analysis of a non-Markovian queueing system with renewal input process and random capacity of customers. Proc. TSU **39**, 30–38 (2017)
29. Naumov, V., Samuylov, K.: On the modeling of queue systems with multiple resources. Proc. RUDN **3**, 60–63 (2014)
30. Klimov, G.P.: Stochastic Service Systems. Nauka, Moscow (1966)
31. Riordan, J.: Stochastic Service Systems. Wiley, New York (1962)
32. Runnenberg, J.: On the use of Collective marks in queueing theory. Proc. Symp. Cong., pp. 339–348. UNC, Hill (1965)
33. Gnedenko, B.V., Danielyan, E.A.: Priority Service Systems. MSU, Moscow (1973)
34. Pacheco, A., Tang, L.C., Prabhu, N.U.: Markov-Modulated Processes and Semi-Regenerative Phenomena. World Scientific, New Jersey (2010)
35. Latouche, G., Ramaswami, V.: Introduction to Matrix Analytic Methods in Stochastic Modeling. SIAM (1999)
36. Neuts, M.: Structured Stochastic Matrices of $M|G|1$ Type and their Applications. Marcel Dekker, N.Y (1989)

37. Lucantoni, D.M.: The *BMAP/G/1* queue: a tutorial. In: Donatiello, L., Nelson, R. (eds.) *Performance/SIGMETRICS -1993*. LNCS, vol. 729, pp. 330–358. Springer, Heidelberg (1993). <https://doi.org/10.1007/BFb0013859>
38. Çinlar, E.: *Introduction to Stochastic Processes*. Prentice-Hall, NJ (1975)



Retrial Queueing Model with Two-Way Communication, Unreliable Server and Resume of Interrupted Call for Cognitive Radio Networks

Svetlana Paul¹(✉) and Tuan Phung-Duc²

¹ National Research Tomsk State University, Tomsk, Russia
paulsv82@mail.ru

² Faculty of Engineering, Information and Systems, University of Tsukuba,
Tsukuba, Japan
tuan@sk.tsukuba.ac.jp

Abstract. In this paper, we consider a single server queueing model $M/GI/GI/1/1$ with two types of calls: incoming calls and outgoing calls. Incoming call enters the system and goes into service if the server is free. If the server is busy, call instantly goes to orbit, after which the call retries to go into service. The server makes an outgoing call in its idle time. We will be reviewing a system with unreliable server. In a free state and while servicing outgoing calls the server is reliable and unable to crash. If while servicing incoming call the server crashes, the incoming call stays at the server and as soon as server recovers the call goes into afterservice. For that system we've obtained probability distribution of server states, condition for the existence of a stationary mode and probability distribution of a number of incoming calls in the system.

Keywords: Retrial queueing system · Incoming and outgoing calls
Unreliable server

1 Introduction

Retrial queueing systems are characterized by the fact that in case the server is busy a new call that comes into the system at that time is not lost, instead it goes to the orbit and tries to enter service again later. This scene appears in various communication systems with random access where several users are using the same communication channel together. These situations also appear in service systems such as call-centers where clients who can not connect to the operator call back later [1, 2]. In cellular communication systems return visit is also common and thus taking it into account is crucial in designing such systems [3, 4].

In service systems such as call-centers due to optimization operator's downtime should be minimized in order to increase productivity. This is the founding

reason for creating mixed call-centers where operator not only does receive calls from outside, but also makes outgoing calls in his downtime [5–8]. These situations are modeled by retrial queueing systems with outgoing calls in which server handles both incoming and outgoing calls.

This type of models was first proposed by Falin [9], who studied the model where the service intensities of both incoming and outgoing calls are the same. Thus in his work Falin obtained integral formulas for partial generating functions and some explicit expressions for characteristics of retrial queue $M/G/1/1$ with two-way communication. Artalejo and Resing [10] expanded Falin’s model for $M/G/1/K$ retrial queueing systems. Falin et al. [11] obtained first moments for the queue length of retrial queue system $M/G/1/1$ where service times of outgoing and incoming calls differ. Martin and Artalejo in their work [12] studied the queueing system $M/G/1/1$ with two-way communication where calls from orbit are going into service in order after an exponential delay.

Artalejo and Phung-Duc [7] are studying the queueing system $M/M/1/1$ with two-way communications and distinct service times of incoming and outgoing requests. In this work the authors obtained an explicit solution for a two-dimensional probability distribution of server states and a number of calls on the orbit. Additionally, the factorial moments were obtained, based on which the proposed numerical and recurrent algorithms may be applicable. Nazarov et al. [13] derive asymptotic results for the same model in [7] under slow retrial rate.

Along with the potential to make outgoing calls the situations could occur where operation of the server could be interrupted by a crash then the repair takes place for some time (recovery period).

Systems with unreliable servers are commonly the subjects of modern research [14]. Work [15] studied systems with calls going to orbit and repeated service. Results in [15] can be applied to work with multimedia applications. In addition, systems with crashing servers in one form or another are commonly occurring while analysing transport systems [16].

In this work we are reviewing retrial queueing system $M/GI/GI/1/1$ with two-way communication, unreliable server and afterservice of interrupted calls. For that system we’ve obtained probability distribution of server states, condition for the existence of a stationary mode and probability distribution of the number of calls in the system.

Our model reflects a real situation in cognitive radio networks where secondary users utilize the licensed channel of primary users when the primary user is not present in the system. Secondary users in cognitive networks correspond to incoming calls in our model. The service of incoming calls may be interrupted by primary calls. This feature is reflected in the breakdown mechanism where the breakdown event corresponds to the arrival of a primary user. The service time of primary user corresponds to the time to repair in our model. The unique feature of this paper is we provide a buffer for the interrupted secondary user so that its service is restarted upon the departure of the primary user.

The rest of our paper is organized as follows. In Sect. 2, we present the model description and definitions of parameters. Section 3 shows the set of Kolmogorov equation describing the dynamics of the system. Section 4 derives the probability

distribution of the state of the server while Sect. 5 shows the characteristic functions of the joint queue length distribution of the number of calls in the orbit and the state of the server. Section 6 demonstrates some numerical examples and concluding remarks are presented in Sect. 7.

2 Model Description and Problem Definition

We consider a single server queueing model with two types of calls: incoming calls and outgoing calls. Incoming calls arrive at the system according to a Poisson process with rate λ .

Incoming call enters the system and goes into service if the server is free, server then starts service for a time duration, distributed with a function $B_1(x)$. If at the moment of entering system the server is busy, call instantly goes to orbit and stays there for a exponentially distributed duration of time with a rate σ , after which the call retries to go into service.

If the server is idle (empty) it starts making outgoing calls to the outside (not from the orbit) with rate α , service time of which has distribution function $B_2(x)$.

We will be reviewing a system with unreliable server [17], which crashes with intensity γ and recovers with intensity μ while servicing incoming calls. In a free state and while servicing outgoing calls the server is reliable and unable to crash.

If while servicing incoming call the server crashes, the incoming call stays at the server and as soon as server recovers the call goes into afterservice. When the server is servicing an incoming call or the server is recovering, incoming calls are going to the orbit.

Let's denote process $i(t)$ as a number of incoming calls in the system at the moment of time t .

Research tasks in this work:

1. Find condition for the existence of a stationary mode in a reviewed retrial queue.
2. Find characteristic function and stationary probability distribution

$$P(i) = P \{i(t) = i\}. \tag{1}$$

3 Kolmogorov System of Equations

Let's denote:

the server states at the moment of time t as $k(t)$: 0 if the server is free, 1 if the server is busy serving an incoming call, 2 if the server is busy serving an outgoing call, 3 if the server is in a state of recovery;

$z(t)$ - remaining time of service, when $k = 1, 2, 3$.

Let's also denote probabilities

$$P \{k(t) = k, i(t) = i, z(t) < z\} = P_k(i, z, t), \quad k = 1, 2, 3,$$

$$P \{k(t) = k, i(t) = i\} = P_k(i, t), \quad k = 0. \tag{2}$$

Since the random process $\{k(t), i(t), z(t)\}, k = 1, 2, 3; \{k(t), i(t)\}, k = 0$ with a variable number of components is a Markov process, then we have to compose a system of Kolmogorov equations for the probability distribution (2).

We denote $P_k(i, \infty, t) = P_k(i, t), k = 1, 2$. When $k = 3$ $z(t)$ is remaining time of servicing a call, that is waiting for the recovery of the server to complete its service.

System of Kolmogorov equations for the probability distribution

$$\begin{aligned} & \{P_0(i, t), P_1(i, z, t), P_2(i, z, t), P_3(i, z, t)\} : \\ & -(\lambda + \alpha + i\sigma)P_0(i, t) + \frac{\partial P_1(i + 1, 0, t)}{\partial z} + \frac{\partial P_2(i, 0, t)}{\partial z} = \frac{\partial P_0(i, t)}{\partial t}, \\ & \frac{\partial P_1(i, z, t)}{\partial z} - \frac{\partial P_1(i, 0, t)}{\partial z} - (\lambda + \gamma)P_1(i, z, t) + \lambda P_1(i - 1, z, t) \\ & + \lambda B_1(z)P_0(i - 1, t) + i\sigma B_1(z)P_0(i, t) + \mu P_3(i, z, t) = \frac{\partial P_1(i, z, t)}{\partial t}, \\ & \frac{\partial P_2(i, z, t)}{\partial z} - \frac{\partial P_2(i, 0, t)}{\partial z} - \\ & \lambda P_2(i, z, t) + \lambda P_2(i - 1, z, t) + \alpha B_2(z)P_0(i, t) = \frac{\partial P_2(i, z, t)}{\partial t}, \\ & -(\lambda + \mu)P_3(i, z, t) + \lambda P_3(i - 1, z, t) + \gamma P_1(i, z, t) = \frac{\partial P_3(i, z, t)}{\partial t}. \end{aligned} \tag{3}$$

Let's write down the last system in stationary mode:

$$\begin{aligned} & -(\lambda + \alpha + i\sigma)P_0(i) + \frac{\partial P_1(i + 1, 0)}{\partial z} + \frac{\partial P_2(i, 0)}{\partial z} = 0, \\ & \frac{\partial P_1(i, z)}{\partial z} - \frac{\partial P_1(i, 0)}{\partial z} - (\lambda + \gamma)P_1(i, z) + \lambda P_1(i - 1, z) \\ & + \lambda B_1(z)P_0(i - 1) + i\sigma B_1(z)P_0(i) + \mu P_3(i, z) = 0, \\ & \frac{\partial P_2(i, z)}{\partial z} - \frac{\partial P_2(i, 0)}{\partial z} - \lambda P_2(i, z) + \lambda P_2(i - 1, z) + \alpha B_2(z)P_0(i) = 0, \\ & -(\lambda + \mu)P_3(i, z) + \lambda P_3(i - 1, z) + \gamma P_1(i, z) = 0. \end{aligned} \tag{4}$$

Let's introduce partial characteristic functions by denoting $j = \sqrt{-1}$:

$$\begin{aligned} H_0(u) &= \sum_{i=0}^{\infty} e^{ju_i} P_0(i), \\ H_k(u, z) &= \sum_{i=1}^{\infty} e^{ju_i} P_k(i, z), \quad k = 1, 2, 3. \end{aligned} \tag{5}$$

Rewriting system (4) in the following form:

$$\begin{aligned}
 & -(\lambda + \alpha)H_0(u) + j\sigma H'_0(u) + e^{-ju} \frac{\partial H_1(u, 0)}{\partial z} + \frac{\partial H_2(u, 0)}{\partial z} = 0, \\
 & \frac{\partial H_1(u, z)}{\partial z} - \frac{\partial H_1(u, 0)}{\partial z} + \mu H_3(u, z) \\
 & + (\lambda(e^{ju} - 1) - \gamma)H_1(u, z) + \lambda B_1(z)e^{ju}H_0(u) - j\sigma B_1(z)H'_0(u) = 0, \\
 & \frac{\partial H_2(u, z)}{\partial z} - \frac{\partial H_2(u, 0)}{\partial z} + \lambda(e^{ju} - 1)H_2(u, z) + \alpha B_2(z)H_0(u) = 0, \\
 & (\lambda(e^{ju} - 1) - \mu)H_3(u, z) + \gamma H_1(u, z) = 0. \tag{6}
 \end{aligned}$$

In the system (6) we'll do a limit as $z \rightarrow \infty$. Denoting

$$H_k(u, \infty) = H_k(u), k = 1, 2,$$

and after summing the resulting equations we will have a following equation

$$\frac{\partial H_1(u, 0)}{\partial z} - \lambda e^{ju} H(u) = 0, \tag{7}$$

where

$$H(u) = H_0(u) + H_1(u) + H_2(u) + H_3(u).$$

In Eq. (7), let us exclude the first equation in the system (6). Then the system of equations for partial characteristic functions (6) could be rewritten in the following form

$$\begin{aligned}
 & \frac{\partial H_1(u, z)}{\partial z} - \frac{\partial H_1(u, 0)}{\partial z} + \mu H_3(u, z) \\
 & + (\lambda(e^{ju} - 1) - \gamma)H_1(u, z) + \lambda B_1(z)e^{ju}H_0(u) - j\sigma B_1(z)H'_0(u) = 0, \\
 & \frac{\partial H_2(u, z)}{\partial z} - \frac{\partial H_2(u, 0)}{\partial z} + \lambda(e^{ju} - 1)H_2(u, z) + \alpha B_2(z)H_0(u) = 0, \\
 & (\lambda(e^{ju} - 1) - \mu)H_3(u, z) + \gamma H_1(u) = 0, \\
 & \frac{\partial H_1(u, 0)}{\partial z} - \lambda e^{ju} H(u) = 0. \tag{8}
 \end{aligned}$$

This system will be the main in further research.

4 Probabilities Distribution of the Server States and Condition of Existence of a Stationary Mode

Let's prove the following assertion.

Theorem 1. *For the considered retrial queue with aftersevice, denoting*

$$b_2 = \int_0^\infty x dB_2(x)$$

the probabilities $r_k = P\{k(t) = k\}$ of the server states have the form

$$\begin{aligned}
 r_0 &= \frac{1}{1 + \alpha b_2} \left(1 - \lambda b_1 \frac{\mu + \gamma}{\mu} \right), \\
 r_1 &= \lambda b_1, \quad r_2 = \alpha b_2 r_0, \quad r_3 = \frac{\gamma}{\mu} r_1.
 \end{aligned}
 \tag{9}$$

Proof. Let's denote

$$\begin{aligned}
 H_k(0, z) &= r_k(z), \quad k = 1, 2, 3; \quad H_0(0) = r_0, \\
 \left. \frac{\partial H_k(u, 0)}{\partial z} \right|_{u=0} &= r'_k(0), \quad k = 1, 2, \quad H'_0(u)|_{u=0} = jm_0,
 \end{aligned}
 \tag{10}$$

then substituting $u = 0$ to (8), we get the following system of equations:

$$\begin{aligned}
 r'_1(z) - r'_1(0) - \gamma r_1(z) + \mu r_3(z) + B_1(z)(\lambda r_0 + \sigma m_0) &= 0, \\
 r'_2(z) - r'_2(0) + \alpha B_2(z)r_0 &= 0, \\
 -\mu r_3(z) + \gamma r_1(z) &= 0, \\
 \lambda - r'_1(0) &= 0.
 \end{aligned}
 \tag{11}$$

By summing the first and the third equations of the system (11) we'll get

$$\begin{aligned}
 r'_1(z) &= r'_1(0) - B_1(z)(\lambda r_0 + \sigma m_0). \\
 r'_2(z) &= r'_2(0) - B_2(z)\alpha r_0.
 \end{aligned}
 \tag{12}$$

From the fourth equation we have

$$r'_1(0) = \lambda.$$

Then we'll get

$$\begin{aligned}
 r'_1(z) &= \lambda - B_1(z)(\lambda r_0 + \sigma m_0), \\
 r'_2(z) &= r'_2(0) - B_2(z)\alpha r_0.
 \end{aligned}
 \tag{13}$$

By tending $z \rightarrow \infty$ we have the following equations

$$\begin{aligned}
 \lambda &= \lambda r_0 + \sigma m_0, \\
 r'_2(0) &= \alpha r_0.
 \end{aligned}$$

Then we'll rewrite system (13) in the following form

$$\begin{aligned}
 r_1(z) &= \lambda \int_0^z (1 - B_1(x)) dx, \\
 r_2(z) &= \alpha r_0 \int_0^z (1 - B_2(x)) dx.
 \end{aligned}
 \tag{14}$$

Sending $z \rightarrow \infty$ we'll get the following expressions from the equations above and the third equation of a system (11)

$$\begin{aligned} r_1 &= \lambda b_1, \\ r_2 &= \alpha b_2 r_0, \\ r_3 &= \frac{\gamma}{\mu} r_1. \end{aligned} \tag{15}$$

We'll find the probability value r_0 from the normalization condition in the form of a first equation in (9). Theorem 1 is proved.

Corollary. The condition for the existence of the stationary mode in the reviewed retrial queue with afterservice is the following inequality

$$\lambda < \frac{\mu}{\mu + \gamma} \cdot \frac{1}{b_1}. \tag{16}$$

Proof. The condition (12) follows from the positivity of the probability r_0 in (9). The corollary is proved.

Let's define the system flow capacity S as a maximum average number of calls that could be serviced in a reviewed system per unit time. By inequality (16) the value S for a reviewed system with outgoing calls and unreliable server is defined by the equality

$$S = \frac{\mu}{\mu + \gamma} \cdot \frac{1}{b_1}. \tag{17}$$

If the value of a parameter λ of an incoming flow is defined by the equality $\lambda = \rho S$, then at any values of a parameter $0 < \rho < 1$ the stationary mode exists in the reviewed system, and the probabilities r_k from (9) server states could be written in the following form

$$\begin{aligned} r_0 &= \frac{1 - \rho}{1 + \alpha b_2}, \\ r_1 &= \rho \frac{\mu}{\mu + \gamma}, \\ r_2 &= \alpha b_2 \frac{1 - \rho}{1 + \alpha b_2}, \\ r_3 &= \rho \frac{\gamma}{\mu + \gamma}, \end{aligned} \tag{18}$$

which does not depend of the form of distribution functions $B_1(x)$ and $B_2(x)$ of a service time of both incoming and outgoing calls. When this happens the intensity λ of the incoming flow linearly depends on S , which by (17) does not depend on the form of distribution functions $B_1(x)$ and $B_2(x)$.

Further we find probability distribution $P(i)$ of the number $i(t)$ of calls in retrial queueing system with afterservice of interrupted calls.

5 Queue Length Distribution

Let's denote the Laplace-Stieltjes transform

$$B_k^*(s) = \int_0^\infty e^{-sx} dB_k(s),$$

$$H_k^*(u, s) = \int_0^\infty e^{-sz} dH_k(u, z), k = 1, 2, 3.$$

Let's rewrite system (8) in the following form

$$\begin{aligned} &-\frac{\partial H_1(u, 0)}{\partial z} \\ &+(\lambda(e^{ju} - 1) - \gamma + s)H_1^*(u, s) + \mu H_3^*(u, s) + B_1^*(s)(\lambda e^{ju} H_0(u) - j\sigma H_0'(u)) = 0, \\ &-\frac{\partial H_2(u, 0)}{\partial z} + (\lambda(e^{ju} - 1) + s)H_2^*(u, s) + \alpha B_2^*(s)H_0(u) = 0, \\ &(\lambda(e^{ju} - 1) - \mu)H_3^*(u, s) + \gamma H_1(u, s) = 0, \\ &\frac{\partial H_1(u, 0)}{\partial z} - \lambda e^{ju} H(u) = 0. \end{aligned} \tag{19}$$

This system will serve as a base in the further researches. Let's write down the third equation of the system (19) in the following form

$$H_3^*(u, s) = \frac{\gamma}{\mu - \lambda(e^{ju} - 1)} H_1^*(u, s),$$

and by substituting this expression into the first equation we'll get

$$\begin{aligned} &-\frac{\partial H_1(u, 0)}{\partial z} + B_1^*(s) (\lambda e^{ju} H_0(u) - j\sigma H_0'(u)) \\ &+ \left\{ \lambda(e^{ju} - 1) - \gamma + s + \mu \frac{\gamma}{\mu - \lambda(e^{ju} - 1)} \right\} H_1^*(u, s) = 0. \end{aligned} \tag{20}$$

Let's denote

$$\begin{aligned} g_1(u) &= \lambda(1 - e^{ju}) + \gamma - \mu \frac{\gamma}{\lambda(1 - e^{ju}) + \mu}, \\ g_2(u) &= \lambda(1 - e^{ju}), \quad g_3(u) = \lambda(1 - e^{ju}) + \mu, \end{aligned} \tag{21}$$

and rewrite system (19) in the following form

$$\begin{aligned} &-\frac{\partial H_1(u, 0)}{\partial z} + (s - g_1(u))H_1^*(u, s) + B_1^*(s)(\lambda e^{ju} H_0(u) - j\sigma H_0'(u)) = 0, \\ &-\frac{\partial H_2(u, 0)}{\partial z} + (s - g_2(u))H_2^*(u, s) + \alpha B_2^*(s)H_0(u) = 0, \end{aligned}$$

$$g_3(u)H_3^*(u, s) - \gamma H_1(u, s) = 0,$$

$$\frac{\partial H_1(u, 0)}{\partial z} - \lambda e^{ju} H(u) = 0. \quad (22)$$

Substituting $s = g_1(u)$ in the first equation and $s = g_2(u)$ in the second equation of the system (22) we'll get these equations

$$\frac{\partial H_1(u, 0)}{\partial z} = B_1^*(g_1(u))(\lambda e^{ju} H_0(u) - j\sigma H_0'(u)),$$

$$\frac{\partial H_2(u, 0)}{\partial z} = \alpha B_2^*(g_2(u))H_0(u). \quad (23)$$

Let's denote $H_k^*(u, 0) = H_k(u), k = 1, 2, 3$. Substituting $s = 0$ in (22), and considering equalities (23) let's rewrite system (22) in the following form

$$-g_1(u)H_1(u) + \{1 - B_1^*(g_1(u))\} \{ \lambda e^{ju} H_0(u) - j\sigma H_0'(u) \} = 0,$$

$$-g_2(u)H_2(u) + \alpha H_0(u) (1 - B_2^*(g_2(u))) = 0,$$

$$g_3(u)H_3(u) = \gamma H_1(u),$$

$$\frac{\partial H_1(u, 0)}{\partial z} - \lambda e^{ju} H(u) = 0. \quad (24)$$

We'll rewrite the first three equations of system (24) in the form:

$$H_1(u) = \frac{1 - B_1^*(g_1(u))}{g_1(u)} \{ \lambda e^{ju} H_0(u) - j\sigma H_0'(u) \},$$

$$H_2(u) = \alpha \frac{1 - B_2^*(g_2(u))}{g_2(u)} H_0(u),$$

$$H_3(u) = \frac{\gamma}{g_3(u)} H_1(u). \quad (25)$$

By substituting these expressions into the fourth equation of system (25) we'll get this equality

$$0 = \lambda e^{ju} H(u) - B_1^*(g_1(u)) \left(\lambda e^{ju} H_0(u) - j\sigma H_0'(u) \right)$$

$$= \lambda e^{ju} \left[H_0(u) \left(1 + \alpha \frac{1 - B_2^*(g_2(u))}{g_2(u)} \right) + H_1(u) \left(1 + \frac{\gamma}{g_3(u)} \right) \right]$$

$$- B_1^*(g_1(u)) \{ \lambda e^{ju} H_0(u) - j\sigma H_0'(u) \}$$

$$= \lambda e^{ju} \left[H_0(u) \left(1 + \alpha \frac{1 - B_2^*(g_2(u))}{g_2(u)} \right) \right]$$

$$+ \frac{1 - B_1^*(g_1(u))}{g_1(u)} \cdot \frac{g_3(u) + \gamma}{g_3(u)} \left(\lambda e^{ju} H_0(u) - j\sigma H_0'(u) \right) \Big]$$

$$- B_1^*(g_1(u)) \{ \lambda e^{ju} H_0(u) - j\sigma H_0'(u) \},$$

which we'll then rewrite in the form

$$\begin{aligned} & \lambda e^{ju} H_0(u) \left(1 + \alpha \frac{1 - B_2^*(g_2(u))}{g_2(u)} \right) \\ &= (\lambda e^{ju} H_0(u) - j\sigma H_0'(u)) \left(B_1^*(g_1(u)) - \lambda e^{ju} \frac{1 - B_1^*(g_1(u))}{g_1(u)} \cdot \frac{g_3(u) + \gamma}{g_3(u)} \right). \end{aligned} \tag{26}$$

Let's denote

$$\begin{aligned} f(u) &= \left(1 + \alpha \frac{1 - B_2^*(g_2(u))}{g_2(u)} \right) \\ &\times \left(B_1^*(g_1(u)) - \lambda e^{ju} \frac{1 - B_1^*(g_1(u))}{g_1(u)} \cdot \frac{g_3(u) + \gamma}{g_3(u)} \right)^{-1}, \end{aligned} \tag{27}$$

and rewrite equality (26) in the following form

$$\lambda e^{ju} H_0(u) f(u) = \lambda e^{ju} H_0(u) - j\sigma H_0'(u), \tag{28}$$

i.e. in the form of an ordinary differential equation

$$H_0'(u) = j \frac{\lambda}{\sigma} e^{ju} H_0(u) (f(u) - 1),$$

with respect to the function $H_0(u)$, satisfying the condition $H_0(0) = r_0$. Solution $H_0(u)$ of this equation will have the following form

$$H_0(u) = r_0 \exp \left\{ j \frac{\lambda}{\sigma} \int_0^u e^{jx} (f(x) - 1) dx \right\}. \tag{29}$$

By substituting last equation into (25) we'll write

$$\begin{aligned} H_1(u) &= \frac{1 - B_1^*(g_1(u))}{g_1(u)} \lambda e^{ju} H_0(u) f(u), \\ H_2(u) &= \alpha \frac{1 - B_2^*(g_2(u))}{g_2(u)} H_0(u), \\ H_3(u) &= \frac{\gamma}{g_3(u)} H_1(u). \end{aligned} \tag{30}$$

Thus the following statement is proved.

Theorem 2. Using $g_1(u)$ and $g_2(u)$ from (22) and also $f(u)$ from (27) then the characteristic function of a number $i(t)$ of calls in a reviewed retrial queueing system with afterservice of repeated calls has the following form

$$H(u) = M e^{ju} = H_0(u) + H_1(u) + H_2(u) + H_3(u),$$

in which the partial characteristic functions $H_k(u), k = \overline{0, 3}$ are defined by equalities (29), (30).

Stationary probabilities distribution $P(i) = P\{i(t) = i\}$ of a number of calls in a reviewed retrial queue is defined by the reverse Fourier transform and has the following form

$$P(i) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ju_i} H(u) du, \tag{31}$$

in which the expression for a characteristic function $H(u)$ is defined in the Theorem 2 formed above. The numerical realization of probabilities distribution $P(i)$ from (31) is effortless at any values of initial parameters $\alpha, \gamma, \mu, \sigma, \lambda$ and of distribution functions $B_1(x)$ and $B_2(x)$ that satisfy the condition (16) of existence of a stationary mode in a reviewed retrial queueing system.

6 Comparison of Two Types of Interruption

For retrial queueing system *M/GI/GI/1/1* with two-way communication, unreliable server and servicing of interrupted calls again, system flow capacity S_1 is defined by the equation

$$S_1 = \frac{\mu}{\mu + \gamma} \cdot \frac{\gamma B_1^*(\gamma)}{1 - B_1^*(\gamma)}.$$

In this system while servicing incoming call the server crashes, the incoming call goes into the orbit. Servicing of call again, at the repeated receipt of call on a server.

For retrial queueing system *M/GI/GI/1/1* with two-way communication, unreliable server and afterservice of interrupted calls system flow capacity S_2 defined by the equations for

$$S_2 = \frac{\mu}{\mu + \gamma} \cdot \frac{1}{b_1}.$$

Thus system flow capacities differ slightly at low γ values and even coincide at $\gamma \rightarrow 0$ and differ significantly at high γ values.

For parameters in the gamma distribution $\alpha_1 = \beta_1$,

$$B_1^*(\gamma) = \left(\frac{\beta_1}{\beta_1 + \gamma} \right)^{\alpha_1},$$

consider the ratio S_1/S_2 at the values of the parameter $\alpha_1 = 0.5; 1; 2$ and the values of the parameter γ (Table 1).

Table 1. A comparison of retrial queueing systems *M/GI/GI/1/1* with two-way communication, unreliable server and afterservice of interrupted calls and servicing interrupted calls again

γ	0.01	0.1	1	10	100
$\alpha_1 = 0.5$	1.005	1.048	1.366	2.791	7.589
$\alpha_1 = 1$	1.000	1.000	1.000	1.000	1000
$\alpha_1 = 2$	0.998	0.976	0.800	0.286	0.038

7 Conclusions

In this paper, we have considered retrial queue $M/GI/GI/1/1$ with two-way communication, unreliable server and afterservice of interrupted calls. We have found probability distribution of server states, the condition for the existence of a stationary mode and probability distribution of a number of calls in the system.

References

1. Artalejo, J.R., Gomez-Corral, A.: *Retrial Queueing Systems: A Computational Approach*. Springer, Heidelberg (2008). <https://doi.org/10.1007/978-3-540-78725-9>
2. Falin, G.I., Templeton, J.G.C.: *Retrial Queues*. Chapman and Hall, London (1997)
3. Bhulai, S., Koole, G.: A queueing model for call blending in call centers. *IEEE Trans. Autom. Control* **48**, 1434–1438 (2003)
4. Deslauriers, A., L'Ecuyer, P., Pichitlamken, J., Ingolfsson, A., Avramidis, A.N.: Markov chain models of a telephone call center with call blending. *Comput. Oper. Res.* **34**, 1616–1645 (2007)
5. Choi, B.D., Choi, K.B., Lee, Y.W.: M/G/1 retrial queueing systems with two types of calls and finite capacity. *Queueing Syst.* **19**, 215–229 (1995)
6. Tran-Gia, P., Mandjes, M.: Modeling of customer retrial phenomenon in cellular mobile networks. *IEEE J. Sel. Areas Commun.* **15**, 1406–1414 (1997)
7. Artalejo, J.R., Phung-Duc, T.: Markovian retrial queues with two way communication. *J. Ind. Manage. Optim.* **8**, 781–806 (2012)
8. Nazarov, A., Phung-Duc, T., Paul, S.: Heavy outgoing call asymptotics for $MMPP/M/1/1$ retrial queue with two-way communication. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) *ITMM 2017. CCIS*, vol. 800, pp. 28–41. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_3
9. Falin, G.I.: Model of coupled switching in presence of recurrent calls. *Eng. Cybern. Rev.* **17**, 53–59 (1979)
10. Artalejo, J.R., Resing, J.A.C.: Mean value analysis of single server retrial queues. *Asia Pac. J. Oper. Res.* **27**, 335–345 (2010)
11. Falin, G.I., Artalejo, J.R., Martin, M.: On the single server retrial queue with priority customers. *Queueing Syst.* **14**, 439–455 (1993)
12. Artalejo, J.R., Phung-Duc, T.: Single server retrial queues with two way communication. *Appl. Math. Model.* **37**(4), 1811–1822 (2003)
13. Nazarov, A., Paul, S., Gudkova, I.: Asymptotic analysis of Markovian retrial queue with two-way communication under low rate of retrials condition. In: *Proceedings - 31st European Conference on Modelling and Simulation, ECMS, Budapest*, pp. 687–693 (2017)
14. Djellab, N.V.: On the M/G/1 retrial queue subjected to breakdowns. *RAIRO Oper. Res.* **36**(4), 299–310 (2002)
15. Sherman, N., Kharoufeh, J., Abramson, M.: An M/G/1 retrial queue with unreliable server for streaming multimedia applications. *Probab. Eng. Inform. Sci.* **23**, 281–304 (2009)
16. Afanasyeva, L.G., Bulinskaya, E.V.: Some problems for flows of interacting particles. In: *Modern Problems of Mathematics and Mechanics*, pp. 55–67 (2009)
17. Samouylov, K., Naumov, V., Sopin, E., Gudkova, I., Shorgin, S.: Sojourn time analysis for processor sharing loss system with unreliable server. In: Wittevrongel, S., Phung-Duc, T. (eds.) *ASMTA 2016. LNCS*, vol. 9845, pp. 284–297. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43904-4_20



Mittag-Leffler Function in Applied Problems of Queuing Theory

Alexander Kirpichnikov¹, Anton Titovtsev^{1(✉)}, and Igor Yakimov²

¹ Kazan National Research Technological University,
K. Marksa str. 68, 420015 Kazan, Russia
kirpichnikov@kstu.ru, notna6683@mail.ru

² Kazan National Research Technical University named after A. N. Tupolev – KAI,
K. Marksa str. 10, 420111 Kazan, Russia
<http://www.kstu.ru>
<http://www.kai.ru>

Abstract. In the present work there has been made an attempt to give the complex description of queuing systems with limited queuing delay of claims, allowing to simplify the majority of the intermediate calculations. The mathematical basis is the introduction of Mittag-Leffler function review. Obtained on this basis, the unified, internally bound system of rather compact formulas allows to describe adequately all main characteristics of steady-state conditions of such multiserver QS. This is probability of system shutdown, load coefficient, queue mean length. And also this system of formulas allow to calculate time characteristics corresponding to these numerical characteristics.

Keywords: Queuing system (QS) · Queuing theory
Numerical characteristic of QS

1 Introduction

The queuing systems (QS) with limited mean queuing delay of claims are the most commonly encountered in applications queuing systems, therefore the calculation of their numerical characteristics is of great practical interest. It is known, however, that this calculation is often rather laborious, and the Markovian models describing steady-state conditions of such queuing systems, as a rule, are poorly adapted for the use in applications as they contain the sums of infinite series which do not reduce to geometrical progressions [1–3] as the final result.

In the present work there has been made an attempt to give the complex description of queuing systems with limited queuing delay of claims, allowing to simplify the majority of the intermediate calculations. The mathematical basis is the introduction of Mittag-Leffler function review

$$E_{\rho}(z, \mu) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma\left(\mu + \frac{k}{\rho}\right)}$$

(Γ - gamma function), well-known by experts in the field of complex variable theory and integral representations, but, as it appeared to be, insufficiently demanded by experts in the field of the applied theory of queuing system. Obtained on this basis, the unified, internally bound system of rather compact formulas allows to describe adequately all main characteristics of steady-state conditions of such multiserver QS - probability of system shutdown p_0 , load coefficient \bar{n} , queue mean length \bar{l} , and also to calculate time characteristics corresponding to these characteristics.

2 Multiserver QS with the Homogeneous Infinite Simple Stream of Claims and Queue of Unlimited Length

Let us assume that we have a multiserver QS with the homogeneous infinite simple stream of claims and queue of unlimited length. Let the intensity of claims stream be equal to λ , and the service rate, i.e. the average number of claims which are served by the device per unit time be μ . The service stream will also be considered as the simple one (with μ intensity).

Let us assume now that the number of places in queue is still not limited, but the staying time of one claim in the queue is limited by some random time t with the mean value \bar{t} . Thereby, each claim in the queue is impacted by some kind of a departure stream with the intensity of $\nu = 1/\bar{t}$.

It is clear, that if this stream has the simplest character, then the process proceeding in QS, will be the Markovian process. Let us find probabilities of steady-state conditions for it.

If there are n claims in the queue, then the total intensity of claims departure from the queue is, apparently, equal to $n\nu$, and then the state graph of the corresponding multi-server queuing system is as it is represented in Fig. 1 (death and reproduction process).

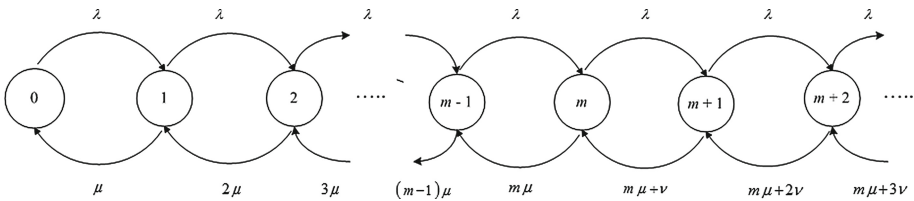


Fig. 1. State graph of queuing system with the limited mean response time of a claim in the queue

Applying general expressions [4] for the probabilities of limit (steady-state) conditions in this scheme, we will obtain

$$p_1 = \frac{\lambda}{\mu} p_0; p_2 = \frac{\lambda^2}{2\mu^2} p_0; p_3 = \frac{\lambda^3}{3!\mu^3} p_0;$$

$$\begin{aligned}
 p_m &= \frac{\lambda^m}{m! \mu^m} p_0; p_{m+1} = \frac{\lambda^{m+1}}{m! \mu^m (m \mu + \nu)} p_0; \\
 p_{m+2} &= \frac{\lambda^{m+2}}{m! \mu^m (m \mu + \nu) (m \mu + 2 \nu)} p_0; \\
 p_{m+3} &= \frac{\lambda^{m+3}}{m! \mu^m (m \mu + \nu) (m \mu + 2 \nu) (m \mu + 3 \nu)} p_0
 \end{aligned}$$

or in designations $\rho = \lambda/\mu$ (the presented intensity of claims stream, i.e. the mean number of claims entering the system during one claim mean service time in it) and $\beta = \nu/\mu$ (the mean number of claims leaving the queue without service during one claim mean service time in it)

$$\begin{aligned}
 p_1 &= \rho p_0; p_2 = \frac{\rho^2}{2} p_0; p_3 = \frac{\rho^3}{3!} p_0; \\
 p_m &= \frac{\rho^m}{m!} p_0; p_{m+1} = \frac{\rho^{m+1}}{m! (m + \beta)} p_0; \\
 p_{m+2} &= \frac{\rho^{m+2}}{m! (m + \beta) (m + 2 \beta)} p_0; \\
 p_{m+3} &= \frac{\rho^{m+3}}{m! (m + \beta) (m + 2 \beta) (m + 3 \beta)} p_0.
 \end{aligned}$$

As a result we have the following formulas for p_k :

$$\begin{aligned}
 p_k &= \frac{\rho^k}{k!} p_0 \text{ at } k \leq m; \\
 p_k &= \frac{\rho^k}{m! (m + \beta) (m + 2 \beta) \dots [m + (k - m) \beta]} p_0 \tag{1}
 \end{aligned}$$

The formulas record (1) for p_k can be simplified. Really, we will divide the numerator and the denominator of the second one of these ratios by β^{k-m} . Then we will obtain

$$p_k = \frac{\rho^k}{k!} p_0 \text{ at } k \leq m; p_k = \frac{\rho^m}{m!} \frac{\alpha^{k-m}}{(m/\beta + 1)_{k-m}} p_0 \text{ at } k \geq m,$$

where $(a)_k = a (a + 1) (a + 2) \dots (a + k - 1)$; $(a)_0 = 1$ – Pochhammer symbol [5]. $\alpha = \rho/\beta = \lambda/\nu$ value obviously shows what mean number of claims enters the system during the mean response time of one “impatient” claim staying in the queue. In this case, due to the normalization requirement $\sum_{k=0}^{\infty} p_k = 1$ we have

$$\begin{aligned}
 p_0 &= \left\{ e_{m-1}(\rho) + \frac{\rho^m}{m!} \left[1 + \frac{\alpha}{m/\beta + 1} \right. \right. \\
 &+ \left. \frac{\alpha^2}{(m/\beta + 1) (m/\beta + 2)} + \frac{\alpha^3}{(m/\beta + 1) (m/\beta + 2) (m/\beta + 3)} + \dots \right\}^{-1}
 \end{aligned}$$

$$\begin{aligned}
 &= \left\{ e_{m-1} \{ \rho \} + \frac{\rho^m}{m!} \left[1 + \frac{\alpha}{(m/\beta + 1)_1} + \frac{\alpha^2}{(m/\beta + 1)_2} + \dots \right] \right\}^{-1} \\
 &= \left[e_{m-1}(\rho) + \frac{\rho^m}{m!} \sum_{k=0}^{\infty} \frac{\alpha^k}{(m/\beta + 1)_k} \right]^{-1} \tag{2}
 \end{aligned}$$

$e_m(\rho) = 1 + \frac{\rho}{1!} + \frac{\rho^2}{2!} + \dots + \frac{\rho^m}{m!}$ - non-complete exponential function (non-complete exponential), and $e_0(\rho) = 1$, and at $m < 0$ we suppose $e_m(\rho) = 0$. It is clear that $e_m(\rho) \rightarrow e^\rho$ at $m \rightarrow \infty$.

Let us consider more closely the sum in the formula (2). The corresponding relations of the M/M/m (the multi-server unit according to Kendall’s classification) and M/M/m/K (model with queue of finite length) models, as we know, contain the sums of infinite or finite numbers of summands reduced to the sums of infinite or finite geometrical progressions respectively. In contrast to these classical cases, the formula (2) contains the sum of the infinite series which is not a progression of this kind. Therefore we will act in the following manner.

Observing that according to the definition

$$(a)_k = \frac{\Gamma(a + k)}{\Gamma(a)},$$

where Γ - gamma-function, we will rewrite the sum that interests us as

$$S = \Gamma(m/\beta + 1) \sum_{k=0}^{\infty} \frac{\alpha^k}{\Gamma(m/\beta + 1 + k)} \tag{3}$$

and then

$$p_0 = \left[e_{m-1}(\rho) + \frac{\rho^m}{m!} \Gamma(m/\beta + 1) E_1(\alpha; m/\beta + 1) \right]^{-1}$$

where

$$E_1(z; \xi) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\xi + k)} \tag{4}$$

- Mittag-Leffler function of the first order (synthesis of the exponential function $\exp z$). This function is well-known to experts in the field of complex variable theory and integral transformations [6,7]. Expression (4) in its turn can be simplified even more. From formula (4) we obviously have

$$\begin{aligned}
 E_1(z; \xi + 1) &= \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\xi + 1 + k)} = \sum_{k=1}^{\infty} \frac{z^{k-1}}{\Gamma(\xi + k)} = \frac{1}{z} \sum_{k=1}^{\infty} \frac{z^k}{\Gamma(\xi + k)} \\
 &= \frac{1}{z} \left[\sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\xi + k)} - \frac{1}{\Gamma(z)} \right],
 \end{aligned}$$

so that

$$E_1(z; \xi) = \frac{1}{\Gamma(\xi)} + z {}_1(z; \xi + 1) \tag{5}$$

– the recurrence formula for $E_1(z, \xi)$, and then from the relation (3) with regard to a well-known recurrent relation $\Gamma(\xi + 1) = \xi \Gamma(\xi)$ it follows

$$\begin{aligned} S &= \Gamma(m/\beta + 1) \frac{\beta}{\rho} \left[{}_1(\alpha; m/\beta) - \frac{1}{\Gamma(m/\beta)} \right] \\ &= \frac{m}{\rho} \Gamma(m/\beta) \left[E_1(\alpha; m/\beta) - \frac{1}{\Gamma(m/\beta)} \right] = \frac{m}{\rho} [\Gamma(m/\beta) {}_1(\alpha; m/\beta) - 1]. \end{aligned}$$

As a result, the relation (4) gives the following formula for p_0 :

$$p_0 = \left[e_{m-2}(\rho) + \frac{\rho^{m-1}}{(m-1)!} \Gamma(m/\beta) {}_1(\alpha; m/\beta) \right]^{-1} \tag{6}$$

(let us remind that $e_0(\rho) = 1$ and $e_m(\rho) = 0$ for all $m < 0$). For single-server QS ($m = 1$) the formula (2.5.7) has an especially simple view

$$p_0 = \frac{1}{\Gamma(1/\beta) {}_1(\alpha; 1/\beta)}.$$

In particular, at $\beta = 1$ (i.e. in the case when $\nu = \mu$) ${}_1(z; 1) = e^z$, and then $p_0 = e^{-\rho}$.

The limit $\beta \rightarrow 0$ corresponds to the case of an ordinary multi-server unit (M/M/m model), and in this case the formula (6) predictably passes to a known relation [4]

$$p_0 = \left[1 + \frac{\rho}{1!} + \frac{\rho^2}{2!} + \dots + \frac{\rho^{m-1}}{(m-1)!} + \frac{\rho^m}{(m-1)!(m-\rho)} \right]^{-1}$$

or

$$p_0 = \left[e_{m-1}(\rho) + \frac{\rho^m}{(m-1)!(m-\rho)} \right]^{-1} = \left[e_m(\rho) + \frac{\rho^{m+1}}{(m-1)!(m-\rho)} \right]^{-1}.$$

3 Numerical Characteristic of Steady-State Operating Conditions

The waiting probability, i.e. the probability that the arriving claim will find all servers engaged (no matter if it waits for the service or not)

$$p_w = \sum_{k=m}^{\infty} p_k = \frac{\rho^m p_0}{m!} \sum_{k=m}^{\infty} \frac{\alpha^{k-m}}{(m/\beta + 1)_{k-m}} = \frac{\rho^m p_0}{m!} \sum_{k=0}^{\infty} \frac{\alpha^k}{(m/\beta + 1)_k}$$

$$\begin{aligned}
 &= \frac{\rho^m p_0}{m!} \Gamma(m/\beta + 1) \sum_{k=0}^{\infty} \frac{\alpha^k}{\Gamma(m/\beta + 1 + k)} \\
 &= \frac{\rho^m p_0}{(m-1)! \beta} \Gamma(m/\beta) E_1(\alpha; m/\beta + 1) \\
 &= \frac{\rho^{m-1} p_0}{(m-1)!} \Gamma(m/\beta) \left[E_1(\alpha; m/\beta) - \frac{1}{\Gamma(m/\beta)} \right] \\
 &= \frac{\rho^{m-1} p_0}{(m-1)!} [\Gamma(m/\beta) E_1(\alpha; m/\beta) - 1] \tag{7}
 \end{aligned}$$

Applying standard procedures, it is possible to show that at small values of the parameter β , close to zero, this value acts like

$$p_W(\beta) \approx p_W(0) \left[1 - \frac{\rho}{(m-\rho)^2} \beta + \frac{\rho(m+2\rho)}{(m-\rho)^4} \beta^2 \right] \tag{8}$$

or

$$p_W(0) - p_W \approx \frac{\rho p_W(0)}{(m-\rho)^2} \left[\beta - \frac{m+2\rho}{(m-\rho)^2} \beta^2 \right].$$

$p_W(0) = p_W(\beta = 0) = \frac{\rho^m p_0}{(m-1)!(m-\rho)}$ - is Erlangian classical formula for the M/M/m model. As we see, at $\beta \neq 0$ p_W is strictly less $p_W(0)$.

The mean number of claims under service (the mean number of engaged servers),

$$\begin{aligned}
 \bar{m} &= \sum_{k=0}^m k p_k + \sum_{k=m+1}^{\infty} m p_k = p_0 \sum_{k=0}^m k \frac{\rho^k}{k!} + \frac{\rho^m p_0}{(m-1)!} \sum_{k=m+1}^{\infty} \frac{\alpha^{k-m}}{(m/\beta + 1)_{k-m}} \\
 &= \rho p_0 \sum_{k=1}^m \frac{\rho^{k-1}}{(k-1)!} + \frac{\rho^m p_0}{(m-1)!} \sum_{k=1}^{\infty} \frac{\alpha^k}{(m/\beta + 1)_k} \\
 &= \rho p_0 e_{m-1}(\rho) + \frac{\rho^m p_0}{(m-1)!} \left[\sum_{k=0}^{\infty} \frac{\alpha^k}{(m/\beta + 1)_k} - 1 \right] \\
 &= \rho p_0 \left[e_{m-2}(\rho) + \frac{\rho^{m-1}}{(m-1)!} \Gamma(m/\beta + 1) E_1(\alpha; m/\beta + 1) \right] \\
 &= \rho p_0 \left\{ e_{m-2}(\rho) + \frac{\rho^{m-1}}{(m-1)!} [\Gamma(m/\beta) E_1(\alpha; m/\beta) - 1] \frac{m}{\rho} \right\},
 \end{aligned}$$

according to the recurrence relations (5). As due to (6)

$$e_{m-2}(\rho) = \frac{1}{p_0} - \frac{\rho^{m-1}}{(m-1)!} \Gamma(m/\beta) E_1(\alpha; m/\beta),$$

thereof it follows

$$\begin{aligned} \bar{m} &= \rho p_0 \left\{ \frac{1}{p_0} + \frac{\rho^{m-1}}{(m-1)!} \frac{m-\rho}{\rho} \Gamma(m/\beta) {}_1(\alpha; m/\beta) - \frac{\rho^{m-1}}{(m-1)!} \frac{m}{\rho} \right\} \\ &= \rho - \frac{\rho^{m-1}}{(m-1)!} [m - (m-\rho) \Gamma(m/\beta) {}_1(\alpha; m/\beta)] p_0 \end{aligned}$$

or

$$\begin{aligned} \bar{m} &= \rho - (m-\rho) \left[\frac{\rho^m}{(m-1)! (m-\rho)} p_0 - p_W \right] \\ &= \rho - (m-\rho) [p_W(0) - p_W] \end{aligned} \tag{9}$$

due to the relation (7). QS load coefficient of this type

$$\begin{aligned} l.c. &= \rho/m - (1 - \rho/m) \left[\frac{\rho^m}{(m-1)! (m-\rho)} p_0 - p_W \right] \\ &= \rho/m - (1 - \rho/m) [p_W(0) - p_W], \end{aligned}$$

shutdown coefficient

$$s.c. = (1 - \rho/m) [1 + p_W(0) - p_W].$$

Let us note that $l.c. < l.c.(0)$, then $s.c. > s.c.(0)$, that is apparent. At small values of parameter β

$$\bar{m} \approx \rho - \beta \frac{\rho p_W(0)}{m-\rho} \approx \rho - \beta \bar{l}(0).$$

According to the common formula, the mean number of claims in queue (queue mean length)

$$\begin{aligned} \bar{l} &= \sum_{k=m+1}^{\infty} (k-m) p_k = \frac{\rho^m}{m!} \sum_{k=m+1}^{\infty} (k-m) \frac{\alpha^{k-m}}{(m/\beta + 1)_{k-m}} p_0 \\ &= \frac{\rho^m p_0}{m!} \sum_{k=1}^{\infty} k \frac{\alpha^k}{(m/\beta + 1)_k} = \frac{\rho^{m+1} p_0}{m! \beta} \frac{d}{d\alpha} \sum_{k=1}^{\infty} \frac{\alpha^k}{(m/\beta + 1)_k} \\ &= \frac{\rho^{m+1} p_0}{m! \beta} \frac{d}{d\alpha} \sum_{k=0}^{\infty} \frac{\alpha^k}{(m/\beta + 1)_k} \\ &= \frac{\rho^{m+1} p_0}{m! \beta} \frac{d}{d\alpha} [\Gamma(m/\beta + 1) E_1(\alpha; m/\beta + 1)] \\ &= \frac{\rho^{m+1} p_0}{(m-1)! \beta^2} \Gamma(m/\beta) \frac{d}{d\alpha} E_1(\alpha; m/\beta + 1) \end{aligned}$$

due to the relation (3). On the other hand, it follows from the same formula

$$\begin{aligned}
 E'_1(z; \xi) &= \frac{d}{dz} \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\xi + k)} = \sum_{k=0}^{\infty} k \frac{z^{k-1}}{\Gamma(\xi + k)} \\
 &= \frac{1}{z} \sum_{k=0}^{\infty} (k - \xi + \xi) \frac{z^k}{\Gamma(\xi + k)} \\
 &= \frac{1}{z} \sum_{k=0}^{\infty} (\xi + k) \frac{z^k}{\Gamma(\xi + k)} - \frac{\xi}{z} \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\xi + k)} \\
 &= \frac{1}{z} \left[\sum_{k=0}^{\infty} (\xi + k - 1 + 1) \frac{z^k}{\Gamma(\xi + k)} - \xi \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\xi + k)} \right] \\
 &= \frac{1}{z} \left[\sum_{k=0}^{\infty} (\xi + k - 1) \frac{z^k}{(\xi + k - 1) \Gamma(\xi + k - 1)} \right. \\
 &\quad \left. + \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\xi + k)} - \xi \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\xi + k)} \right] \\
 &= \frac{1}{z} \left[\sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\xi + k - 1)} - (\xi - 1) \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\xi + k)} \right] \\
 &= \frac{1}{z} [E_1(z; \xi - 1) - (\xi - 1) E_1(z; \xi)],
 \end{aligned}$$

from where

$$\begin{aligned}
 \bar{l} &= \frac{\rho^m p_o}{(m - 1)! \beta} (m/\beta) [{}_1(\alpha; m/\beta) - m/\beta {}_1(\alpha; m/\beta + 1)] \\
 &= \frac{\rho^m p_o}{(m - 1)! \beta} \Gamma(m/\beta) \left\{ {}_1(\alpha; m/\beta) - m/\rho \left[{}_1(\alpha; m/\beta) - \frac{1}{\Gamma(m/\beta)} \right] \right\} \\
 &= \frac{\rho^{m-1} p_o}{(m - 1)! \beta} [(\rho - m) \Gamma(m/\beta) {}_1(\alpha; m/\beta) + m] \\
 &= \frac{\rho^{m-1} p_o}{(m - 1)! \beta} [m - (m - \rho) \Gamma(m/\beta) {}_1(\alpha; m/\beta)] \\
 &= \frac{(m - \rho)}{\beta} [p_W(0) - p_W] \tag{10}
 \end{aligned}$$

At small values of parameter β in accordance with formula (8)

$$\bar{l} \approx \frac{\rho p_W(0)}{m - \rho} \left[1 - \frac{m + 2\rho}{(m - \rho)^2} \beta \right] = \bar{l}(0) \left[1 - \frac{m + 2\rho}{(m - \rho)^2} \beta \right].$$

Further, as we know, every claim in the queue is impacted by some kind of the “departure stream” with the intensity, and this, in its turn, means that in average ν - so-called “impatient” unserved claims will leave the mean number of claims \bar{l} in the queue per unit time. Thereby, in total the system will serve per unit time

$$A = \lambda - \nu \bar{l}$$

claims (A – an absolute system flow capacity). In this case, the relative system capacity of such QS, i.e. a share of served claims among all entered the system, will apparently be

$$q = \frac{A}{\lambda} = \frac{\lambda - \nu \bar{l}}{\lambda} = 1 - \frac{\nu}{\lambda} \bar{l} \text{ or } q = 1 - \bar{l} / \alpha.$$

The mean number of engaged servers \bar{m} , as usual, can be obtained by the division of an absolute flow capacity A by one claim service speed μ , there comes the connection

$$\bar{m} = \frac{A}{\mu} = \frac{\lambda - \nu \bar{l}}{\mu} = \rho - \beta \bar{l},$$

which is easy to verify by the relations (9) and (10) obtained above.

In this case the mean number of claims in the system is on the whole

$$\bar{k} = \bar{m} + \bar{l} = \rho + \frac{1 - \beta}{\beta} (m - \rho) [p_W(0) - p_W] = \rho + (1 - \beta) \bar{l}.$$

Time characteristics corresponding to the numerical characteristics obtained above, i.e. the claim mean staying time in the queue and its service mean time, as well as models, are defined by corresponding Little’s formulas

$$\bar{t}_l = \frac{\bar{l}}{A} = \frac{\bar{l}}{\lambda - \nu \bar{l}} \text{ and } \bar{t}_n = \frac{\bar{n}}{A} = \frac{\bar{n}}{\lambda - \nu \bar{l}} = \frac{1}{\mu},$$

then, one claim total staying time in the system is $\bar{t} = \frac{\bar{k}}{A}$.

It is easy to check that at $\beta \rightarrow 0$ the obtained in this work system of formulas pass to the corresponding relations of the M/M/m model. Thus, it should be noted, however, the following fundamental difference between these two queuing systems.

In the M/M/m system the steady-state limiting mode exists only in the case when $\rho < m$, as at $\rho \geq m$ the geometrical progressions in common ratios of this model formulas diverge, physically it corresponds to the unlimited queue growth at $t \rightarrow \infty$.

On the contrary, in QS with “impatient” claims all the claims leave the queue sooner or later (either as a result of being serviced, or without being serviced) and therefore, the steady-state mode of service at $t \rightarrow \infty$ in it is always reached, independently of the given intensity of claims stream ρ . From the point of view of mathematics, it follows from the fact that series in corresponding formulas denominators in this case meet at any positive values of ρ and β . Let us also note that for QS with “impatient” claims, the concept of refusal probability does

not make sense as every claim entering the system stands in the queue, but it can be unserved if it leaves the queue before its due time.

In conclusion we will note that this model formalization can be carried out with the application of Kummer's confluent hypergeometric function [8]

$${}_1F_1(a; b; z) = \sum_{k=0}^{\infty} \frac{(a)_k}{(b)_k} \frac{z^k}{k!},$$

from where ${}_1F_1(1; b; z) = \Gamma(b) E_1(z, b)$, as $(1)_k = k!$. In this case we obviously have

$$S = {}_1F_1(1; m/\beta + 1; \alpha) = \frac{m}{\rho} [{}_1F_1(1; m/\beta; \alpha) - 1],$$

so that

$$p_0 = \left[e_{m-2}(\rho) + \frac{\rho^{m-1}}{(m-1)!} {}_1F_1(1; m/\beta; \alpha) \right]^{-1};$$

$$p_W = 1 - e_{m-1}(\rho) p_0 = \frac{\rho^{m-1} p_0}{(m-1)!} [{}_1F_1(1; m/\beta; \alpha) - 1].$$

and all other formulas remain unchanged.

It is also easy to see that relations obtained above remain true for the open queuing system model with the claim limited mean staying time in the system on the whole (i.e. both in queue, and under service) as well. The obvious replacement thus consists in the replacement of μ for $\mu + \xi$, where $\xi = 1/\bar{\tau}$, and $\bar{\tau}$ is the mean staying time of one unserved or not fully served claim in the system. It is clear, that in the obtained above relations ρ should be replaced for $\hat{\rho} = \frac{\lambda}{\mu + \xi}$, β for $\hat{\beta} = \frac{\xi}{\mu + \xi}$ and α for $\hat{\alpha} = \frac{\lambda}{\xi}$.

The first of these values obviously means the mean number of claims entering the system during the mean time when the claim is in the system (including served, unserved and not fully served claims). The second value is the mean number of claims leaving the system unserved during the same time, including not fully served claims. The third is the mean number of claims entering the system during the mean staying time of one unserved or not fully served claim in the system.

In the future we plan to study the use of Mittag-Leffler function to the problem of queues in queuing systems of multicomponent flows, presented in a series of works [9–15].

References

1. Saaty, T.L.: Elements of Queueing Theory with Applications. McGraw-Hill Book Company Inc., New York, Toronto, London (1961)
2. Kleinrock, L.: Teoriya massovogo obsluzhivaniya [The Queueing Theory]. Mashinostroyenie Publ., Moscow (1979). (in Russian)

3. Kirpichnikov, A.P.: *Metody prikladnoy teorii massovogo obsluzhivaniya* [Methods of Applied Queuing Theory]. Publishing Office of KSU Publ., Kazan (2011). (in Russian)
4. Feller, W.: *Vvedenie v teoriyu veroyatnostey i eyo prilozheniya* [Introduction to Probability Theory and its Applications]. Mir Publ., Moscow (1964). (in Russian)
5. Prudnikov, A.P., Brychkov, Yu.A., Marichev, O.I.: *Integraly i ryady. Spetsial'niye funktsii* [Integrals and Series. Special Functions]. Nauka Publ., Moscow (1983). (in Russian)
6. Dzhrbashyan, M.M.: *Integral'niye preobrazovaniya i predstavleniya funktsiy* [Integral Transformations and Representations of Functions]. Nauka Publ., Moscow (1966). (in Russian)
7. Goldberg, A.A., Ostrovsky, I.V.: *Raspredeleniye znacheniy meromorfnykh funktsiy* [Value Distribution of Meromorphic Functions]. Nauka Publ., Moscow (1970). (in Russian)
8. Prudnikov, A.P. Brychkov, Yu.A. Marichev, O.I.: *Integraly i ryady. Dopolnitel'niye glavy* [Integrals and Series. Complementary Chapters]. Nauka Publ., Moscow (1985). (in Russian)
9. Kirpichnikov, A.P., Titovtsev, A.S.: Open systems of multicomponent flows differentiated service. *Ciência e Técnica Vitivinícola* **29**(7), 108–122 (2014)
10. Kirpichnikov, A., Titovtsev, A.: Mathematical model of a queuing system with arbitrary quantity of sources and size-limited queue. *Int. J. Pure Appl. Math.* **106**(2), 649–661 (2016)
11. Kirpichnikov, A., Titovtsev, A.: Mathematical model of open queuing system with full set of memories. *Int. J. Pure Appl. Math.* **107**(1), 139–143 (2016)
12. Kirpichnikov, A., Titovtsev, A.: Physical and mathematical queues in the applied queuing theory. *Int. J. Pure Appl. Math.* **108**(2), 409–418 (2016)
13. Titovtsev, A.: The concept of higher orders queues in the queuing theory. *Int. J. Pure Appl. Math.* **109**(2), 451–457 (2016)
14. Kirpichnikov, A., Titovtsev, A.: Generalized Little's formulas and classification of higher orders queues in the queuing theory. *Int. J. Pure Appl. Math.* **114**(4), 819–822 (2017)
15. Kirpichnikov, A., Titovtsev, A.: On the problems of queues in mixed type queuing systems with random quantity of sources and size-limited queues. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) *ITMM 2017. CCIS*, vol. 800, pp. 68–82. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_6



A Contribution to Modeling Two-Way Communication with Retrial Queueing Systems

Attila Kuki^(✉), János Sztrik, Ádám Tóth, and Tamás Bérczes

University of Debrecen, Debrecen, Hungary

{attila.kuki,janos.sztrik,adam.toth,tamas.berczes}@inf.unideb.hu

Abstract. The goal of this paper is to investigate the two-way communication by the help of finite source retrial queueing systems. Incoming calls from sources (primary calls) arrive to the server according to a Poisson process. If an incoming call finds the server idle, its service starts. Otherwise, if the server is busy, an arriving (primary or secondary - from the orbit) call moves into the orbit and after some exponentially distributed time it retries to enter to the server. When the server is idle it generates an outgoing call after an exponentially distributed time with different parameters to the calls in the orbit and in the sources, respectively. The service time of the incoming and outgoing calls are exponentially distributed with different rates. Results on two-way communication assume, that after the service an outgoing call (primary or secondary) is sent back to the source. The novelty of this paper is investigating two cases. In Case 1 the secondary outgoing call is sent back to the orbit, thus the pending incoming call will not be lost. In Case 2 after service of the secondary outgoing call its incoming service request will be started immediately. This means a two-phase service. The balance equations are solved by the help of MOSEL-2 tool. Graphical results and comparisons of the cases are presented.

Keywords: Finite-source queueing system · Retrial queues
Call centers · Two-way communication

1 Introduction

This paper deals with investigations on systems with two-way communication. These systems can be modeled effectively by the help of retrial queueing systems. The research on two-way communications has been becoming more and more popular topic of investigations for the last years. The main reason is that there are a many application fields which can be modeled by this type of systems. For example, in business organizations, e.g. in call centers where the agents could perform outgoing calls to sell, advertise and promote products and services of the business. It is very important to increase the utilization the server, see for example [1, 2, 8, 13, 17, 20]. The most important characteristics of two-way

communication is that an idle server can look for calls inside and outside of the system. In other words, it can perform outgoing calls to the source (primary calls) or to the orbit (secondary calls). The first results on infinite source retrial queueing systems with two-way communication was published by Falin [12], followed by some recent ones, see for example [3, 6, 9, 10, 14–16, 18, 19].

Authors have investigated the case, when a secondary outgoing call after servicing is sent back to the source [11]. The novelty of this paper is, that a more realistic case is considered regarding secondary outgoing calls from the orbit. A call being in the orbit implies that the call still has an unserved incoming request. So far, the server makes a secondary outgoing call from the orbit, serves the request, and sends back the call to the source. In this case the original incoming request of this call remains unserved. In the model presented here the served secondary outgoing call (an outgoing call from the orbit) is sent back to the orbit again, where the call is able to retry his request for servicing the original incoming call. In addition, in this model an other operational mode is investigated. When a secondary outgoing call from the orbit arrives to the server, after serving the outgoing call, the pending incoming request will be served immediately, as well. When this two-phase service is finished, the call is sent back to the source.

The rest of the paper is organized as follows. In Sect. 2 description of the model is given, the corresponding 2-dimensional Markov process is defined. In Sect. 3 the most interesting results obtained by MOSEL-2 tool are presented. Finally, the paper ends with a Conclusion.

2 Model Description and Notations

This paper deals with a finite source retrial queueing model with one server. The work flow of the model can be seen on Fig. 1.

In the source there are N calls. Each call can make a primary incoming call (incoming calls in the system) towards the server. The inter-request times are exponentially distributed with parameter λ_1 . When the server is idle, it starts serving the call immediately with an exponentially distributed service time with parameter μ_1 . After the service is finished, the call goes back to the source. When the incoming call finds the server busy, it is forwarded to the orbit. This secondary incoming jobs from the orbit may retry their requests for service after a random waiting time. The distribution of this period is exponential with parameter ν_1 . In the other hand, the idle server after some exponentially distributed period can make an outgoing calls towards the sources (outgoing calls in the system). Two types of outgoing calls are distinguished:

- After an exponentially distributed idle period with parameter λ_2 the server may call a call from the source to be served (primary outgoing call),
- the server is able to make a call from the orbit, as well (secondary outgoing call). It is performed after an exponentially distributed idle period with parameter ν_2 .

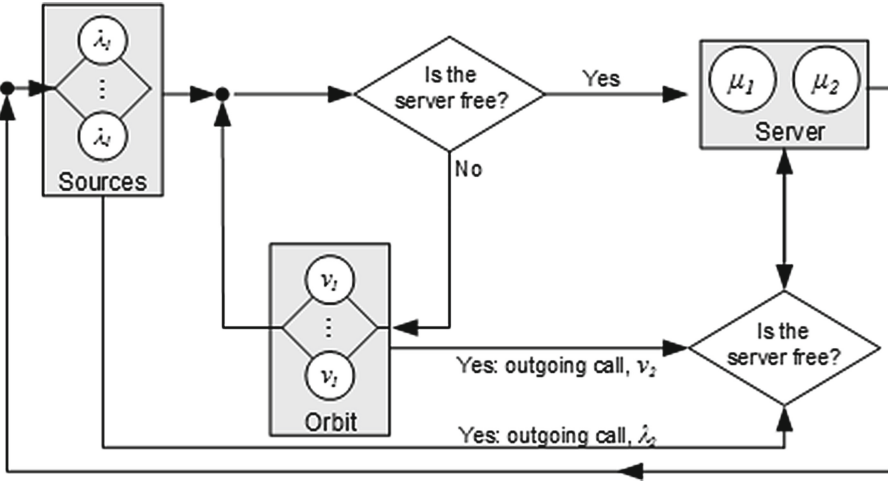


Fig. 1. A retrial queue with components

The outgoing calls (primary and secondary) are served at the server in an exponentially distributed service period with parameter μ_2 . A primary outgoing call (called from the source) goes back to the source after the service. When an outgoing call comes from the orbit (secondary outgoing call), two cases can be considered.

- Case 1. The call came from the orbit, which means this call has an unserved incoming request. After the outgoing call is served, this incoming request remains unserved. This call is sent back to the orbit after the outgoing service is finished, thus the call will be able to retry its incoming call,
- Case 2. As in the previous case, after the outgoing call is processed, it has an unserved primary call. In this case the server is able to serve the incoming request immediately after the outgoing job was finished. That means a two-phase service. First the outgoing call is served, after the incoming one. When both of the service phases has been finished, the call is sent back to the source.

It is assumed that the arrivals of primary incoming calls, retrial intervals of secondary incoming calls, service times of incoming and outgoing calls, and the time to make outgoing calls are mutually independent.

We denote the number of calls in orbit and the server state at time t by $O(t)$ and $S(t)$, respectively.

Obviously, when the server is busy the number of calls in the orbit cannot be equal to N , i.e. $O(t) < N$. As stated above in Case 2, after the service both incoming and the two-phase outgoing calls go to a free state. This means that when the server is idle, there will be at least one call in free state, i.e. again $O(t) < N$. Thus, the state space of the process $(S(t), O(t))$ is the set of $\{0, 1, 2\} \times \{0, 1, 2, \dots, N - 1\}$. In Case 1 after the service the secondary outgoing call goes back to the orbit. This means that when the server is idle, the source

can be empty. Thus, the state space of the process $(S(t), O(t))$ is the set of $\{0, 1, 2\} \times \{0, 1, 2, \dots, N\}$.

Because of the finite state space these two-dimensional Markov processes are always stable.

Let us define the state of the server by $S(t)$, that is

$$S(t) = \begin{cases} 0, & \text{when the server is idle} \\ 1, & \text{when an first order request is in service} \\ 2, & \text{when a second order request from source is in service} \\ 3, & \text{when a second order request from orbit is in service} \end{cases}$$

The used numerical values of the parameters can be seen in Table 1. Some special values of the parameters in the model described above give back models which have been investigated earlier by authors.

- $\lambda_2 = \nu_2 = 0$ provides a classical single server retrial queue studied by e.g. [4, 5].
- $\lambda_2 = 0, \mu_2 = \mu_1$ provides a single server retrial queue with two-way communication with search of the customers from the orbit. The reason of the outgoing calls is to shorten the idle period of the server.
- $\mu_2 = \mu_1$ provides a single server retrial queue with two-way communication with search of the customers. The reason of the outgoing calls is again to shorten the idle period of the server.

Table 1. Numerical values of model parameters

Parameter	Symbol	Value
Number of calls in source	N	10
Incoming generation rate	λ_1	[0.1..5.1]
Outgoing generation rate	λ_2	0.2
Incoming retrial rate	ν_1	0.1
Outgoing generation rate from orbit	ν_2	0.2
Incoming service rate	μ_1	1
Outgoing service rate	μ_2	1

It is not difficult to see that the system of balance equations for the stationary probabilities in Case 1 are

$$\begin{aligned}
 p_{i,j} &= \lim_{t \rightarrow \infty} P(S(t) = i, O(t) = j), i = 0, 1, 2, 3 \text{ and } j = 0, 1, \dots, N \\
 [(N - j)(\lambda_1 + \lambda_2) + j(\nu_1 + \nu_2)] p_{0,j} &= \mu_1 p_{1,j} + \mu_2 p_{2,j} + \mu_2 p_{3,j-1} \\
 [(N - j - 1)\lambda_1 + \mu_1] p_{1,j} &= (N - j)\lambda_1 p_{0,j} + (j + 1)\nu_1 p_{0,j+1} + (N - j)\lambda_1 p_{1,j-1} \\
 [(N - j - 1)\lambda_1 + \mu_2] p_{2,j} &= (N - j)\lambda_2 p_{0,j} + (N - j)\lambda_1 p_{2,j-1}
 \end{aligned}$$

$$[(N - j - 1)\lambda_1 + \mu_2] p_{3,j} = (j + 1)\nu_2 p_{0,j+1} + (N - j)\lambda_1 p_{3,j-1}$$

with $p_{1,-1} = p_{2,-1} = p_{3,-1} = 0$.

Similarly, the system of balance equations for the stationary probabilities in Case 2 can be written as

$$p_{i,j} = \lim_{t \rightarrow \infty} P(S(t) = i, O(t) = j), i = 0, 1, 2, 3, 4 \text{ and } j = 0, 1, \dots, N - 1$$

$$[(N - j)(\lambda_1 + \lambda_2) + j(\nu_1 + \nu_2)] p_{0,j} = \mu_1 p_{1,j} + \mu_2 p_{2,j}$$

$$[(N - j - 1)\lambda_1 + \mu_1] p_{1,j} =$$

$$= (N - j)\lambda_1 p_{0,j} + (j + 1)\nu_1 p_{0,j+1} + (N - j)\lambda_1 p_{1,j-1} + \mu_2 p_{3,j}$$

$$[(N - j - 1)\lambda_1 + \mu_2] p_{2,j} = (N - j)\lambda_2 p_{0,j} + (N - j)\lambda_1 p_{2,j-1}$$

$$[(N - j - 1)\lambda_1 + \mu_2] p_{3,j} = (j + 1)\nu_2 p_{0,j+1} + (N - j)\lambda_1 p_{3,j-1}$$

with $p_{0,N} = p_{1,-1} = p_{2,-1} = p_{3,-1} = 0$.

As soon as we have calculated the distributions defined above (by the help of MOSEL-2 tool, see the next section), the most important steady-state system performance measures can be obtained in the following way.

– *Utilization 1*

$$U_1 = \sum_{o=0}^N P(1, o)$$

– *Utilization 2*

$$U_2 = \sum_{s=2}^3 \sum_{o=0}^N P(s, o)$$

– *Average number of jobs in the orbit*

$$\bar{O} = \sum_{s=0}^3 \sum_{o=0}^N o P(s, o)$$

– *Average number of active primary users*

$$\bar{M} = N - \bar{O} - U_1 - U_2$$

– *Average generation rate of primary users*

$$\bar{\lambda}_1 = \lambda_1 \bar{M}$$

– *Mean time spent in orbit by using Little-formula*

$$\bar{W} = \frac{\bar{O}}{\bar{\lambda}_1}$$

3 Numerical Results

Investigating the functionality and the behavior of the system several numerical calculations were performed. Solving the system balance equations described above the MOSEL-2 tool was used. For Markov-processes it is a very efficient tool. MOSEL-2 is a model description language and are equipped with several model translators. Using these translators third-party performance evaluation tools can be used. For obtaining the stationary system probabilities, here the SPNP tool is used. SPNP performs numerical calculations instead of simulation (see in [7]). From the probabilities the well known system characteristics are also be calculated. The most interesting performance characteristics obtained by these tools are graphically presented in this section. On the figures the lines represent different working assumptions or cases. The applied values for the parameters are listed in Table 2.

Table 2. Numerical values of model parameters

Case studies											
No.	N	λ_1	λ_2	ν_1	ν_2	μ_1	μ_2	C_W	C_1	C_2	C_P
Figure 2	10	0.1..5.1	0.2	0.1	0.2	1	1				
Figure 3	10	0.1..5.1	0.2	0.1	0.2	1	1				
Figure 4	10	0.1..5.1	0.2	0.1	0.2	1	1				
Figure 5	10	0.1..5.1	0.2	0.1	0.2	1	1				
Figure 6	10	0.1..5.1	0.2	0.1	0.2	1	1	100	10	10	1
Figure 7	10	0.1..5.1	0.2	0.1	0.2	1	1	100	10	10	1

On Figs. 2, 3, and 4 the mean waiting time of the calls are represented in function of the incoming generation rate for Case 1, Case 2 and comparing the two cases, respectively.

On the first two figures four cases are displayed. “No outgoing” means, that there are only incoming calls in the system. This is a common finite source retrial system. “Outgoing - Only from source” is for the case, when only primary outgoing calls are performed. The line “Outgoing - Only from orbit” is for secondary outgoing calls only. The fourth line represents the investigated Case 1. The similar lines are on the figure for Case 2. Note that, the “No outgoing” lines are the same. The reason of the virtual deviation is the different scale of axes y .

For these values of parameters except the “Outgoing - Only from source” case an interesting maximum value of the mean waiting time can be observed. When we consider a simple retrial queueing system, it can be found a parameter setting, where this maximum feature can be observed. This is a general characteristics of the retrial queues. With slightly modifications of the parameters the same maximum effect also appears here. On Fig. 4 the Case 1 and Case 2 are compared

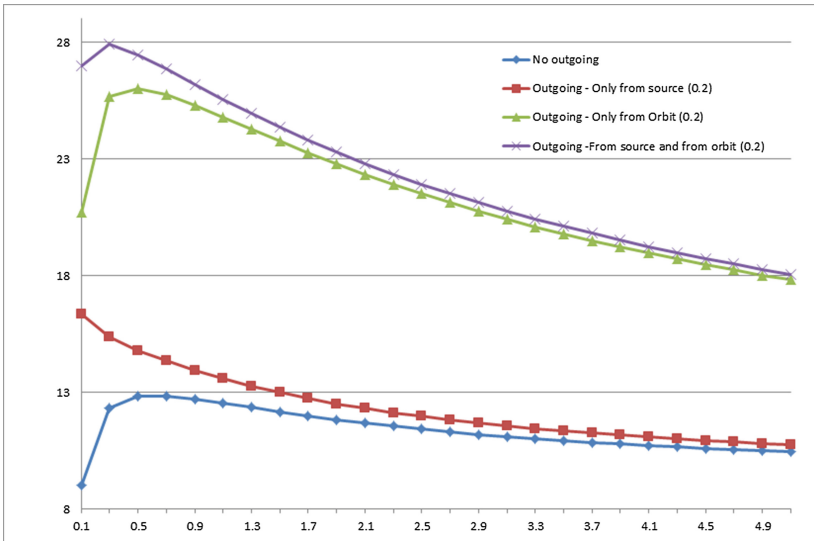


Fig. 2. Mean waiting time (Case 1) vs. λ_1

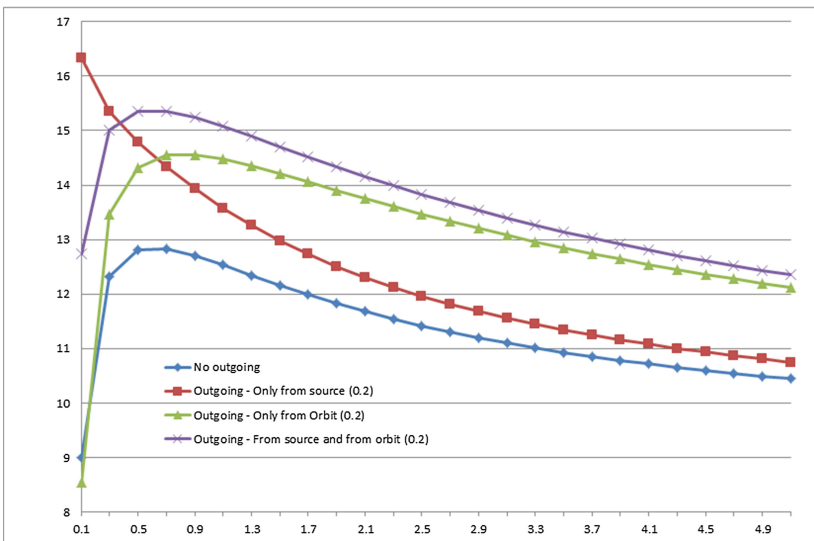


Fig. 3. Mean waiting time (Case 2) vs. λ_1

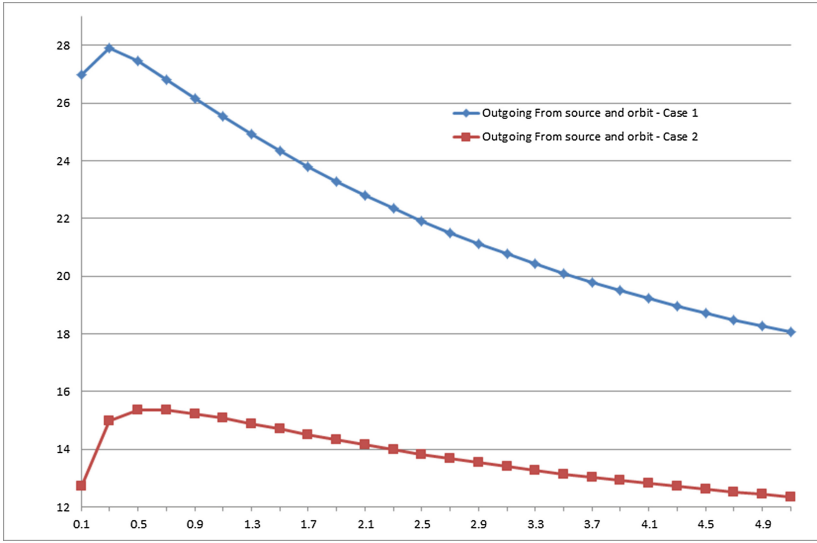


Fig. 4. Mean waiting time vs. λ_1

with all of the incoming and outgoing calls. This figure reflects and ensures the expected behaviour of Case 1 and 2.

Figure 5 displays the probability of the busy server in Case 1. That means, the server state can be $S(t) = 1, 2, 3$. The running parameter is the incoming arrival rate, λ_1 again. Pairs of lines can be observed on this figure. One pair is the outgoing and outgoing from orbit only lines, while the other pair is the no outgoing and the outgoing from source only lines. For this set of parameters for smaller values of λ_1 the first pair while for larger values of λ_1 the second pair has larger values, i.e. higher server utilization.

In this type of service or production systems the waiting time of calls and the utilization of the server are singular quantities. They cannot be optimized at the same time. Optimizing the server utilization will increase the waiting time of calls. Some balance or some combined indicator has to be involved. The following expected loss $E(L)$ function enables the “fair” investigation of the system.

$$E(L) = C_w(1 - U_1 - U_2) + C_1\mu_1U_1 + C_2\mu_2U_2 + C_P(E(O) + U_1 + U_2).$$

The first component is the loss on idle state of the system. The second and the third components are the cost of servicing incoming and outgoing calls, respectively. Here the speed of the service has to be taken into consideration, thus beside the cost weights the service rates are present as multiplicative factors. The last component states the loss of the system not in production from the point of view of the calls, i.e. it is the sojourn time of the call: it is under service or it is in the virtual waiting facility (orbit). For a given set of parameters (listed in Table 2) the shape of loss functions can be seen on Figs. 6 and 7.

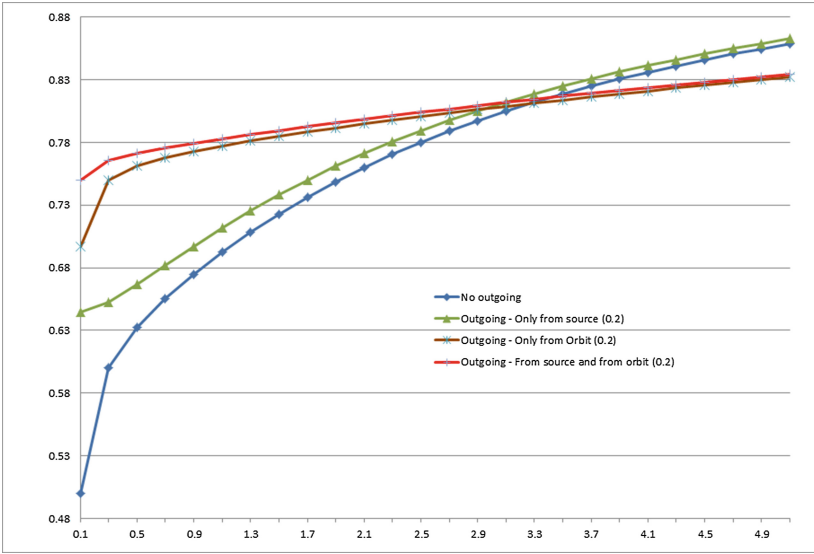


Fig. 5. Probability of server is busy vs. λ_1

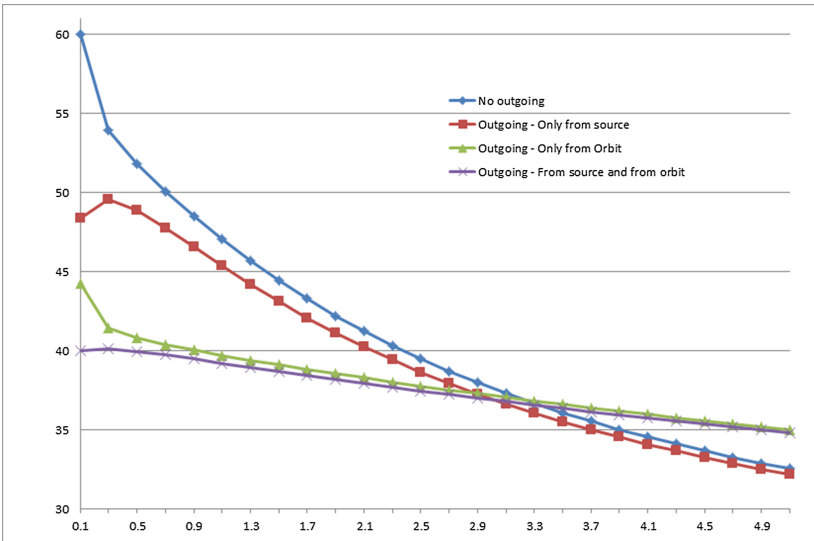


Fig. 6. The loss function in Case 1 vs. λ_1

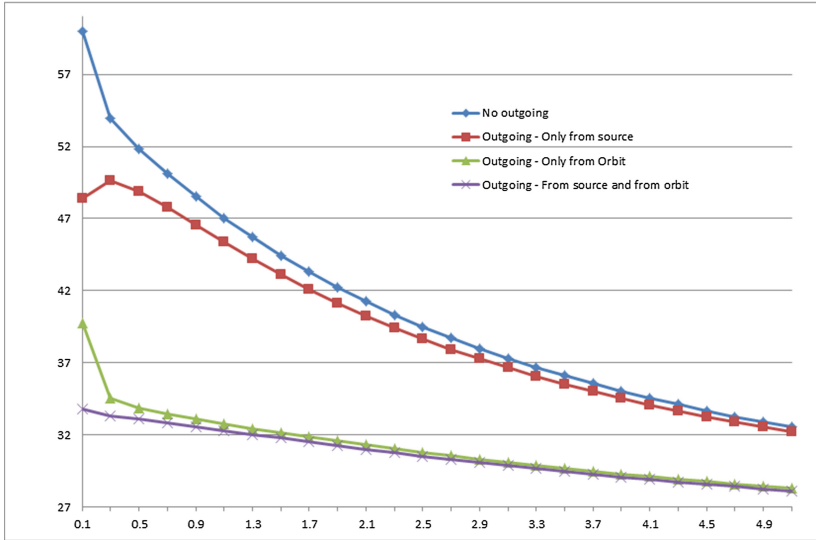


Fig. 7. The loss function in Case 2 vs. λ_1

In Case 1 and Case 2 it can be observed that $E(L)$ values are much lower in the two-way cases, especially when the orbit is involved. In Case 1 these values are higher than in Case 2. It is the effect of the two-phase service of the calls. For the one-phase service (Case 1) it is interesting, that large values of the incoming rate implies lower values of loss functions for incoming only and outgoing from source only cases (see the line intersections on Fig. 6). The same effect can be seen on Fig. 5, as well. The reason of this similar behavior is that the first three components of the loss function contain the effect of the utilizations, and these components have large weights.

4 Conclusion

This paper gives a contribution to the model described in [11]. The original model stated, that a secondary outgoing call is sent back to the source after service. Let's consider a bank, where the calls are called to give some signature sample (outgoing calls). These calls can be outside the bank (in free state) or inside the bank, waiting for some transaction (incoming calls in the orbit). When the call connected from the orbit for the outgoing call, it is quite natural not send it outside the bank but keep it inside (Case 1) or after the signature perform its original transaction request (Case 2). The numerical results proof that in Case 2 the most important performance measures (waiting time, utilization etc.) are better than in Case 1. A loss function keeping balance between utilization and waiting times has been also introduced. In the future it would be interesting to investigate the sensitivity of the loss function to the parameter changing.

Acknowledgments. The research work of Attila Kuki, János Sztrik, and Tamás Bérczes was granted by Austrian-Hungarian Bilateral Cooperation in Science and Technology project 2017-2.2.4-TÉT-AT-2017-00010.

The research work of Ádám Tóth was supported by the construction EFOP-3.6.3-VEKOP-16-2017-00002. The project was supported by the European Union, co-financed by the European Social Fund.

References

1. Aguir, S., Karaesmen, F., Akşin, O.Z., Chauvet, F.: The impact of retrials on call center performance. *OR Spectr.* **26**(3), 353–376 (2004)
2. Aksin, Z., Armony, M., Mehrotra, V.: The modern call center: a multi-disciplinary perspective on operations management research. *Prod. Oper. Manag.* **16**(6), 665–688 (2007)
3. Artalejo, J.R., Phung-Duc, T.: Markovian retrial queues with two way communication. *J. Ind. Manag. Optim.* **8**(4), 781–806 (2012)
4. Artalejo, J.: Retrial queues with a finite number of sources. *J. Korean Math. Soc.* **35**, 503–525 (1998)
5. Artalejo, J., Corral, A.G.: *Retrial Queueing Systems: A Computational Approach*. Springer, Heidelberg (2008). <https://doi.org/10.1007/978-3-540-78725-9>
6. Artalejo, J., Phung-Duc, T.: Single server retrial queues with two way communication. *Appl. Math. Model.* **37**(4), 1811–1822 (2013)
7. Begain, K., Bolch, G., Herold, H.: *Practical Performance Modeling. Application of the MOSEL Language*. Kluwer Academic Publisher, Boston (2001)
8. Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., Zhao, L.: Statistical analysis of a telephone call center: a queueing-science perspective. *J. Am. Stat. Assoc.* **100**(469), 36–50 (2005)
9. Dimitriou, I.: A retrial queue to model a two-relay cooperative wireless system with simultaneous packet reception. In: Wittevrongel, S., Phung-Duc, T. (eds.) *ASMTA 2016*. LNCS, vol. 9845, pp. 123–139. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43904-4_9
10. Dragieva, V., Phung-Duc, T.: Two-way communication M/M/1 retrial queue with server-orbit interaction. In: *Proceedings of the 11th International Conference on Queueing Theory and Network Applications*, p. 11. ACM (2016)
11. Dragieva, V., Phung-Duc, T.: Two-way communication M/M/1//N retrial queue. In: Thomas, N., Forshaw, M. (eds.) *ASMTA 2017*. LNCS, vol. 10378, pp. 81–94. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61428-1_6
12. Falin, G.: Model of coupled switching in presence of recurrent calls. *Eng. Cybern.* **17**(1), 53–59 (1979)
13. Gans, N., Koole, G., Mandelbaum, A.: Telephone call centers: tutorial, review, and research prospects. *Manuf. Serv. Oper. Manag.* **5**(2), 79–141 (2003)
14. Nazarov, A., Phung-Duc, T., Paul, S.: Heavy outgoing call asymptotics for *MMPP/M/1/1* retrial queue with two-way communication. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) *ITMM 2017*. CCIS, vol. 800, pp. 28–41. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_3
15. Nazarov, A.A., Paul, S., Gudkova, I., et al.: Asymptotic analysis of Markovian retrial queue with two-way communication under low rate of retrials condition. In: *Proceedings 31st European Conference on Modelling and Simulation* (2017)

16. Phung-Duc, T., Rogiest, W.: Two way communication retrieval queues with balanced call blending. In: Al-Begain, K., Fiems, D., Vincent, J.-M. (eds.) ASMTA 2012. LNCS, vol. 7314, pp. 16–31. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-30782-9_2
17. Pustova, S.: Investigation of call centers as retrieval queuing systems. *Cybern. Syst. Anal.* **46**(3), 494–499 (2010)
18. Sakurai, H., Phung-Duc, T.: Two-way communication retrieval queues with multiple types of outgoing calls. *Top* **23**(2), 466–492 (2015)
19. Sakurai, H., Phung-Duc, T.: Scaling limits for single server retrieval queues with two-way communication. *Ann. Oper. Res.* **247**(1), 229–256 (2016)
20. Wolf, T.: System and method for improving call center communications, US Patent App. 15/604,068, 30 November 2017



Steady State Probabilistic Characteristics of the On/Off Production Rate Control Production-Inventory System with MMPP Demand Arrivals

Klimentii Livshits^(✉), Anna Kitaeva, and Ekaterina Ulyanova

National Research Tomsk State University, Tomsk, Russia
kim47@mail.ru, kit1157@yandex.ru, ulyanovaeks@gmail.com

Abstract. Approximate steady state probability distribution of the stock level for the single product production-inventory system when the demand is a Markov-modulated Poisson process (MMPP) with finite number of states under the on/off production rate control is obtained. The control causes the stock level to fluctuate around a given value by reducing the production rate when the stock exceeds this value. The asymptotic distributions of the stock-out and overproduction periods are also obtained. Exact steady state distribution of the stock level for MMPP demand arrivals with two states and exponential batch size distribution is compared with the approximate one.

Keywords: Production-inventory system · On/off control
Markov-modulated Poisson process demand arrivals
Steady state distribution

1 Introduction and Problem Statement

A systematic study of inventory models incorporated uncertainly and dynamics began in the early 50s from the works by Arrow et al. [1] and Dvoretzky et al. [2]. Nowadays a set of stochastic models are available to solve the inventory control problem under various conditions encountered in practice, for example, Ross [3], Chopra and Meindl [4], Beyer et al. [5], Nazarov and Broner [6].

The purpose of this paper is to stabilize a supply system performance by regulating the production rate.

Let $S(t)$ be a stock level at time t , the product flow (input flow) be continuous with rate $C(S)$, the demand (output flow) be a Markov-modulated Poisson process (MMPP), i.e. the intensity of a Poisson process of the customers arrivals is defined by the state of a Markov chain with n states: if at time t the Markov process has value $i = 1, 2, \dots, n$ then the customers are arriving according to a Poisson process with intensity $\lambda_i > 0$. The amounts required at each arrival are distributed according to probability density function (PDF) $\varphi(\cdot)$ and are

independent of everything else. Denote the first and second moments of the distribution respectively as a_1 and a_2 .

MMPP processes submit a flexible way of modeling demand for inventory systems, see, for example, Abhyankar and Graves [7]. A collection of results about Markov-modulated Poisson processes is given in Fischer and Meier-Hellstern [8]. Here we are going to consider MMPP only in the steady state.

Denote $Q = [q_{ij}]$ an infinitesimal generator of a n -state continuous-time Markov chain determining the customers' arrival rate in a Poisson process, here $q_{ij} \geq 0$, if $i \neq j$, and

$$\sum_{j=1}^n q_{ij} = 0. \tag{1}$$

Let $\gamma_i, i = \overline{1, n}$ be simple eigenvalues of $Q, \gamma_n = 0, \gamma_i < 0, i = \overline{1, n-1}$. Then system

$$\sum_{i=1}^n q_{ij} \pi_i = 0, \tag{2}$$

$$\pi_1 + \pi_2 + \dots + \pi_n = 1. \tag{3}$$

has an unique solution that gives us the steady state distribution of the Markov chain.

Suppose that for some reasons we are interested in keeping the stock level near some base-stock level S_0 . Let us consider the on/off control of the production rate, that is, the product flow has constant rate C_0 , if the inventory level $S(t)$ is below S_0 ; otherwise, we decrease the rate of product flow to $C_1 < C_0$.

Thus

$$C(S) = \begin{cases} C_0, & S < S_0, \\ C_1, & S \geq S_0. \end{cases} \tag{4}$$

The aim of such type of control is maybe to avoid the overflow ($S(t) \gg S_0$) and stock-out ($S(t) < 0$).

Let us denote $\lambda_0 = \sum_{i=1}^n \lambda_i \pi_i$ and assume that

$$C_0 > \lambda_0 a, \quad C_1 < \lambda_0 a. \tag{5}$$

The condition $C_0 > \lambda_0 a$ means that if the inventory level is below the base-stock level then the stock level is replenished in the mean, that is, the resources are accumulated, and the second inequality means that if the inventory level is above the base-stock level then the resources are expended. Under these conditions a stock level $S(t)$ is a stationary process.

2 Approximate Stationary Distribution of the Stock Level

Let us denote

$$P_i(s, t) ds = P \{s < S(t) \leq s + ds; \lambda(t) = \lambda_i\}, \quad i = \overline{1, n}.$$

Consider time interval $\Delta t \ll 1$. Let at time $t + \Delta t$ the system is in state i , i.e. the intensity of the customers' flow is λ_i , and $S(t + \Delta t) = s$. Then following backward scenarios are possible:

1. The intensity of the customers' flow have not changed and we have had no purchases at time interval Δt . The probability of the event is equal to $1 + (q_{ii} - \lambda_i)\Delta t + o(\Delta t)$.
2. The intensity have changed, suppose that $\lambda(t) = \lambda_j \neq \lambda_i$, and we have had no purchases at time interval Δt . The probability of the event is equal to $q_{ji}\Delta t + o(\Delta t)$.
3. The intensity have not changed and we have had a purchase x at time interval Δt . The probability of the event is equal to $\lambda_i\Delta t\varphi(x)dx + o(\Delta t)$.
4. The rest possibilities have probabilities $o(\Delta t)$ as $\Delta t \rightarrow 0$.

So, we get equation

$$P_i(s, t + \Delta t) = (1 + (q_{ii} - \lambda_i) \Delta t)P_i(s - C(s)\Delta t, t) + \sum_{j \neq i} q_{ji}P_j(s - C(s)\Delta t, t)\Delta t + \lambda_i\Delta t \int_0^\infty P_i(s - C(s)\Delta t + x, t)\varphi(x)dx + o(\Delta t).$$

Using the Taylor expansion, dividing the equation by Δt and tending Δt to zero, we get

$$\frac{\partial P_i(s, t)}{\partial t} + C(s)\frac{\partial P_i(s, t)}{\partial s} = -\lambda_i P_i(s, t) + \sum_{j \neq i} q_{ji}P_j(s, t) + \lambda_i \int_0^\infty P_i(s + x, t)\varphi(x)dx.$$

Denote stationary distributions of the stock level below and above the base level

$$P_i^0(s) = \lim_{t \rightarrow \infty} P_i(s, t), \quad s < S_0, \quad P_i^1(s) = \lim_{t \rightarrow \infty} P_i(s, t), \quad s \geq S_0. \quad (6)$$

Functions $P_i^0(s)$ and $P_i^1(s)$ are satisfied the equations

$$C_1 \dot{P}_i^1(s) = -\lambda_i P_i^1(s) + \sum_{j=1}^n q_{ji}P_j^1(s) + \lambda_i \int_0^\infty P_i^1(s + x)\varphi(x)dx, \quad s \geq S_0, \quad (7)$$

$$C_0 \dot{P}_i^0(s) = -\lambda_i P_i^0(s) + \sum_{j=1}^n q_{ji}P_j^0(s) + \lambda_i \int_0^{S_0-s} P_i^0(s + x)\varphi(x)dx + \lambda_i \int_{S_0-s}^\infty P_i^1(s + x)\varphi(x)dx, \quad s < S_0, \quad (8)$$

given that

$$\int_{-\infty}^{S_0} P_j^0(s)ds + \int_{S_0}^\infty P_j^1(s)ds = \pi_j \quad (9)$$

and

$$C_0 P_i^0(S_0) = C_1 P_i^1(S_0), \tag{10}$$

which can be obtained by integrating the Eqs. (7) and (8) in the domain of their definition.

Let us consider the case

$$C_0 = (1 + \theta)\lambda_0 a, \quad C_1 = (1 - \alpha\theta)\lambda_0 a, \tag{11}$$

were $\alpha > 0$ and $0 < \theta \ll 1$, $\lambda_0 a$ is an average product amount selling in time unit. Thus, process $S(t)$ will oscillate around S_0 and approximately we can consider $S(t)$ as a diffusion process.

Diffusion methods have been applied in a variety of domains; see Janssen et al. [9]. In inventory modelling beginning with the papers by Bather [10] and Puterman [11], the Brownian motion process is one of the most commonly used demand processes; see, e.g. Rao [12], Rudi et al. [13], and Avinadav [14]. Benkhrouf et al. [15] consider a demand process as a Brownian motion in an inventory system with deterioration.

On/off control model has been investigated in Livshits and Ulyanova [16, 17], Kitaeva [18] and Kitaeva et al. [19], where the stock level process has been considered asymptotically as a diffusion process and its stationary distribution has been obtained. There single-product inventory model with both random and controllable demand and continuous uncontrolled production rate under finite storage capacity has been considered.

But here we are going to consider the other method of obtaining the stationary distribution of the stock level process.

To solve system (7) and (8) let us use the similar method to the one in Livshits and Public [20, 21].

Consider the case $s > S_0$. Let us find the solution in the following form

$$P_i(s) = \theta f_i(\alpha\theta s, \theta), \tag{12}$$

where $f_i(z, \theta)$ is twice differentiable function with respect to z , except, maybe, point $z_0 = \theta S_0$. Let $S_0 = S_0(\theta)$ and $S_0(\theta) \rightarrow \infty$ as $\theta \rightarrow +0$ such a way that

$$\lim_{\theta \rightarrow 0} \theta S_0(\theta) = z_0. \tag{13}$$

Substituting (12) into (7), after the change of variables $z = \alpha\theta s$ we get for $z \geq \alpha z_0$

$$C_1 \alpha \theta \dot{f}_i(z, \theta) = -\lambda_i f_i(z, \theta) + \sum_{j=1}^n q_{ji} f_j(z, \theta) + \lambda_i \int_0^\infty f_i(z + \alpha\theta x, \theta) \varphi(x) dx. \tag{14}$$

It follows from (14) for $\theta \rightarrow 0$ that

$$\sum_{j=1}^n q_{ji} f_j(z, 0) = 0. \tag{15}$$

Therefore, we can take the functions of interest, taking into account (1), in the following form:

$$f_j(z, 0) = \pi_j f(z), \tag{16}$$

where $f(\cdot)$ is some function, which will be defined below.

Define $f_j(z, \theta)$ as follows:

$$f_j(z, \theta) = \pi_j f(z) + h_j(z)\theta + o(\theta). \tag{17}$$

Substituting (17) into (14) and using the linear Taylor expansions of functions $f_j(\cdot, \theta)$, we get as $\theta \rightarrow 0$

$$\sum_{j=1}^n g_{ji} h_j(z) = -(\lambda_i - \lambda_0) a \pi_i \dot{f}(z). \tag{18}$$

Let

$$f_j(z, \theta) = \pi_j f(z) + h_j(z)\theta + g_i(z)\theta^2 + o(\theta^2). \tag{19}$$

Substituting (19) into (14), using the quadratic Taylor expansions of functions $f_j(\cdot, \theta)$, and taking into account (16) and (18), we obtain as $\theta \rightarrow 0$

$$-\sum_{j=1}^n g_{ji} g_j(z) = \frac{\lambda_i a_2}{2} \pi_i \ddot{f}(z) + (\lambda_i - \lambda_0) a \dot{h}_i(z) + \lambda_0 a \pi_i \dot{f}(z). \tag{20}$$

If we sum up equations (20) taking into account (1), we get

$$\frac{\lambda_i a_2}{2} \ddot{f}(z) + \lambda_0 a \dot{f}(z) - \sum_{j=1}^n (\lambda_i - \lambda_0) a \dot{h}_i(z) = 0. \tag{21}$$

It follows from (18) that

$$\sum_{j=1}^n g_{ji} \dot{h}_j(z) = -(\lambda_i - \lambda_0) a \pi_i \dot{f}(z). \tag{22}$$

Consider

$$V = Q^T = R\gamma P, \tag{23}$$

where $R = [R_{ij}]$ is matrix of eigenvectors of V , $P = [P_{ij}] = R^{-1}$, $\gamma = \text{diag}(\gamma_1, \gamma_2, \dots, \gamma_{n-1}, 0)$ is a diagonal matrix of eigenvalues of V . From (2) follows that $R_{in} = \pi_i$ and from relations (23) and (1) we get $\sum_{k,j=1}^n R_{jk} \gamma_k P_{ki} = 0$.

Since the columns of P are linearly independent, it follows that $\sum_{j=1}^n R_{jk} = 0$ for $k = \overline{1, n-1}$. Elements of the n -th row of P satisfy the equations

$$\sum_{j=1}^n P_{nj} R_{jk} = 0, \quad k = \overline{1, n-1}, \quad \sum_{j=1}^n P_{nj} \pi_j = 1,$$

and it follows that $P_{nk} = 1, k = \overline{1, n}$.

Thus, (22) can be written as

$$\gamma_k \sum_{j=1}^n P_{kj} \dot{h}_j(z) = - \sum_{i=1}^n P_{ki} (\lambda_i - \lambda_0) \pi_i a \ddot{f}(z)$$

or

$$\sum_{j=1}^n P_{kj} \dot{h}_j(z) = - \frac{1}{\gamma_k} \sum_{i=1}^n P_{ki} (\lambda_i - \lambda_0) \pi_i a \ddot{f}(z), \quad k = \overline{1, n-1},$$

$$\sum_{i=1}^n P_{ni} \dot{h}_i(z) = c(z),$$

where $c(\cdot)$ is some function. It follows that

$$\dot{h}_i(z) = \sum_{k=1}^{n-1} R_{ik} \frac{1}{\gamma_k} \sum_{j=1}^n P_{kj} (\lambda_0 - \lambda_j) \pi_j a \ddot{f}(z) + \pi_i c(z). \tag{24}$$

Now we can write an equation for function $f(\cdot)$ by substituting (24) into (21):

$$A_2 \ddot{f}(z) + A_1 \dot{f}(z) = 0, \tag{25}$$

where

$$A_1 = \lambda_0 a, \tag{26}$$

$$A_2 = \frac{\lambda_0 a^2}{2} - a^2 \sum_{k=1}^{n-1} \frac{1}{\gamma_k} \sum_{i=1}^n (\lambda_0 - \lambda_i) R_{ik} \sum_{j=1}^n P_{kj} (\lambda_0 - \lambda_j) \pi_j. \tag{27}$$

Let us show that the following quadratic form is negative definite

$$W = \sum_{t=1}^{n-1} \frac{1}{\gamma_t} \sum_{i=1}^n (\lambda_0 - \lambda_i) R_{it} \sum_{j=1}^n P_{tj} (\lambda_0 - \lambda_j) \pi_j.$$

Denote $x_t = \sum_{i=1}^n (\lambda_0 - \lambda_i) R_{it}$, then $\lambda_0 - \lambda_j = \sum_{t=1}^n x_t P_{tj}$; and we can rewrite W as following

$$W = \sum_{t=1}^{n-1} \frac{x_t}{\gamma_t} \sum_{k=1}^n x_k \omega_{tk} = \sum_{t=1}^{n-1} \frac{x_t}{\gamma_t} \sum_{k=1}^{n-1} x_k \omega_{tk},$$

where $\omega_{tk} = \sum_{j=1}^n P_{tj} P_{kj} \pi_j$ and $\omega_{tn} = 0$ since $P_{nj} = 1$ and $\sum_{j=1}^n P_{tj} \pi_j = 0, t \neq n$.

Since matrix P is not degenerate and $\pi_j \geq 0, \sum_j \pi_j = 1$, it follows that $\omega = [\omega_{ij}] > 0$, that is, all the principal minors of this matrix are positive. Denote $\Delta_k(\omega)$ the minor of the k -th order of ω , then the minors of the k -th order of form W we can wright as $\Delta_k = \prod_{j=1}^k \frac{1}{\gamma_j} \Delta_k(\omega)$. Since $\gamma_k < 0$, the signs of the minors Δ_k alternate. Therefore, quadratic form W is negative definite.

Solution (25) has the form

$$f(z) = B_1 + B_2 e^{-\frac{A_1}{A_2} z},$$

where B_1 and B_2 are some constants. Since $\lim_{z \rightarrow +\infty} f(z) = 0$, we have

$$f(z) = B_2 e^{-\frac{A_1}{A_2} (z - z_0)}, \quad z \geq \alpha z_0. \tag{28}$$

Thus, for $s > S_0$

$$P_i^1(s) = B_2 \pi_i \theta e^{-\frac{A_1}{A_2} \alpha \theta (s - S_0)} + o(\theta). \tag{29}$$

Consider the case $s < S_0$. Let us find the solution of (8) in the same form as in the previous case. Substituting (12) into (8), we get

$$C_0 \theta \dot{f}_i(z, \theta) = -\lambda_i f_i(z, \theta) + \sum_{j=1}^n q_{ji} f_j(z, \theta) + \lambda_i \int_0^\infty f_i(z + \theta x, \theta) \varphi(x) dx + \lambda_i R_i(z, \theta), \tag{30}$$

where

$$R_i(z, \theta) = \int_{\frac{z_0 - z}{\theta}}^\infty f_i(z + \theta x, \theta) \varphi(x) dx - \int_{\frac{z_0 - z}{\theta}}^\infty \left[B_2 \pi_i \theta e^{-\frac{A_1}{A_2} \alpha (z - z_0 + \theta x)} + o(\theta) \right] \varphi(x) dx.$$

Since functions $f_i(z, \theta)$ are bounded and the second moment of distribution $\varphi(\cdot)$ exists,

$$\begin{aligned} \int_{\frac{z_0 - z}{\theta}}^\infty f_i(z + \theta x, \theta) \varphi(x) dx &\leq \int_{\frac{z_0 - z}{\theta}}^\infty \varphi(x) dx \cdot \max_z f_i(z, \theta) \\ &\leq \frac{\theta^2}{(z_0 - z)^2} \int_{\frac{z_0 - z}{\theta}}^\infty x^2 \varphi(x) dx \cdot \max_z f_i(z, \theta) < o(\theta^2) \end{aligned}$$

as $\theta \rightarrow 0$. Analogously, the second terms in $R_i(z, \theta)$ tend to zero faster than θ^2 as $\theta \rightarrow 0$. Neglecting the last term in (30), analogously the first case, we obtain that for $z < z_0$

$$f_i(z, \theta) = D \pi_i e^{\frac{A_1}{A_2} (z - z_0)} + O(\theta)$$

and

$$P_i^0(s) = D \pi_i \theta e^{\frac{A_1}{A_2} \theta (s - S_0)} + o(\theta) \tag{31}$$

where D is a constant Since $C_0P_i^0(S_0) = C_1P_i^1(S_0)$, it follows that $D = B$. Taking into account the normalization condition (9) we get the stationary distribution of the stock level

$$P_i(s) = \begin{cases} \pi_i \frac{A_1\alpha}{A_2(1+\alpha)} \theta e^{-\frac{A_1}{A_2}\alpha\theta(s-S_0)} + o(\theta), & s \geq S_0, \\ \pi_i \frac{A_1\alpha}{A_2(1+\alpha)} \theta e^{-\frac{A_1}{A_2}\theta(S_0-s)} + o(\theta), & s < S_0. \end{cases} \tag{32}$$

3 Exact Stationary Distribution of the Stock Level for MMPP Two-State Demand with Exponential Batch Size’s Distribution

Here we compare the exact stationary distribution for MMPP two-state demand and an exponential batch size distribution of the purchases with the approximate one, which are obtained above. Thus, the amounts required at each arrival (batch sizes) are i.i.d. random variables with exponential distribution

$$\varphi(x) = \frac{1}{a} \exp\left(-\frac{x}{a}\right), \quad x \geq 0.$$

It can be shown that the exact distribution is defined by the following way

$$P_i(s) = \begin{cases} A_{i1}e^{-\alpha_1s} + A_{i2}e^{-\alpha_2s}, & s > S_0, \\ B_{i1}e^{-\beta_1s} + B_{i2}e^{-\beta_2s}, & s \leq S_0, \end{cases} \tag{33}$$

where $i = 1, 2$; α_1 and α_2 are positive roots of equation

$$f(z) = z \left(C_1 - \frac{\lambda_1 a}{1 + az} \right) \left(C_1 - \frac{\lambda_2 a}{1 + az} \right) + (q_{11} + q_{22}) \left(C_1 - \frac{\lambda_0 a}{1 + az} \right) = 0,$$

and β_1 and β_2 are positive roots of equation

$$f(z) = z \left(C_0 - \frac{\lambda_1 a}{1 - az} \right) \left(C_0 - \frac{\lambda_2 a}{1 - az} \right) - (q_{11} + q_{22}) \left(C_0 - \frac{\lambda_0 a}{1 - az} \right) = 0.$$

Constants A_{ij} and B_{ij} are defined by system

$$A_{2i} = -\frac{A_{1i}}{q_{21}} \left(q_{11} + C_1\alpha_i - \frac{\lambda_1 a\alpha_i}{1 + a\alpha_i} \right), B_{2i} = -\frac{B_{1i}}{q_{21}} \left(q_{11} + \frac{\lambda_1 a\beta_i}{1 - a\beta_i} - C_0\beta_i \right),$$

$$\sum_{j=1}^2 B_{ij} \frac{\exp\left(\frac{\beta_j a - 1}{a} S_0\right)}{\beta_j - 1} + \sum_{j=1}^2 A_{ij} \frac{\exp\left(-\frac{1 + a\alpha_j}{a} S_0\right)}{1 + a\alpha_j} = 0,$$

and normalization equations (9).

Figures 1 and 2 illustrate a quality of the approximation. Here the solid line corresponds to the exact stationary distribution derived from (33):

$$P(s) = \pi_1 P_1(s) + \pi_2 P_2(s),$$

and the dashed line corresponds to the approximate distribution (32).

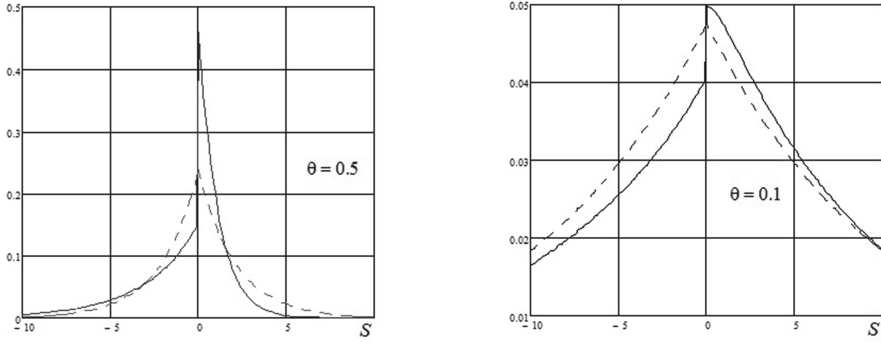


Fig. 1. Exact (solid line) and approximate (dashed line) PDF of the stock level for MMPP two-state demand with exponential batch size distribution; $\theta = 0.5$ and 0.1 , $\alpha = 1$, $q_{11} = q_{22} = -5$, $\lambda_1 = 15$, $\lambda_2 = 10$, $a = 1$, $S_0 = 0$.

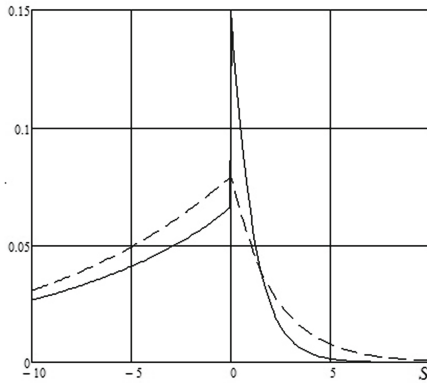


Fig. 2. Exact (solid line) and approximate (dashed line) PDF of the stock level for MMPP two-state demand with exponential batch size distribution; $\theta = 0.1$, $\alpha = 5$, $q_{11} = q_{22} = -5$, $\lambda_1 = 15$, $\lambda_2 = 10$, $a = 1$, $S_0 = 0$.

4 The Stationary Distribution of the Duration of an Overproduction Situation ($S > S_0$)

Let us find asymptotic conditional distribution $u(\cdot)$ of the length of time $t_i(s)$ when $S > S_0$ given that at the beginning of the period of overproduction $\lambda(t) = \lambda_i$ and $S = s(t)$. Denote conditional moment generating function of the period

$$H_i(u, s) = E \left\{ e^{-ut_i(s)} \right\} \tag{34}$$

Consider time interval $\Delta t \ll 1$. It follows

$$H_i(u, s) = e^{-u\Delta t} \left[(1 + q_{ii}\Delta t) E_{\Delta s} \{ H_i(u, s + \Delta s) \} + \sum_{j \neq i} q_{ij}\Delta t E_{\Delta s} \{ H_j(u, s + \Delta s) \} \right] + o(\Delta t). \tag{35}$$

Taking expectation with respect to Δs , tending Δt to zero, and taking into account that if the value of the purchase x is greater than the difference $s - S_0$ then the period of overproduction is over and, therefore, $H_i(u, s - x) = 1$, we obtain the following system of equations

$$\begin{aligned}
 -C_1 \frac{\partial H_i(u, s)}{\partial s} &= -(\lambda_i + u)H_i(u, s) + \sum_{j=1}^n q_{ij}H_j(u, s) \\
 &+ \lambda_i \int_0^{s-S_0} H_i(u, s-x)\varphi(x)dx + \lambda_i \int_{s-S_0}^{\infty} \varphi(x)dx
 \end{aligned}
 \tag{36}$$

To solve (36), consider again the case $C_1 = (1 - \alpha\theta)\lambda_0 a$, where $0 < \theta \ll 1$. Here and farther we will consider the case $\alpha = 1$. The solution will be found in the form

$$H_i(u, s) = f_i\left(\frac{u}{\theta^2}, \theta s, \theta\right),
 \tag{37}$$

where $f_i(\omega, z, \theta)$ are twice differentiable functions with respect to z and uniformly continuous with respect to ω and θ ; also $\lim_{\theta \rightarrow 0} \theta S_0(\theta) = z_0$.

Substitute (37) into (36) and denote $\omega = \frac{u}{\theta^2}$, $z = \theta s$, $z_0 = \theta S_0$. Rewrite (37) in the following form

$$\begin{aligned}
 C_1 \theta \frac{\partial f_i(\omega, z, \theta)}{\partial z} &= (\lambda_i + \omega \theta^2)f_i(\omega, z, \theta) - \sum_{j=1}^n q_{ij}f_j(\omega, z, \theta) \\
 &- \lambda_i \int_0^{\infty} f_i(\omega, z - \theta x, \theta)\varphi(x)dx + R_i(z, \theta)
 \end{aligned}
 \tag{38}$$

where

$$R_i(z, \theta) = \lambda_i \int_{\frac{z-z_0}{\theta}}^{\infty} f_i(\omega, z - \theta x, \theta)\varphi(x)dx + \lambda_i \int_{\frac{z-z_0}{\theta}}^{\infty} \varphi(x)dx$$

Analogously the previous, we get that $R_i(z, \theta) < o(\theta^2)$ as $\theta \rightarrow 0$ for $z > z_0$. So, we can neglect this term in (38).

From (38) we get as $\theta \rightarrow 0$

$$\sum_{i=1}^n q_{ij}f_j(\omega, z, 0) = 0.
 \tag{39}$$

Since $Rang Q = n - 1$ and $\sum_{j=1}^n q_{ij} = 0$, it follows that

$$f_j(\omega, z, 0) \equiv f(\omega, z),
 \tag{40}$$

where $f(\omega, z)$ is some function.

Let's

$$f_i(\omega, z, \theta) = f(\omega, z) + h_i(\omega, z)\theta + o(\theta)
 \tag{41}$$

as $\theta \rightarrow 0$.

From (38), we get

$$\sum_{j=1}^n q_{ij} h_j(\omega, z) = (\lambda_i - \lambda_0) a f_z(\omega, z). \tag{42}$$

Now consider

$$f_i(\omega, z, \theta) = f(\omega, z) + h_i(\omega, z)\theta + g_i(\omega, z)\theta^2 + o(\theta^2) \tag{43}$$

as $\theta \rightarrow 0$.

From (38), we have as $\theta \rightarrow 0$

$$\begin{aligned} \frac{\lambda_i a_2}{2} \ddot{f}_z(\omega, z) - \lambda_0 a_1 \dot{f}_z(\omega, z) - \omega f(\omega, z) \\ - (\lambda_i - \lambda_0) a_1 \dot{h}_{i,z}(\omega, z) = \sum_{j=1}^n q_{ij} g_j(\omega, z). \end{aligned} \tag{44}$$

Multiplying equations of system (44) to π_i and summing them up, we get

$$\frac{\lambda_0 a_2}{2} \ddot{f}_z(\omega, z) - \lambda_0 a_1 \dot{f}_z(\omega, z) - \omega f(\omega, z) = \sum_{i=1}^n (\lambda_i - \lambda_0) \pi_i a_1 \dot{h}_{i,z}(\omega, z). \tag{45}$$

From (42), it follows

$$\sum_{j=1}^n q_{ij} \dot{h}_j(\omega, z) = (\lambda_i - \lambda_0) a_1 \ddot{f}_z(\omega, z)$$

or taking into account that $Q = P^T \gamma R^T$,

$$\gamma_t \sum_{j=1}^n R_{jt} \dot{h}_{j,z}(\omega, z) = \sum_{i=1}^n R_{it} (\lambda_i - \lambda_0) a_1 \ddot{f}_z(\omega, z).$$

Note, that the last equation of the above system is satisfied by any sum $\sum_{i=1}^n R_{in} \dot{h}_{i,z}(\omega, z)$ because $R_{jn} = \pi_j$ and $\gamma_n = 0$. Thus, we obtain the system of equations

$$\begin{aligned} \sum_{j=1}^n R_{jt} \dot{h}_{j,z}(\omega, z) = \frac{1}{\gamma_t} \sum_{i=1}^n R_{it} (\lambda_i - \lambda_0) a_1 \ddot{f}_z(\omega, z) \quad t = \overline{1, n-1}, \\ \sum_{i=1}^n R_{in} \dot{h}_{i,z}(\omega, z) = c(z), \end{aligned}$$

where $c(\cdot)$ is any function.

Since $P_{nk} = 1$, it follows from the above that

$$\dot{h}_{k,z}(\omega, z) = \sum_{i=1}^{n-1} P_{ik} \frac{1}{\gamma_i} \sum_{j=1}^n R_{ji} (\lambda_j - \lambda_0) a_1 \ddot{f}_z(\omega, z) + c(z),$$

and

$$\sum_{i=1}^n (\lambda_i - \lambda_0) \pi_i a \dot{h}_{i,z}(\omega, z) = a_1^2 \sum_{k=1}^n (\lambda_k - \lambda_0) \pi_k \sum_{i=1}^{n-1} \frac{P_{ik}}{\gamma_i} \sum_{j=1}^n R_{ji} (\lambda_j - \lambda_0) \ddot{f}_z \omega, z).$$

Thus, function $f(z, \omega)$ satisfies equation

$$A_2 \ddot{f}_z(\omega, z) - A_1 \dot{f}_z(\omega, z) - \omega f(\omega, z) = 0, \tag{46}$$

where A_1 and A_2 are defined by (26) and (27). Roots of the characteristic equation for (46) are

$$t_1(\omega) = \frac{A_1 + \sqrt{A_1^2 + 4A_2\omega}}{2A_2}, \quad t_2(\omega) = \frac{A_1 - \sqrt{A_1^2 + 4A_2\omega}}{2A_2}, \tag{47}$$

therefore,

$$f(\omega, z) = D_1(\omega) e^{t_1(\omega)(z-z_0)} + D(\omega) e^{t_2(\omega)(z-z_0)}.$$

Since $|H_i(u, s)| \leq 1$, it follows that $|f(\omega, z)| \leq 1$ for $z \geq z_0$. Therefore, $D_1(\omega) = 0$ and

$$f(\omega, z) = D(\omega) e^{t_2(\omega)(z-z_0)}. \tag{48}$$

Let's find $D(\omega)$. For $z = z_0$, it follows from (36) that

$$C_1 \theta \frac{\partial f_i(\omega, 0, \theta)}{\partial z} = (\lambda_i + \omega \theta^2) f_i(\omega, 0, \theta) - \sum_{j=1}^n q_{ij} f_j(\omega, 0, \theta) + \lambda_i = 0. \tag{49}$$

From the above we get that $D(\omega) = 1$ as $\theta \rightarrow 0$ and

$$H_i(\omega, s) = e^{t_2(\frac{\omega}{\theta^2})\theta(s-S_0)} + O(\theta). \tag{50}$$

For $s > S_0$, stock level S is distributed according the following PDF

$$P(s | S > S_0) = \frac{P(s)}{P\{S > S_0\}} = \frac{\theta A_1}{A_2} e^{-\frac{\theta A_1}{A_2}(s - S_0)}$$

Taking expectation with respect to the Markov chain's state and the stock level, we obtain unconditional moment generating function

$$H(\omega) = \sum_{i=1}^n \pi_i \int_{S_0}^{\infty} H_i(\omega, s) P(s | S > S_0) ds = \frac{\beta}{1 + \sqrt{1 + \beta\omega}}, \tag{51}$$

where

$$\beta = \frac{4A_2}{\theta^2 A_1^2} \tag{52}$$

Finding the inverse Laplace transform (Bateman and Erdely [22]), from the above we get PDF of the period of overproduction

$$u(t) = \frac{2}{\sqrt{\pi\beta t}} e^{-\frac{t}{\beta}} - \frac{2}{\beta} \text{Erfc} \left(\sqrt{\frac{t}{\beta}} \right). \tag{53}$$

5 The Stationary Distribution of the Duration of a Stock-Out Situation ($S(t) < 0$)

Let us find asymptotic conditional distribution $v(\cdot)$ of the length of time $\tau_i(s)$ when $S(t) < 0$ (the stock-out period) given that at the beginning of the period $\lambda(t) = \lambda_i$ and $S = s(t) < 0$. Denote conditional moment generating functions of the period

$$\Psi_i(u, s) = E \left\{ e^{-u\tau_i(s)} \right\}. \tag{54}$$

Analogously the previous part we obtain the system of equations with respect to $\Psi_i(\cdot, \cdot)$

$$C_0 \frac{\partial \Psi_i(u, s)}{\partial s} = (\lambda_i + u)\Psi_i(u, s) - \sum_{j=1}^n q_{ij}\Psi_j(u, s) - \lambda_i \int_0^\infty \Psi_i(u, s - x)\varphi(x)dx. \tag{55}$$

To solve (55), consider the case $C_0 = (1 + \theta)\lambda_0 a$, where $0 < \theta \ll 1$, and the following form of the solution

$$\Psi_i(u, s) = f_i \left(\frac{u}{\theta^2}, \theta s, \theta \right). \tag{56}$$

Analogously the previous, we obtain that $f_j(\omega, z, 0) \equiv f(\omega, z)$ satisfies the following equation

$$A_2 \ddot{f}_z(\omega, z) + A_1 \dot{f}_z(\omega, z) - \omega f(\omega, z) = 0 \tag{57}$$

where A_1 and A_2 are defined by (26) and (27).

The solution of (57)

$$f(\omega, z) = D_1(\omega)e^{-t_1(\omega)z} + D_2(\omega)e^{-t_2(\omega)z},$$

where $t_1(\omega)$ and $t_2(\omega)$ are defined by (47). Since $|f(\omega, z)| \leq 1$ for $z < 0$, it follows $D_1(\omega) = 0$. Thus

$$f(\omega, z) = D_2(\omega)e^{-t_2(\omega)z}. \tag{58}$$

From $\Psi_i(\omega, 0) = 1$ it follows that $f(\omega, 0) = 1$, therefore, $D_2(\omega) = 1$. Thus,

$$\Psi_i(\omega, s) = e^{-t_2(\frac{\omega}{\theta^2})\theta s} + O(\theta). \tag{59}$$

Giving $S < 0$ stock level S is distributed according the following PDF

$$P(s | S < 0) = \frac{P(s)}{P\{S < 0\}} = \frac{\theta A_1}{A_2} e^{\frac{A_1 \theta}{A_2} s}.$$

Taking expectation with respect to the Markov chain's state and the stock level we receive unconditional moment generating function

$$\Psi(\omega) = \sum_{i=1}^n \pi_i \int_{-\infty}^0 \Psi_i(\omega, s)P(s | S < 0)ds = \frac{\beta}{1 + \sqrt{1 + \beta\omega}},$$

where β is defined by (52). Finding the inverse Laplace transform (Bateman and Erdely [22]), we get from the above that PDF of the period of stock-out

$$v(t) = \frac{2}{\sqrt{\pi\beta t}} e^{-\frac{t}{\beta}} - \frac{2}{\beta} \text{Erfc} \left(\sqrt{\frac{t}{\beta}} \right). \quad (60)$$

Thus, the asymptotic distributions of the durations of the periods of overproduction and stock out coincide. That is quite natural because of the symmetric changes of the production rate for $\alpha = 1$ when the threshold is crossed.

6 Conclusion

Thus, the main part of the stationary distribution of the stock level of the production-inventory system under consideration has been found. We assume that the difference between the rate of production and mean sale per time unit is a small value θ , which is positive, if the stock is below the base stock level S_0 and negative, otherwise. Also we assume that $S_0 \gg 1$ as $\theta \ll 1$ so that θS_0 keeps a constant value. Under the same assumptions, the approximate distributions of the durations of the period of overproduction and the period of unmet demand were obtained.

The similar approximate method can be used for analysis of more complex systems, for example, to simultaneously account for the dependence of the intensity of the customer's flow on the retail price. Other challenging problems are maximization the approximate expected profit, investigation its sensitivity to the parameters of the model, and giving recommendations to practitioners.

References

1. Arrow, K.J., Harris, Th.E., Marschak, J.: Optimal inventory policy. *Econometrica* **19**(3), 205–272 (1951)
2. Dvoretzky, A., Kiefer, J., Wolfowitz, J.: On the optimal character of the (S, s) policy in inventory theory. *Econometrica* **21**, 586–596 (1953)
3. Ross, Sh.M.: *Applied Probability Models with Optimization Applications*, 224 p. Dover Publications, New York (1992)
4. Chopra, S., Meindl, P.: *Supply Chain Management: Strategy, Planning and Operation*, 529 p. Pearson Education, New Jersey (2013)
5. Beyer, D., Cheng, F., Sethi, S.P., Taksar, M.: *Markovian Demand Inventory Models*, 255 p. Springer, New York (2010). <https://doi.org/10.1007/978-0-387-71604-6>
6. Nazarov, A., Broner, V.: Inventory management system with on/off control of input product flow. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) *ITMM 2017*. CCIS, vol. 800, pp. 370–381. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_30
7. Abhyankar, H.S., Graves, S.C.: Creating an inventory hedge for Markov-modulated poisson demand. An application and model. *Manuf. Serv. Oper. Manag.* **3**(4), 306–320 (2001)

8. Fischer, W., Meier-Hellstern, K.: The Markov-modulated Poisson process (MMPP) cookbook. *Perform. Eval.* **18**, 149–171 (1993)
9. Janssen, J., Manca, O., Manca, R.: *Applied Diffusion Processes from Engineering to Finance*, 394 p. Wiley, London (2013)
10. Bather, J.A.: A continuous time inventory model. *J. Appl. Probab.* **3**, 538–549 (1966)
11. Puterman, M.: A diffusion process model for a storage system. In: Geisler, M.A. (ed.) *Studies in the Management Sciences, Logistics, I*, pp. 143–159. North-Holland Press, Amsterdam (1975)
12. Rao, U.: Properties of the periodic review (R, T) inventory control policy for stationary, stochastic demand. *Manuf. Serv. Oper. Manag.* **5**(1), 37–53 (2003)
13. Rudi, N., Groenevelt, H., Randall, T.R.: End-of-period vs. continuous accounting of inventory-related costs. *Oper. Res.* **57**(6), 1360–1366 (2009)
14. Avinadav, T.: Continuous accounting of inventory costs with Brownian-motion and Poisson demand processes. *Ann. Oper. Res.* **229**, 85–102 (2015)
15. Benkherouf, L., Boumenir, A., Aggoun, L.: A diffusion inventory model for deteriorating items. *Appl. Math. Comput.* **138**, 21–39 (2003)
16. Livshits, K., Ulyanova, E.: Diffusion approximation of the production and selling of perishable products. *Russ. Phys. J.* **58**(11/2), 281–285 (2015)
17. Livshits, K., Ulyanova, E.: Switch-hysteresis control of the production process in a model with perishable goods. *Commun. Comput. Inf. Sci.* **638**, 192–206 (2015)
18. Kitaeva, A.V.: Stabilization of inventory system performance: on/off control. In: *The 19th World Congress of the International Federation of Automatic Control, IFAC Proceedings Volumes (IFAC-PapersOnline)*, vol. 19, pp. 10748–10753 (2014)
19. Kitaeva, A., Subbotina, V., Zmeev, O.: Diffusion approximation in inventory management with examples of application. *Commun. Comput. Inf. Sci.* **487**, 189–196 (2014)
20. Livshits, K.I., Bublic, Ya.S.: Ruin probability of an insurance company under double stochastic payment current. *J. Control Comput. Sci.* **1**(10), 66–77 (2010). Tomsk State University (in Russian)
21. Livshits, K.I., Bublic, Ya.S.: Ruin probability of an insurance company under double stochastic flows of insurance premium and insurance payments. *J. Control Comput. Sci.* **4**(17), 64–73 (2011). Tomsk State University (in Russian)
22. Bateman, H., Erdelyi, A.: *Tables of Integral Transforms*, vol. 1, 391 p. McGraw-Hill, New York (1954)



System State Distribution of a Finite-Source Retrieval Queue with Subscribed Customers

Velika Dragieva^(✉)

University of Forestry, 10 Kliment Ohridsky, 1756 Sofia, Bulgaria
dragievav@yahoo.com

Abstract. The paper deals with a single server, finite-source retrieval queue where the server serves two types of customers, called regular customers and subscribed customers. The service times of both types customers follow two distinct arbitrary probability distributions. In addition, the subscribed customers do not join the orbit of repeated regular customers if the server is busy at the time of their arrival. Instead, such an unsuccessful subscribed customer waits till the current regular service is over, and then is accepted for service. Using the supplementary variable approach and the discrete transformations technique we derive formulas for computing the stationary joint distribution of the server state and the orbit size.

Keywords: Finite queue · Retrials · Subscribed customers

1 Introduction

Queueing systems with finite source and retrials combine two features of great practical importance. In the queueing models with finite source (also called closed queueing models) it is assumed that the server/servers service a finite number of customers as it is in most of the real situations. Each of these customers produces its one flow of demands which means that the generalized input flow depends on the number of customers able to produce demands, i.e. the customers not being under service or not waiting for service. These models have been used to analyze the performance of telephone, computer, communication and other systems [4, 5, 7, 15, 16]. The characteristic feature of queueing systems with retrials concerns the behavior of those unsuccessful demands whose service cannot start at the moment of their arrival. In the models with retrials it is assumed that these customers are not lost or allowed to queue. Instead, they repeat their attempts for service until find the server idle. Between trials the customers are said to be in the orbit, or to be sources of repeated (secondary) calls, secondary subscribers. Retrial queues arise in diverse real situations including our daily activity, telephone switching systems, telecommunication and computer networks, call centers, cellular and local area networks, etc. [1, 3, 6, 8, 20, 21] A systematic

account of the fundamental methods and the latest results, as well as an classified bibliography on this topic can be found, for example in [3, 11, 12, 17], and references therein. Single server retrial queue with a finite number of customers has been studied in a number of articles by a number of authors: Ohmura and Takahashi [19], Falin and Artalejo [13], Amador [2], Dragieva [9] and others. Recently, such models, extended with different additional features of the service have been extensively studied. This includes, service with an unreliable server [22, 24], service with two phases of the service times [23], service with collisions [18], service with random access [14], service with two-way communication [10], etc. To the best of our knowledge there are no investigations about a retrial queue with one server that serves a finite number of customers, some of which have a special status and are called subscribed or special customers. The motivation for studying such model are many real situations like call centers, repair centers, or medical centers. Usually in these centers along with the regular customers there is a special group of subscribed customers, or customers (patients) under special care whose service consists mainly of preventive activities, initiated by the server (operator) when being idle. In addition, we assume in our model that if a special customer arrives and finds the server busy with a regular customer service, the special customer waits till the end of the current service and then is immediately accepted for service.

Further the paper is organized as follows. In Sect. 2 we describe the model in detail. Section 3 contains the main results of the paper, namely formulas for calculation of the stationary joint distribution of the server state and the orbit size. Section 4 closes the paper and presents some possible further investigations.

2 Model Description and Notations

We consider in this article a queueing model with one server (an operator, or a company) that serves customers of two types - regular customers, and subscribed customers that also will be called special customers. The numbers of both types customers are fixed - K regular and $(N - K)$ subscribers, $K < N$. Each of these customers produces a Poisson flow of demands with intensity λ_1 and λ_2 , respectively.

At any time t the server can be in one of three possible states - idle, busy with service of a regular customer (regular service) or busy with a special customer (special service). This will be indicated by the variable $C(t)$, equal to 0, 1 or 2, respectively.

If the server is idle at the time of a regular customer arrival, the customer starts to be served. Otherwise it enters a virtual waiting room, called orbit and after an exponentially distributed interval repeats its attempt for service. These attempts are repeated until the customer finds the server idle. Thus, each regular customer in the orbit produces a Poisson flow of demands with intensity μ . The customers in the orbit are called secondary or repeated customers, while those that are outside it - primary regular customers or regular customers in free state. The service duration of primary and secondary regular customers follows

the same arbitrary law with common probability distribution function $B_1(x)$, hazard rate function $b_1(x) = B_1'(x)[1 - B_1(x)]^{-1}$, Laplace-Stieltjes transform $\beta_1(s)$ and mean $1/\nu_1$. After the service is over the regular customers of both types (primary or secondary) move to a free state, i.e. can produce a Poisson flow of demands with intensity λ_1 .

The behaviour of the subscribed customers is as follows. If the server is idle at the time t of a subscribed customer arrival, $C(t) = 0$, the service of this customer starts. If $C(t) = 1$, i.e. if the server is busy with a regular service, then the subscribed customer waits till the current regular service is over and then is accepted immediately for service. We assume that no more than one special customer is allowed to wait for the next service, i.e. if at the time moment of a special customer arrival, $C(t) = 1$ and if one special customer is waiting, the system state does not change. Finally, if the server is busy with a subscribed service, and a subscribed customer arrives, the system state does not change, i.e. we assume that the rejected subscribed customers do not join the orbit. The service duration of subscribers follows an arbitrary law with common probability distribution function $B_2(x)$, hazard rate function $b_2(x)$, Laplace-Stieltjes transform $\beta_2(s)$ and mean $1/\nu_2$. After the service any subscribed customer is free to produce his/her usual demands that form a Poisson flow with intensity λ_2' .

Thus, the flow of subscribed customers demands can be either with intensity $(N - K)\lambda_2'$ (when the server is idle or serving a regular customer and no subscriber is waiting for the next service) or $(N - K - 1)\lambda_2'$ (when a subscribed customer is under service or waiting for the next service). Since the last flow does not change the system state it is convenient to accept that the subscribed customers arrive in the system according to a Poisson flow with intensity $\lambda_2 = (N - K)\lambda_2'$.

Introducing a supplementary variable $z(t)$, equal to the elapsed service time, the state of the system at time t can be described by the Markov process

$$X(t) = \{C(t), R(t), z(t)\}$$

where $C(t)$ denotes the server state at time t and $R(t)$ is the number of repeated regular customers at time t .

3 Steady State Analysis

Because of the finite state space of the Markov process X the stationary regime exists and we can define the limiting probabilities (densities)

$$p_{i,j}(x)dx = \lim_{t \rightarrow \infty} P \{C(t) = i, R(t) = j, x \leq z(t) < x + dx\}, i = 1, 2,$$

$$p_{i,j} = \lim_{t \rightarrow \infty} P \{C(t) = i, R(t) = j\}, i = 0, 1, 2, j = 0, 1, \dots, K.$$

In a general way we obtain the equations of statistical equilibrium

$$p_{0,n} [(K - n) \lambda_1 + \lambda_2 + n\mu]$$

$$= (1 - \delta_{n,K}) \int_0^\infty p_{1,n}(x)b_1(x)e^{-\lambda_2x}dx + \int_0^\infty p_{2,n}(x)b_2(x)dx, \tag{1}$$

$$\frac{dp_{1,n}(x)}{dx} = -[(K - n - 1)\lambda_1 + b_1(x)]p_{1,n}(x) + (K - n)\lambda_1p_{1,n-1}(x), \tag{2}$$

$$\frac{dp_{2,n}(x)}{dx} = -[(K - n)\lambda_1 + b_2(x)]p_{2,n}(x) + (K - n + 1)\lambda_1p_{2,n-1}(x), \tag{3}$$

$$p_{1,n}(0) = (K - n)\lambda_1p_{0,n} + (1 - \delta_{n,K})(n + 1)\mu p_{0,n+1}, \tag{4}$$

$$p_{2,n}(0) = \lambda_2p_{0,n} + (1 - \delta_{n,K}) \int_0^\infty p_{1,n}(x)b_1(x)(1 - e^{-\lambda_2x})dx, \tag{5}$$

$n = 0, 1, \dots, K$, with $p_{1,-1}(x) = p_{2,-1}(x) = p_{1,K}(x) = p_{0,-1} = 0$. Here $e^{-\lambda_2x}$ is the probability that during a time interval x no subscribed call arrives, $(1 - e^{-\lambda_2x})$ is the probability that during a time interval x at least one subscribed call arrives. To solve Eqs. (2) and (3) we apply the discrete transformations method, common in the investigation of finite queues [9, 13, 16, 19, 22–24]. To this end we write the equations in a matrix form,

$$[\theta_i I_i - A_i] \bar{p}_i(x) = 0,$$

where θ_i are given as,

$$\theta_i = b_i(x) + \frac{d}{dx},$$

I_i is the identity matrix of order $K - 2 + i$, A_i are constructed from (2) and (3), respectively in the usual way and $\bar{p}_i(x)$, ($i = 1, 2$) is the column vector of the unknown functions $p_{i,j}(x)$,

$$\bar{p}_i(x) = (p_{i0}(x), \dots, p_{i,K-2+i}(x))^T.$$

Then we find the matrices Y_i and Λ_i , such that $Y_i^{-1}A_iY_i = \Lambda_i$. Thus, applying the transformation $\bar{p}_i(x) = Y_i\bar{q}_i(x)$, in the equation $[\theta_i I - A_i] \bar{p}_i(x) = 0$ we obtain it in the form $\theta_i q_i(x) = \Lambda_i q_i(x)$ which is easy to solve.

The matrices Y_1 and Λ_1 , which simplify Eq. (2) are well known in the theory of finite source queues [13, 16, 19, 23]. Λ_1 is a diagonal one, $\Lambda_1 = \text{diag}\{0, -\lambda_1, \dots, -(K - 1)\lambda_1\}$ where the diagonal elements are exactly the eigenvalues of A_1 with corresponding eigenvectors

$$\bar{y}_1^{(k)} = \left(y_{1,0}^{(k)}, \dots, y_{1,K-1}^{(k)} \right),$$

where

$$y_{1,n}^{(k)} = \begin{cases} (-1)^{k-(K-n-1)} \binom{k}{K-n-1}, & \text{for } k + n \geq K - 1, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The matrix A_2 is of the same type as A_1 , but of order $K + 1$, while A_1 is of order K . This means that $A_2 = \text{diag}\{0, -\lambda_1, \dots, -K\lambda_1\}$ and the corresponding eigenvectors

$$\bar{y}_2^{(k)} = \left(\bar{y}_{2,0}^{(k)}, \dots, \bar{y}_{2,K}^{(k)} \right)$$

with entries

$$y_{2,n}^{(k)} = \begin{cases} (-1)^{k-(K-n)} \binom{k}{K-n}, & k+n \geq K, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Thus, with the help of transformations

$$\begin{aligned} p_{1,n}(x) &= \sum_{k=0}^{K-1} y_{1,n}^{(k)} q_{1,k}(x) \\ &= \sum_{m=0}^n (-1)^m \binom{K-1-(n-m)}{m} q_{1,K-1-(n-m)}(x), \end{aligned} \tag{8}$$

$n = 0, \dots, K-1,$

$$\begin{aligned} p_{2,n}(x) &= \sum_{k=0}^K y_{2,n}^{(k)} q_{2,k}(x) = \sum_{m=0}^n (-1)^m \binom{K-(n-m)}{m} q_{2,K-(n-m)}(x), \end{aligned} \tag{9}$$

$n = 0, \dots, K,$

we can solve (2) and (3), and then, using (1), (4) and (5) we can derive formulas for the densities $p_{i,n}(x)$, $i = 1, 2$ and the corresponding probabilities $p_{i,n}$ and $p_{0,n}$, $n = 0, \dots, K$. This is obtained in the following proposition.

Proposition 1. *The stationary joint distribution of the server state and the orbit size can be calculated according to the following formulas:*

$$\begin{aligned} p_{i,n}(x) &= \\ & \sum_{k=K-n-2+i}^{K-2+i} (-1)^{k-(K-n-2+i)} \binom{k}{K-n-2+i} \bar{q}_{i,k} (1 - B_i(x)) e^{-k\lambda_i x}, \end{aligned} \tag{10}$$

$$\begin{aligned} p_{i,n} &= \sum_{k=K-n-2+i}^{K-2+i} (-1)^{k-(K-n-2+i)} \binom{k}{K-n-2+i} \bar{q}_{i,k} \frac{1 - \bar{\beta}_{i,k}}{k\lambda_i}, \end{aligned} \tag{11}$$

$i = 1, 2,$

$$\begin{aligned} ((K-n)\lambda_1 + \lambda_2 + n\mu) p_{0,n} &= \sum_{k=K-n}^K (-1)^{k-(K-n)} \binom{k}{K-n} \bar{q}_{2,k} \bar{\beta}_{2,k} + \\ & (1 - \delta_{n,K}) \sum_{k=K-n-1}^{K-1} (-1)^{k-(K-n-1)} \binom{k}{K-n-1} \bar{q}_{1,k} \bar{\beta}_{12,k}, \end{aligned} \tag{12}$$

where

$$\begin{aligned} \bar{\beta}_{i,k} &= \beta_i (k\lambda_i) \quad (i = 1, 2), \quad \bar{\beta}_{12,k} = \beta_1 (k\lambda_1 + \lambda_2), \\ \frac{1 - \bar{\beta}_{i,k}}{k\lambda_i} &= \frac{1}{\nu_i} \text{ for } k = 0. \end{aligned} \tag{13}$$

The quantities $\bar{q}_{i,k}$ are connected by the linear equations

$$\sum_{k=K-n-1}^{K-1} \bar{q}_{1,k} A_{n,k} + \bar{q}_{1,K-n-2} (n+1) \bar{\gamma}_{K-n-2} \mu = \sum_{k=K-n-1}^K \bar{q}_{2,k} B_{n,k}, \tag{14}$$

$$\begin{aligned}
 n &= 0, \dots, K - 1, \\
 \sum_{k=K-n-1}^{K-1} \bar{q}_{1,k} C_{n,k} &= \sum_{k=K-n}^K \bar{q}_{2,k} D_{n,k}, \\
 n &= 0, \dots, K,
 \end{aligned} \tag{15}$$

where

$$\begin{aligned}
 A_{n,k} &= (-1)^{k-(K-n-1)} \left\{ \binom{k}{K-n-1} [\lambda_2 + \bar{\gamma}_k (K-n) \lambda_1] \right. \\
 &\quad \left. - (1 - \delta_{n,K-1}) (n+1) \bar{\gamma}_k \mu \binom{k}{K-n-2} \right\}, \\
 B_{n,k} &= (-1)^{k-(K-n)} \binom{k}{K-n-1} [\lambda_1 (k+n+1-K) - (n+1) \mu], \\
 C_{n,k} &= (-1)^{k-(K-n-1)} \binom{k}{K-n-1} \{ [(K-n) \lambda_1 + n\mu] \bar{\gamma}_k + \lambda_2 \bar{\beta}_{12,k} \}, \\
 D_{n,k} &= (-1)^{k-(K-n)} \binom{k}{K-n} [(K-n) \lambda_1 + n\mu + \lambda_2 (1 - \bar{\beta}_{2,k+1})],
 \end{aligned}$$

with

$$\bar{\gamma}_k = \bar{\beta}_{1,k} - \bar{\beta}_{12,k},$$

$$\bar{q}_{1,-1} = 0.$$

Proof. As stated above, applying in (2) and (3) transformations (8) and (9), respectively we get them in the simpler form

$$\begin{aligned}
 \frac{dq_{1,m}(x)}{dx} &= -[m\lambda_1 + b_1(x)] q_{1,m}(x), \quad m = 0, \dots, K - 1, \\
 \frac{dq_{2,m}(x)}{dx} &= -[m\lambda_2 + b_2(x)] q_{2,m}(x), \quad m = 0, \dots, K,
 \end{aligned}$$

with solutions

$$\begin{aligned}
 q_{1,m}(x) &= q_{1,m}(0) (1 - B_1(x)) \exp \{-m\lambda_1 x\}, \\
 q_{2,m}(x) &= q_{2,m}(0) (1 - B_2(x)) \exp \{-m\lambda_2 x\}.
 \end{aligned}$$

This leads to the following expressions for the functions $p_{i,n}(x)$, ($i = 1, 2$, $n = 0, \dots, K - 2 + i$)

$$p_{i,n}(x) = \sum_{k=0}^{K-2+i} y_{i,n}^{(k)} \bar{q}_{i,k} (1 - B_i(x)) e^{-k\lambda_i x}, \tag{16}$$

where $\bar{q}_{i,k} = q_{i,k}(0)$. Substituting according to these equations in (1) we express $p_{0,n}$ in terms of the quantities $\bar{q}_{i,k}$

$$p_{0,n} = \frac{\sum_{k=0}^{K-1} \left(y_{1,n}^{(k)} \bar{q}_{1,k} \bar{\beta}_{12,k} + y_{2,n}^{(k)} \bar{q}_{2,k} \bar{\beta}_{2,k} \right) + y_{2,n}^{(K)} \bar{q}_{2,K} \bar{\beta}_{2,K}}{(K-n) \lambda_1 + \lambda_2 + n\mu}, \tag{17}$$

where $\bar{\beta}_{12,k}, \bar{\beta}_{i,k}$ ($i = 1, 2, k = 0, \dots, K$) are determined by (13). Formulas (16) and (17), compared with formulas (7) and (6) for $y_{i,n}^{(k)}$ ($i = 1, 2, n+k \leq K-2+i$) prove (10) and (12). Formulas (11) follow from the relations

$$p_{i,n} = \int_0^\infty p_{i,n}(x)dx, \quad i = 1, 2.$$

Now, using the initial conditions (4) and (5) we can derive a system of linear equations for $\bar{q}_{i,k}$. Substituting in (4) and (5) $p_{i,n}(x)$ according to (16) we obtain

$$\begin{aligned} \sum_{k=0}^{K-1} y_{1,n}^{(k)} \bar{q}_{1,k} &= (K-n) \lambda_1 p_{0,n} + (n+1) \mu p_{0,n+1}, \\ n &= 0, \dots, K-1, \\ \sum_{k=0}^K y_{2,n}^{(k)} \bar{q}_{2,k} &= \lambda_2 p_{0,n} + (1 - \delta_{n,K}) \sum_{k=0}^{K-1} y_{1,n}^{(k)} \bar{q}_{1,k} (\bar{\beta}_{1,k} - \bar{\beta}_{12,k}), \\ n &= 0, 1, \dots, K. \end{aligned} \tag{18}$$

Excluding $p_{0,n}$ we get to the following equations

$$\begin{aligned} \lambda_2 \sum_{k=0}^{K-1} y_{1,n}^{(k)} \bar{q}_{1,k} &= (K-n) \lambda_1 \left[\sum_{k=0}^K y_{2,n}^{(k)} \bar{q}_{2,k} - \sum_{k=0}^{K-1} y_{1,n}^{(k)} \bar{q}_{1,k} (\bar{\beta}_{1,k} - \bar{\beta}_{12,k}) \right] \\ + (n+1) \mu &\left[\sum_{k=0}^K y_{2,n+1}^{(k)} \bar{q}_{2,k} - (1 - \delta_{n,K-1}) \sum_{k=0}^{K-1} y_{1,n+1}^{(k)} \bar{q}_{1,k} (\bar{\beta}_{1,k} - \bar{\beta}_{12,k}) \right], \end{aligned}$$

which can be also written in the form

$$\begin{aligned} &\sum_{k=0}^{K-1} \bar{q}_{1,k} \left\{ y_{1,n}^{(k)} [\lambda_2 + (K-n) \lambda_1 (\bar{\beta}_{1,k} - \bar{\beta}_{12,k})] \right. \\ &\left. + (1 - \delta_{n,K-1}) (n+1) \mu (\bar{\beta}_{1,k} - \bar{\beta}_{12,k}) y_{1,n+1}^{(k)} \right\} \\ &= \sum_{k=0}^K \left[(K-n) \lambda_1 y_{2,n}^{(k)} + (n+1) \mu y_{2,n+1}^{(k)} \right] \bar{q}_{2,k}, \\ n &= 0, \dots, K-1. \end{aligned}$$

Substituting here $y_{i,n}^{(k)}$ according to (7) and (6) we obtain the equations

$$\begin{aligned} \sum_{k=K-n-1}^{K-1} \bar{q}_{1,k} &\left\{ (-1)^{k-(K-n-1)} \binom{k}{K-n-1} [\lambda_2 + (K-n) \lambda_1 (\bar{\beta}_{1,k} - \bar{\beta}_{12,k})] \right. \\ &\left. + (1 - \delta_{n,K-1}) (n+1) \mu (\bar{\beta}_{1,k} - \bar{\beta}_{12,k}) (-1)^{k-(K-n-2)} \binom{k}{K-n-2} \right\} \\ &+ \bar{q}_{1,K-n-2} (n+1) \mu (\bar{\beta}_{1,K-n-2} - \bar{\beta}_{12,K-n-2}) = (n+1) \mu \bar{q}_{2,K-n-1} \end{aligned}$$

$$\begin{aligned}
 &+ \sum_{k=K-n}^K \left[(K-n) \lambda_1 (-1)^{k-(K-n)} \binom{k}{K-n} \right. \\
 &+ \left. (n+1) \mu (-1)^{k-(K-n-1)} \binom{k}{K-n-1} \right] \bar{q}_{2,k}, \\
 &n = 0, \dots, K-1,
 \end{aligned}$$

which after some transformations prove formulas (14).

Another relations between these quantities come from Eqs. (17) and (18):

$$\begin{aligned}
 &[(K-n) \lambda_1 + \lambda_2 + n\mu] \left\{ \sum_{k=0}^{K-1} \left[y_{2,n}^{(k)} \bar{q}_{2,k} - y_{1,n}^{(k)} \bar{q}_{1,k} (\bar{\beta}_{1,k} - \bar{\beta}_{12,k}) \right] + y_{2,n}^{(K)} \bar{q}_{2,K} \right\} \\
 &= \lambda_2 \left[\sum_{k=0}^{K-1} \left(y_{1,n}^{(k)} \bar{q}_{1,k} \bar{\beta}_{12,k} + y_{2,n}^{(k)} \bar{q}_{2,k} \bar{\beta}_{2,k} \right) + y_{2,n}^{(K)} \bar{q}_{2,K} \bar{\beta}_{2,K} \right],
 \end{aligned}$$

which can be simplified to the form

$$\begin{aligned}
 &\sum_{k=0}^K y_{2,n}^{(k)} \bar{q}_{2,k} [(K-n) \lambda_1 + n\mu + \lambda_2 (1 - \bar{\beta}_{2,k+1})] \\
 &= \sum_{k=0}^{K-1} y_{1,n}^{(k)} \bar{q}_{1,k} \{ [(K-n) \lambda_1 + n\mu + \lambda_2] \bar{\beta}_{1,k} - [(K-n) \lambda_1 + n\mu] \bar{\beta}_{12,k} \}.
 \end{aligned}$$

Here we apply formulas (7) and (6) for $y_{i,n}^{(k)}$ and obtain formulas (15). This finishes the proof of the proposition.

Proposition 1 provides convenient formulas for calculation of the probabilities $p_{i,n}$ ($i = 0, 1, 2, n = 0, \dots, K$). We can see that all variables $\bar{q}_{i,k}$ ($i = 1, 2$) (consequently and the probabilities $p_{i,n}$) are proportional to $\bar{q}_{2,K}$. Indeed, from Eq. (15) for $n = 0$ we can express $\bar{q}_{1,K-1}$ in terms of $\bar{q}_{2,K}$,

$$\bar{q}_{1,K-1} = \bar{q}_{2,K} \frac{D_{0,K}}{C_{0,K-1}}.$$

Then, from (14) for $n = 0$ we express $\bar{q}_{1,K-2}$ in terms of $\bar{q}_{1,K-1}$ (and consequently of $\bar{q}_{2,K}$) and of $\bar{q}_{2,K-1}$,

$$\bar{q}_{1,K-2} \bar{\gamma}_{K-2} \mu = \bar{q}_{2,K-1} B_{0,K-1} + \bar{q}_{2,K} B_{0,K} - \bar{q}_{1,K-1} A_{0,K-1},$$

and, substituting with this expression in (15) for $n = 1$ we can express $\bar{q}_{2,K-1}$ in terms of $\bar{q}_{2,K}$:

$$\begin{aligned}
 &\bar{q}_{2,K-1} \left(D_{1,K-1} - \frac{C_{1,K-2} B_{0,K-1}}{\bar{\gamma}_{K-2} \mu} \right) \\
 &= \bar{q}_{1,K-1} \left(C_{1,K-1} + \frac{C_{1,K-2} A_{0,K-1}}{\bar{\gamma}_{K-2} \mu} \right) - \bar{q}_{2,K} \left(D_{1,K} - \frac{C_{1,K-2} B_{0,K}}{\bar{\gamma}_{K-2} \mu} \right).
 \end{aligned}$$

Thus, using (14) and (15) for $n = 0$, and (15) for $n = 1$ we express $\bar{q}_{1,K-1}$, $\bar{q}_{1,K-2}$ and $\bar{q}_{2,K-1}$ in terms of $\bar{q}_{2,K}$.

Further we continue in the same way: if suppose that all quantities $\bar{q}_{1,k}$ ($k = K - n - 1, \dots, K - 1$) and $\bar{q}_{2,k}$ ($k = K - n, \dots, K$) have been expressed in terms of $\bar{q}_{2,K}$, then from (14) for n we express $\bar{q}_{1,K-n-2}$ by the following equation:

$$\bar{q}_{1,K-n-2} (n + 1) \bar{\gamma}_{K-n-2} \mu = \sum_{k=K-n-1}^K \bar{q}_{2,k} B_{n,k} - \sum_{k=K-n-1}^{K-1} \bar{q}_{1,k} A_{n,k}.$$

Then we substitute with this expression in (15) for $n + 1$, which allows to express $\bar{q}_{2,K-n-1}$ in terms of $\bar{q}_{i,k}$ ($i = 1, 2, k = K - n + i - 2, \dots, K + 2 - i$), i.e. in terms of $\bar{q}_{2,N}$:

$$\begin{aligned} & \bar{q}_{2,K-n-1} \left(D_{n+1,K-n-1} - \frac{C_{n+1,K-n-2} B_{n,K-n-1}}{(n + 1) \bar{\gamma}_{K-n-2} \mu} \right) \\ &= \sum_{k=K-n-1}^{K-1} \bar{q}_{1,k} \left(C_{n+1,k} + \frac{C_{n+1,K-n-2} A_{n,k}}{(n + 1) \bar{\gamma}_{K-n-2} \mu} \right) - \sum_{k=K-n}^K \bar{q}_{2,k} \left(D_{n+1,k} - \frac{C_{n+1,K-n-2} B_{n,k}}{(n + 1) \bar{\gamma}_{K-n-2} \mu} \right). \end{aligned}$$

Thus, repeating this procedure for $n = 0, \dots, K - 1$ we can see that all quantities $\bar{q}_{i,k}$ are proportional to $\bar{q}_{2,K}$. The coefficients of proportionality can be recursively computed from the above described procedure by putting $\bar{q}_{2,K} = 1$. Then $\bar{q}_{2,K}$ can be found with the help of (11), (12) and the normalization condition

$$P_0 + P_1 + P_2 = 1$$

where P_i , ($i = 0, 1, 2$) is the stationary server state distribution,

$$P_i = \lim_{t \rightarrow \infty} P(C(t) = i).$$

Once the probabilities $p_{i,n}$ have been computed we can calculate the main stationary macro characteristics of the system performance:

- server state distribution, P_i ,

$$P_i = \sum_{n=0}^K p_{i,n}, \quad i = 0, 1, 2, \quad p_{1,K} = 0;$$

- mean orbit size,

$$E[R] = \lim_{t \rightarrow \infty} E[R(t)] = \sum_{i=0}^2 \sum_{n=0}^K n p_{i,n};$$

- mean rate of generation of primary regular calls,

$$\Lambda = (K - E[R] - P_1) \lambda_1;$$

- mean waiting time of an regular customer in the orbit,

$$E[W] = \frac{E[R]}{\Lambda};$$

- the blocking probability P_B that one regular customer will find the server busy at the time of its arrival and will join the orbit

$$P_B = \frac{\sum_{n=0}^{K-2} (K-n-1)\lambda_1 p_{1,n} + \sum_{n=0}^{K-1} (K-n)\lambda_1 p_{2,n}}{\sum_{n=0}^{K-2} (K-n-1)\lambda_1 p_{1,n} + \sum_{n=0}^{K-1} (K-n)\lambda_1 (p_{0,n} + p_{2,n})}.$$

4 Conclusions

In this paper we derive formulas for computing the joint distribution of the server state and the orbit size in one single server retrial queue where the server serves a finite number of customers of two types - K regular and $(N - K)$ subscribed. The service times of both types follow two distinct arbitrary laws with common probability distributions. The obtained formulas allow to extend further the investigation of this system by studying the influence of the system input parameters on the macro characteristics of the system performance, on the basis of numerical examples. We also plan to consider the waiting time process, the busy period distribution and other descriptors of the system performance.

References

1. Aguir, S., Karaesmen, E., Aksin, O., Chauvet, F.: The impact of retrials on call center performance. *OR Spectr.* **26**, 353–376 (2004)
2. Amador, J.: On the distribution of the successful and blocked events in retrial queues with finite number of sources. In: *Proceedings of the 5th International Conference on Queueing Theory and Network Applications*, pp. 15–22 (2010)
3. Artalejo, J., Gómez-Corral, A.: *Retrial Queueing Systems: A Computational Approach*. Springer, Heidelberg (2008). <https://doi.org/10.1007/978-3-540-78725-9>
4. Balazsfalvi, G., Sztrik, J.: A tool for modeling distributed protocols. *PIK* **31**(1), 39–44 (2008)
5. Biro, J., Bérczes, T., Kőrösi, A., Heszberger, Z., Sztrik, J.: Discriminatory processor sharing from optimization point of view. In: Dudin, A., De Turck, K. (eds.) *ASMTA 2013*. LNCS, vol. 7984, pp. 67–80. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39408-9_6
6. Choi, B., Shin, Y.W., Ahn, W.C.: Retrial queues with collision arising from unslotted CSMA/CD protocol. *Queueing Syst.* **11**(4), 335–356 (1992)
7. Cooper, R.: *Introduction to Queueing Theory*, 2nd edn. Edward Arnold, London (1981)
8. Deslauriers, A., L'Ecuyer, P., Pichitlamken, J., Ingolfsson, A., Avramidis, A.: Markov chain models of a telephone call center with call blending. *Comput. Oper. Res.* **34**, 1616–1645 (2007)
9. Dragieva, V.: A finite source retrial queue: number of retrials. *Commun. Stati. Theory Methods* **42**(5), 812–829 (2013)

10. Dragieva, V., Phung-Duc, T.: Two-way communication M/M/1//N retrieval queue. In: Thomas, N., Forshaw, M. (eds.) ASMTA 2017. LNCS, vol. 10378, pp. 81–94. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61428-1_6
11. Gómez-Corral, A., Phung-Duc, T.: Retrieval queues and related models. *Ann. Oper. Res.* **247**(1), 1–2 (2016)
12. Falin, G., Templeton, J.: *Retrieval Queues*. Chapman and Hall, London (1997)
13. Falin, G., Artalejo, J.: A finite source retrieval queue. *Eur. J. Oper. Res.* **108**, 409–424 (1998)
14. Fiems, D., Phung-Duc, T.: Light-traffic analysis of random access systems without collisions. *Ann. Oper. Res.* (2017). <https://doi.org/10.1007/s10479-017-2636-7>
15. Jain, R.: *The Art of Computer Systems Performance Analysis*. Wiley&Sons, New York (1991)
16. Jaiswal, N.: *Priority Queues*. Academic press, New York (1969)
17. Kim, J., Kim, B.: A survey of retrieval queueing systems. *Ann. Oper. Res.* **247**(1), 3–36 (2016)
18. Nazarov, A., Sztrik, J., Kvach, A.: Some features of a finite-source M/GI/1 retrieval queueing system with collisions of customers. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) DCCN 2017. CCIS, vol. 700, pp. 186–200. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66836-9_16
19. Ohmura, H., Takahashi, Y.: An analysis of repeated call model with a finite number of sources. *Electron. Commun. Jpn.* **68**, 112–121 (1985)
20. Tran-Gia, P., Mandjes, M.: Modeling of customer retrieval phenomenon in cellular mobile networks. *IEEE J. Sel. Areas Commun.* **15**, 1406–1414 (1997)
21. Van Do, T., Wochner, P., Berches, T., Sztrik, J.: A new finite-source queueing model for mobile cellular networks applying spectrum renting. *Asia-Pac. J. Oper. Res.* **31**, 14400004 (2014)
22. Wang, J., Zhao, L., Zhang, F.: Analysis of the finite source retrieval queues with server breakdowns and repairs. *J. Ind. Manag. Optim.* **7**(3), 655–676 (2011)
23. Wang, J., Wang, F., Sztrick, J., Kuki, A.: Finite source retrieval queue with two phase service. *Int. J. Oper. Res.* **3**(4), 421–440 (2017)
24. Zhang, F., Wang, J.: Performance analysis of the retrieval queues with finite number of sources and service interruption. *J. Korean Stat. Soc.* **42**, 117–131 (2013)



Modeling of a Multi-link Transport Connection by a Network of Queuing Systems

Pavel Mikheev, Anastasiya Pichugina, and Sergey Suschenko^(✉)

National Research Tomsk State University, Lenina Street, 36, 634050 Tomsk, Russia
ssp.inf.tsu@gmail.com

Abstract. A model of a multi-link homogeneous transport connection with limited buffer storage in transit nodes as a network of queuing systems with discrete time is proposed. An estimate of the lower bound of the capacity of the transport connection throughput is obtained. The effect of a set of successive absolutely reliable data links located between two nondeterministic retransmission sections of a given quality on the operational characteristics of a transport connection is found.

Keywords: Transport connection · Transport protocol
Limited buffer storage · Operating characteristics
Network of QS with discrete time

1 Introduction

The most important operational characteristics of multi-link virtual channels are their throughput and average end-to-end delay of protocol data units. These figures are determined not only by the reliability of data transmission in each section of the retransmissions, but also the number of buffer storages for receiving data packets at transit nodes. Known approaches to the analysis of these performance indicators and the results in this area [1–9] are focused on models of QS networks with continuous time for given distributions of input flows and transmission time of protocol data units that do not take into account the specifics of linear control protocols. These methods lead to approximate results, obtained, as a rule, by time-consuming numerical calculations. Modeling using models with discrete time is performed in [10–13], but the results are obtained only at the linear level for a two-link data transmission path. Since the algorithms with decisive feedback are the basis of the control procedures for the linear and transport level protocols, systems with discrete time are the more adequate description of real information transfer processes [11, 13]. However, analysis of QS networks with discrete time is a nontrivial task, since the output flows of discrete Markov QS in most cases lose Markov properties [14]. The model proposes a data transmission path model consisting of several retransmissions sites with limited storage in transit nodes, taking into account the discrete nature of the information transfer process.

2 The Model of the Path in the Form of an Open QS Network

Consider a data transmission path consisting of D consecutive links. We will assume that the exchange in each link is performed by complete information packets in accordance with the start-stop protocol procedure. The duration of the transmission cycle of the packet t from the beginning of its output to the communication line until confirmation of the receipt will be assumed to be the same on all the retransmission sections, and the buffer stores of the transit nodes of the path are bounded by the sizes $K_d, d = \overline{1, D - 1}$.

We also believe that the reliability of packet transmission in the d link is F_d , and the number of retransmissions due to distortions of information packets and acknowledgments, as well as buffer memory locks is not limited. At the same time, the error-free transmission time of a packet for each link is a random variable that is a multiple of t and has a geometric distribution law with the parameter F_d . We assume, in addition, that the transmitting node of the first link always has packets for sending along the considered path, and in transit nodes “external” flows are not added to the main traffic. Then the behavior of the multilinked data transmission path is described by an open Markov network of $D - 1$ discrete QS [13], the intensity of the input flow to which is determined by the value F_d , and the service intensity in each d QS ($d = \overline{1, D - 1}$) is the value of F_{d+1} .

Since we are considering a path with limited queue sizes in transit nodes, the output streams of each discrete QS will not be Markovs [14]. Therefore, such a network cannot be analyzed as a collection of independent Markov discrete QS, but should be described by an enclosed Markov chain in a $D - 1$ -dimensional space with the number of states equal to the product $\prod_{d=1}^{D-1} (K_d + 1)$.

We denote by π_A^B the transition probabilities of the Markov chain from the state A to the state B , where $A = i_{D-1}i_{D-2} \dots i_1; B = j_{D-1}j_{D-2} \dots j_1; i_d = \overline{0, K_d}; j_d = \overline{0, K_d}; d = \overline{1, D - 1} - (D - 1)$ are the bit numbers of the initial and changed states of the Markov chain in a $(D - 1)$ -dimensional space with the cardinality of the set of values in d digit (d dimension of space) equal to $K_d + 1$, and P_A are the probabilities of states of the Markov chain. The throughput of a path of length D , is denoted by $Z_D(K_1, \dots, K_{D-1})$, and the average end-to-end delay — $T_D(K_1, \dots, K_{D-1})$. Since the model in question assumes that the sender always has packets to transmit, this operational indicator corresponds to the average upper limit of the delay. The throughput of the multi-link path is determined by the average value of the acknowledged (serviced) flow:

$$Z_D(K_1, \dots, K_{D-1}) = F_D \sum_{i_1=0}^{K_1} \dots \sum_{i_{D-2}=0}^{K_{D-2}} \sum_{i_{D-1}=1}^{K_{D-1}} P_{i_{D-1} \dots i_1}.$$

The indicator of the average end-to-end packet delay, measured in durations of t , is composed of the time of entry into the QS network (the transmission time at the first link) and the service time in the QS network (the transmission time for

the remaining links before reaching the destination node of the D section of the retransmitting taking into account the presence of queues in transit nodes) [4]:

$$T_D(K_1, \dots, K_{D-1}) = \frac{1 + \bar{K}_D}{Z_D(K_1, \dots, K_{D-1})},$$

where \bar{K}_D is the average number of packets in all transit nodes of the data path (in the QS network):

$$\bar{K}_D = \sum_{i_1=0}^{K_1} \cdots \sum_{i_{D-1}=0}^{K_{D-1}} \sum_{d=1}^{D-1} i_d P_{i_{D-1} \dots i_1}.$$

3 Analysis of the Three-Link Path

Lets start with the data transfer path, which consists of three retransmission areas with buffers of arbitrary size in transit nodes. The type of transition probabilities of the Markov chain describing the transport process along such a path is given in Table 1.

Table 1. Transitional probabilities for a three-link path

$\pi_{i_2 i_1}^{j_2 j_1}$	i_2	i_1	j_2	j_1
F_1	0	0	0	1
$F_1(1 - F_2)$	0	$\bar{1}, \bar{K}_1 - 1$	0	$i_1 + 1$
$F_2(1 - F_1)$	0	$\bar{1}, \bar{K}_1$	1	$i_1 - 1$
$F_1 F_2$	0	$\bar{1}, \bar{K}_1$	1	i_1
$F_3(1 - F_1)$	$\bar{1}, \bar{K}_2$	0	$i_2 - 1$	0
$F_1(1 - F_3)$	$\bar{1}, \bar{K}_2$	0	i_2	1
$F_1(1 - F_3)$	K_2	$\bar{1}, \bar{K}_1 - 1$	K_2	$i_1 + 1$
$F_1 F_3$	$\bar{1}, \bar{K}_2$	0	$i_2 - 1$	1
$F_3(1 - F_1)(1 - F_2)$	$\bar{1}, \bar{K}_2$	$\bar{1}, \bar{K}_1 - 1$	$i_2 - 1$	i_1
$F_3(1 - F_2)$	$\bar{1}, \bar{K}_2$	K_1	$i_2 - 1$	K_1
$F_2 F_3(1 - F_1)$	$\bar{1}, \bar{K}_2$	$\bar{1}, \bar{K}_1$	i_2	$i_1 - 1$
$F_2(1 - F_1)(1 - F_3)$	$\bar{1}, \bar{K}_2 - 1$	$\bar{1}, \bar{K}_1$	$i_2 + 1$	$i_1 - 1$
$F_1 F_2(1 - F_3)$	$\bar{1}, \bar{K}_2 - 1$	$\bar{1}, \bar{K}_1$	$i_2 + 1$	i_1
$F_1(1 - F_2)(1 - F_3)$	$\bar{1}, \bar{K}_2 - 1$	$\bar{1}, \bar{K}_1 - 1$	i_2	$i_1 + 1$
$F_1 F_3(1 - F_2)$	$\bar{1}, \bar{K}_2$	$\bar{1}, \bar{K}_1 - 1$	$i_2 - 1$	$i_1 + 1$

For $K_1 = K_2 = 1$, the solution of the system of local equilibrium equations for the Markov chain describing the three-link transport connection has the form:

$$\begin{aligned}
 P_{00} &= \frac{F_2 F_3^2 (1 - F_1)^2}{F_3 (F_1 + F_3 (1 - F_1)) (F_1 + F_2 (1 - F_1)) + F_1^2 F_2 (1 - F_3)}; \\
 P_{01} &= P_{00} \frac{F_1 (F_1 (1 - F_2) + F_3 (1 - F_1))}{F_2 F_3 (1 - F_1)^2}; \\
 P_{10} &= P_{00} \frac{F_1}{F_3 (1 - F_1)}; \quad P_{11} = P_{00} \frac{F_1^2}{F_3^2 (1 - F_1)^2}.
 \end{aligned}$$

The throughput of the three-link path is determined by the value:

$$Z_3(1, 1) = \frac{F_1 F_2 F_3 (F_1 + F_3 (1 - F_1))}{F_3 (F_1 + F_3 (1 - F_1)) (F_1 + F_2 (1 - F_1)) + F_1^2 F_2 (1 - F_3)}.$$

Let us consider particular cases of this solution. It is not difficult to see that for two absolutely reliable channels ($F_1 = F_2 = 1$, or $F_2 = F_3 = 1$, or $F_1 = F_3 = 1$), the throughput of a three-link path is determined by the reliability of the transmission in the third (F_3 , or F_1 , or F_2 , respectively).

For the case when the first retransmission area is absolutely reliable ($F_1 = 1$), the throughput assumes the form coinciding with the expression of this indicator for the two-link path [12]:

$$Z_3(1, 1) = \frac{F_2 F_3}{F_2 + F_3 (1 - F_2)}. \quad (1)$$

With the statistically homogeneous second and third links of the data transfer path ($F_2 = F_3 = F$), this relation is transformed to the form:

$$Z_3(1, 1) = \frac{F}{2 - F}. \quad (2)$$

The throughput of the path with the deterministic average channel ($F_2 = 1$) takes the following form:

$$Z_3(1, 1) = \frac{F_1 F_3 (F_1 + F_3 (1 - F_1))}{F_3 (F_1 + F_3 (1 - F_1)) + F_1^2 (1 - F_3)}.$$

In this case the values $F_1 = F_3 = F$ lead to the relation:

$$Z_3(1, 1) = \frac{F(2 - F)}{1 + 2(1 - F)}. \quad (3)$$

For $F_3 = 1$ we have

$$Z_3(1, 1) = \frac{F_1 F_2}{F_1 + F_2 (1 - F_1)}.$$

It is not difficult to see that this relation is exactly the same as (1). The comparison (2) and (3) shows that (3) exceeds (2) by $\Delta = \frac{F(1-F)^2}{(3-2F)(2-F)}$, assuming

the maximum value when $F = 0.468$. This fact is easily explained by the fact that the absolutely reliable channel of the second link of data transfer serves as an additional buffer for storing packets between the first and third sections of the retransmission, thereby reducing the probability of blocking buffer memory. For a statistically homogeneous data transmission path ($F_1 = F_2 = F_3 = F$) we have:

$$Z_3(1, 1) = F \frac{1 + (1 - F)}{1 + 3(1 - F) + (1 - F)^2}. \tag{4}$$

In this case, the indicator of the average end-to-end packet delay expressed in the transmission cycle duration of the t packet will be

$$T_3(1, 1) = \frac{3 + 6(1 - F) + (1 - F)^2}{F(1 + (1 - F))}.$$

Now let us consider a statistically homogeneous path for $K_1 = 1$ and an arbitrary K_2 . For a given K_2 , writing out the equilibrium equations, taking into account the normalization condition, one can find the state probabilities and operational parameters of the path. For $K_2 = \overline{2, 5}$, the throughput values and average end-to-end delay are as follows:

$$\begin{aligned} Z_3(1, 2) &= \frac{F\{3 + 3(1 - F) + (1 - F)^2\}}{3 + 7(1 - F) + 4(1 - F)^2 + (1 - F)^3}; \\ Z_3(1, 3) &= \frac{F\{7 + 8(1 - F) + 4(1 - F)^2 + (1 - F)^3\}}{7 + 16(1 - F) + 12(1 - F)^2 + 5(1 - F)^3 + (1 - F)^4}; \\ Z_3(1, 4) &= \frac{F\{15 + 20(1 - F) + 13(1 - F)^2 + 5(1 - F)^3 + (1 - F)^4\}}{15 + 36(1 - F) + 33(1 - F)^2 + 18(1 - F)^3 + 6(1 - F)^4 + (1 - F)^5}; \\ Z_3(1, 5) &= F\left\{31 + 48(1 - F) + 38(1 - F)^2 + 19(1 - F)^3 + 6(1 - F)^4 + (1 - F)^5\right\} / \left\{31 + 80(1 - F) + 86(1 - F)^2 + 57(1 - F)^3 + 25(1 - F)^4 + 7(1 - F)^5 + (1 - F)^6\right\}; \\ T_3(1, 2) &= \frac{10 + 15(1 - F) + 6(1 - F)^2 + (1 - F)^3}{F\{3 + 3(1 - F) + (1 - F)^2\}}; \\ T_3(1, 3) &= \frac{25 + 37(1 - F) + 21(1 - F)^2 + 7(1 - F)^3 + (1 - F)^4}{F\{7 + 8(1 - F) + 4(1 - F)^2 + (1 - F)^3\}}; \\ T_3(1, 4) &= \frac{56 + 89(1 - F) + 63(1 - F)^2 + 29(1 - F)^3 + 8(1 - F)^4 + (1 - F)^5}{F\{15 + 20(1 - F) + 13(1 - F)^2 + 5(1 - F)^3 + (1 - F)^4\}}; \\ T_3(1, 5) &= \left\{119 + 209(1 - F) + 180(1 - F)^2 + 101(1 - F)^3 + 38(1 - F)^4 + 9(1 - F)^5 + (1 - F)^6\right\} / F\left\{31 + 48(1 - F) + 38(1 - F)^2 + 19(1 - F)^3 + 6(1 - F)^4 + (1 - F)^5\right\}. \end{aligned}$$

Under the assumption that K_1 is arbitrary and $K_2 = 1$ it is easy to obtain the values P_{ij} and the ratios for the throughput that satisfy the equality

$$Z_3(K_1, 1) = Z_3(1, K_2) \tag{5}$$

if k_1 and K_2 coincide here. Thus, the throughput rate is invariant to the order of transit nodes with buffer storage of different volumes along a statistically homogeneous data path. At the same time, the average end-to-end delay is dependent on this order:

$$\begin{aligned}
 T_3(2, 1) &= \frac{11 + 20(1 - F) + 10(1 - F)^2 + 2(1 - F)^3}{F\{3 + 3(1 - F) + (1 - F)^2\}}; \\
 T_3(3, 1) &= \frac{31 + 61(1 - F) + 43(1 - F)^2 + 17(1 - F)^3 + 3(1 - F)^4}{F\{7 + 8(1 - F) + 4(1 - F)^2 + (1 - F)^3\}}; \\
 T_3(4, 1) &= \left\{79 + 173(1 - F) + 153(1 - F)^2 + 81(1 - F)^3 + 26(1 - F)^4 \right. \\
 &\quad \left. + 4(1 - F)^5\right\} / F\left\{15 + 20(1 - F) + 13(1 - F)^2 \right. \\
 &\quad \left. + 5(1 - F)^3 + (1 - F)^4\right\}; \\
 T_3(5, 1) &= \left\{191 + 465(1 - F) + 488(1 - F)^2 + 317(1 - F)^3 + 136(1 - F)^4 \right. \\
 &\quad \left. + 37(1 - F)^5 + 5(1 - F)^6\right\} / F\left\{31 + 48(1 - F) + 38(1 - F)^2 \right. \\
 &\quad \left. + 19(1 - F)^3 + 6(1 - F)^4 + (1 - F)^5\right\}.
 \end{aligned}$$

Numerical analysis shows that with the growth of k_2 , the throughput of the three-link path $Z_3(1, K_2)$ rapidly tends to the theoretical limit $Z_2(1)$. A study of a homogeneous path with $K_1 = K_2 = 2$ shows that operational characteristics are determined by expressions:

$$\begin{aligned}
 Z_3(2, 2) &= F\left\{24 + 26(1 - F) + 17(1 - F)^2 + 9(1 - F)^3 \right. \\
 &\quad \left. + 2(1 - F)^4\right\} / \left\{24 + 46(1 - F) + 35(1 - F)^2 + 21(1 - F)^3 \right. \\
 &\quad \left. + 8(1 - F)^4 + (1 - F)^5\right\}; \\
 T_3(2, 2) &= \left\{96 + 140(1 - F) + 96(1 - F)^2 + 55(1 - F)^3 + 17(1 - F)^4 \right. \\
 &\quad \left. + (1 - F)^5\right\} / F\left\{24 + 26(1 - F) + 17(1 - F)^2 \right. \\
 &\quad \left. + 9(1 - F)^3 + 2(1 - F)^4\right\}.
 \end{aligned}$$

With the further increase of k_1 and K_2 , the structural complexity of the analytical solution is rapidly increasing. A comparative analysis of the throughput of $Z_3(K_1, K_2)$ with different relations between k_1 and k_2 shows that the uniform distribution of buffers along the data path provides the best values of this

operating characteristic. This fact should be considered when building multi-link connections.

Assume now that $f_1 = 1$, $F_2 = F_3 = F$, and k_1 and k_2 are arbitrary. Then the set of probable States is formed by a set of two adjacent geometric figure: rectangle ($i = \overline{0, 1}$; $j = \overline{1, K_2}$) and line segment ($i = \overline{0, K_1}$; $j = K_2$). Operational indicators thus have the following form:

$$Z_3(K_1, K_2) = \frac{F\{K_1 + K_2 - F\}}{1 + K_1 + K_2 - 2F}; \tag{6}$$

$$T_3(K_1, K_2) = \frac{K_1(3 + K_1 + 2K_2) + K_2(3 + K_2) + 2 - 4F - 2F^2}{2F\{K_1 + K_2 - F\}}.$$

It follows from relation (6) that for an unbounded growth of K_1 or K_2 , the throughput of $Z_3(K_1, K_2)$ tends to the value F .

4 Analysis of the Multi-link Path

Consider a virtual connection with the number of packet storage locations in transit nodes equal to one. Because of this, the condition numbers of Markov chain represent the binary number with number of digits equal to $D1$. Since each link of the path is controlled by a start-stop protocol procedure, the permissible state changes correspond to a single shift to the left of a subset of the bits of the initial Markov chain state number. In this case, the “non-transmission” of the data package (as a result of distortion in the communication channel) from one transit node to the next leads to the effect of “blocking” (locking) the package located in the previous node. Transition probabilities of the Markov chain that describes the transportation process in the transmitting tract of the four units have a dependency on the parameters of the inter-nodal sections of the retransmission that are listed in the Table 2.

For $F_d = F$, $d = \overline{1, 4}$, the probabilities of states of a given Markov chain are determined by the relations:

$$\begin{aligned}
 P_{000} &= \frac{(1 - F)^3}{1 + 6(1 - F) + 6(1 - F)^2 + (1 - F)^3}; \\
 P_{001} &= P_{000} \frac{3}{1 - F}; \quad P_{010} = P_{000} \frac{2}{1 - F}; \quad P_{100} = P_{000} \frac{1}{1 - F}; \\
 P_{011} &= P_{000} \frac{3}{(1 - F)^2}; \quad P_{101} = P_{000} \frac{2}{(1 - F)^2}; \\
 P_{110} &= P_{000} \frac{1}{(1 - F)^2}; \quad P_{111} = P_{000} \frac{1}{(1 - F)^3}.
 \end{aligned}$$

Table 2. Transitional probabilities for a four-link path for $K_1 = K_2 = K_3 = 1$

$\pi_{i_3 i_2 i_1}^{j_3 j_2 j_1}$	i_3	i_2	i_1	j_3	j_2	j_1
F_1	0	0	0	0	0	1
$F_2(1 - F_1)$	0	0	1	0	1	0
$F_1 F_2$	0	0	1	0	1	1
$F_1(1 - F_3)$	0	1	0	0	1	1
$F_3(1 - F_1)$	0	1	0	1	0	0
$F_1 F_3$	0	1	0	1	0	1
$F_3(1 - F_2)$	0	1	1	1	0	1
$F_2 F_3(1 - F_1)$	0	1	1	1	1	0
$F_1 F_2 F_3$	0	1	1	1	1	1
$F_4(1 - F_1)$	1	0	0	0	0	0
$F_1 F_4$	1	0	0	0	0	1
$F_1(1 - F_4)$	1	0	0	1	0	1
$F_4(1 - F_2)$	1	0	1	0	0	1
$F_2 F_4(1 - F_1)$	1	0	1	0	1	0
$F_1 F_2 F_3$	1	0	1	0	1	1
$F_2(1 - F_1)(1 - F_4)$	1	0	1	1	1	0
$F_1 F_2(1 - F_4)$	1	0	1	1	1	1
$F_4(1 - F_1)(1 - F_3)$	1	1	0	0	1	0
$F_1 F_4(1 - F_3)$	1	1	0	0	1	1
$F_3 F_4(1 - F_1)$	1	1	0	1	0	0
$F_1 F_3 F_4$	1	1	0	1	0	1
$F_1(1 - F_4)$	1	1	0	1	1	1
$F_4(1 - F_3)$	1	1	1	0	1	1
$F_3 F_4(1 - F_2)$	1	1	1	1	0	1
$F_2 F_3 F_4(1 - F_1)$	1	1	1	1	1	0

Dependence of operational indicators on the parameters of the data transmission path has the form:

$$\begin{aligned}
 Z_4(1, 1, 1) &= \frac{F\{1 + 3(1 - F) + (1 - F)^2\}}{1 + 6(1 - F) + 6(1 - F)^2 + (1 - F)^3}; \\
 T_4(1, 1, 1) &= \frac{4 + 18(1 - F) + 12(1 - F)^2 + (1 - F)^3}{F\{1 + 3(1 - F) + (1 - F)^2\}}.
 \end{aligned}
 \tag{7}$$

Table 3. Transitional probabilities of the Markov chain for the four-link data path for $F_2 = F_3 = 1$

$\pi_{i_3 i_2 i_1}^{j_3 j_2 j_1}$	i_3	i_2	i_1	j_3	j_2	j_1
F_1	0	0	0	1	0	0
$1 - F_1$	0	0	1	0	1	0
F_1	0	0	1	0	1	1
$1 - F_1$	0	1	0	1	0	0
F_1	0	1	0	1	0	1
$1 - F_1$	0	1	1	1	1	0
F_1	0	1	1	1	1	1
$F_1(1 - F_4)$	$\overline{1, K_3}$	0	0	i_3	0	1
$F_4(1 - F_1)$	$\overline{1, K_3}$	0	0	$i_3 - 1$	0	0
$F_1 F_4$	$\overline{1, K_3}$	0	0	$i_3 - 1$	0	1
$F_4(1 - F_1)$	$\overline{1, K_3}$	0	1	$i_3 - 1$	1	0
$F_1 F_4$	$\overline{1, K_3}$	0	1	$i_3 - 1$	1	1
$(1 - F_1) \times (1 - F_4)$	$\overline{1, K_3}$	0	1	i_3	1	0
$F_1(1 - F_4)$	$\overline{1, K_3}$	0	1	i_3	1	1
$F_4(1 - F_1)$	$\overline{1, K_3}$	1	0	i_3	0	0
$F_1 F_4$	$\overline{1, K_3}$	1	0	i_3	0	1
$(1 - F_1) \times (1 - F_4)$	$\overline{1, K_3 - 1}$	1	0	$i_3 + 1$	0	0
$F_1(1 - F_4)$	$\overline{1, K_3 - 1}$	1	0	$i_3 + 1$	0	1
$F_4(1 - F_1)$	$\overline{1, K_3}$	1	0	i_3	1	0
$(1 - F_1) \times (1 - F_4)$	$\overline{1, K_3 - 1}$	1	1	$i_3 + 1$	1	0
$F_1(1 - F_4)$	$\overline{1, K_3 - 1}$	1	1	$i_3 + 1$	1	1
$F_1(1 - F_4)$	K_3	$\overline{1, K_2}$	0	K_3	i_2	1
$F_4(1 - F_1)$	K_3	$\overline{1, K_2}$	0	K_3	$i_2 - 1$	0
$F_1 F_4$	K_3	$\overline{1, K_2}$	0	K_3	$i_2 - 1$	1
$F_4(1 - F_1)$	K_3	$\overline{1, K_2}$	1	K_3	i_2	0
$(1 - F_1) \times (1 - F_4)$	K_3	$\overline{1, K_2 - 1}$	1	K_3	$i_2 + 1$	0
$F_1(1 - F_4)$	K_3	$\overline{1, K_2 - 1}$	1	K_3	$i_2 + 1$	1
$F_1(1 - F_4)$	K_3	K_2	$\overline{1, K_1 - 1}$	K_3	K_2	$i_1 + 1$
$F_4(1 - F_1)$	K_3	K_2	$\overline{1, K_1}$	K_3	K_2	$i_1 - 1$

Let us consider a transfer path at $F_2 = F_3 = 1$, $F_1 = F_4 = F$, and arbitrary K_d , $d = \overline{1, 3}$. The space of probable states is formed by the combination of three adjacent geometric figures: rectangular parallelepiped ($i = 0, 1$; $j = 0, 1$; $k = \overline{0, K_3}$), the rectangle ($i = 0, 1$; $j = \overline{0, K_2}$; $k = K_3$) and line segment

($i = \overline{0, K_1}$; $j = K_2$; $k = K_3$). The transient probabilities of Markov chain for this path are given in Table 3.

The probability of states are determined by the following dependencies:

$$\begin{aligned}
 P_{000} &= \frac{(1-F)^3}{1+K_1+K_2+K_3-3F}; & P_{K_3K_2i} &= P_{000} \frac{1}{(1-F)^3}, & i &= \overline{1, K_1}; \\
 P_{k00} &= P_{000} \frac{1}{1-F}, & k &= \overline{1, K_3}; & P_{k01} &= P_{000} \frac{F}{(1-F)^2}, & k &= \overline{1, K_3}; \\
 P_{k10} &= P_{000} \frac{F}{(1-F)^2}, & k &= \overline{1, K_3-1}; & P_{k11} &= P_{000} \frac{F^2}{(1-F)^3}, & k &= \overline{1, K_3-1}; \\
 P_{K_3j0} &= P_{000} \frac{1}{(1-F)^2}, & j &= \overline{1, K_2}; & P_{K_3j1} &= P_{000} \frac{F}{(1-F)^3}, & j &= \overline{1, K_2-1}.
 \end{aligned}$$

For throughput and end-to-end delay, its fair

$$Z_4(K_1, K_2, K_3) = \frac{F\{K_1 + K_2 + K_3 - 2F\}}{1 + K_1 + K_2 + K_3 - 3F}; \quad (8)$$

$$\begin{aligned}
 T_4(K_1, K_2, K_3) &= \left\{ K_1(3 + K_1 + 2K_2 + 2K_3) + K_2(3 + K_2 + 2K_3) \right. \\
 &\quad \left. + K_3(3 + K_3) + 2 - 6F - 6F^2 \right\} / 2F \left\{ K_1 + K_2 + K_3 - 2F \right\}.
 \end{aligned}$$

Let us consider a data path of length $D = 5$ with a single buffer storage size in transit nodes. The rules for constructing transient probabilities of Markov chain describing such a path correspond to the principles given above. These probabilities for states with 0000 to 0111 coincide with transient probabilities for a path of length $D = 4$ (see Table 2) given that the shift to the left of the third bit of the state number at $D = 5$ is not lost, but remains within the bit grid. For the remaining states, the transition probabilities of Markov chain are given in Table 4.

For $F_d = F$, $d = \overline{1, 5}$, the probability states are:

$$\begin{aligned}
 P_{0000} &= \frac{(1-F)^3}{1+10(1-F)+20(1-F)^2+10(1-F)^3+(1-F)^4}; \\
 P_{0001} &= P_{0000} \frac{4}{1-F}; & P_{0010} &= P_{0000} \frac{3}{1-F}; \\
 P_{0100} &= P_{0000} \frac{2}{1-F}; & P_{1000} &= P_{0000} \frac{1}{1-F}; \\
 P_{0011} &= P_{0000} \frac{6}{(1-F)^2}; & P_{0101} &= P_{0000} \frac{5}{(1-F)^2}; \\
 P_{0110} &= P_{0000} \frac{3}{(1-F)^2}; & P_{1001} &= P_{0000} \frac{3}{(1-F)^2}; \\
 P_{1010} &= P_{0000} \frac{2}{(1-F)^2}; & P_{1100} &= P_{0000} \frac{1}{(1-F)^2};
 \end{aligned}$$

$$\begin{aligned}
 P_{0111} &= P_{0000} \frac{4}{(1-F)^3}; & P_{1011} &= P_{0000} \frac{3}{(1-F)^3}; \\
 P_{1101} &= P_{0000} \frac{2}{(1-F)^3}; & P_{1110} &= P_{0000} \frac{1}{(1-F)^3}; \\
 P_{1111} &= P_{0000} \frac{1}{(1-F)^4}.
 \end{aligned}$$

The structure of the solution of the system of local equilibrium equations has the form of the relation of the integer coefficient to the probability of distortion of the data $1 - F$, raised to the power equal to the number of requirements in the network QS (the number of units in the state number). Unknown integer coefficients are easily determined by direct substitution of the solution into equilibrium equations. Under this scheme, you can define a probability distribution for an arbitrary length path with a single buffer pool at transit nodes.

The throughput of a statistically homogeneous five-link path is defined by the expression:

$$Z_5(1, 1, 1, 1) = \frac{F\{1 + 6(1 - F) + 6(1 - F)^2 + (1 - F)^3\}}{1 + 10(1 - F) + 20(1 - F)^2 + 10(1 - F)^3 + (1 - F)^4}, \tag{9}$$

and the average end-to-end delay — by dependency:

$$T_5(1, 1, 1, 1) = \frac{5 + 40(1 - F) + 60(1 - F)^2 + 20(1 - F)^3 + (1 - F)^4}{F\{1 + 6(1 - F) + 6(1 - F)^2 + (1 - F)^3\}}.$$

For $F_2 = F_3 = F_4 = 1$, $F_1 = F_5 = F$, the probabilities of states take the form:

$$\begin{aligned}
 P_{0000} &= \frac{(1 - F)^3}{5 - 4F}; & P_{0001} &= P_{0010} = P_{0100} = P_{0000} \frac{F}{1 - F}; \\
 P_{1000} &= P_{0000} \frac{1}{1 - F}; & P_{0011} &= P_{0101} = P_{0110} = P_{0000} \frac{F^2}{(1 - F)^2}; \\
 P_{1001} &= P_{1010} = P_{0000} \frac{F}{(1 - F)^2}; \\
 P_{1100} &= P_{0000} \frac{1}{(1 - F)^2}; & P_{0111} &= P_{0000} \frac{F^3}{(1 - F)^3}; \\
 P_{1011} &= P_{0000} \frac{F^2}{(1 - F)^3}; & P_{1101} &= P_{0000} \frac{F}{(1 - F)^3}; \\
 P_{1110} &= P_{0000} \frac{1}{(1 - F)^3}; & P_{1111} &= P_{0000} \frac{1}{(1 - F)^4}.
 \end{aligned}$$

Operating indicators in this case are determined by the relation:

$$\begin{aligned}
 Z_5(1, \dots, 1) &= F \frac{1 + 3(1 - F)}{1 + 4(1 - F)}, \\
 T_5(1, \dots, 1) &= \frac{5 + 4(1 - F) + 6(1 - F)^2}{F\{1 + 3(1 - F)\}}.
 \end{aligned} \tag{10}$$

For data transmission paths consisting of 6 and 7 retransmission sections, transition probabilities, probabilities of states of Markov chain and operational parameters are obtained in a similar way:

$$\begin{aligned}
 Z_6(1, \dots, 1) = & F \left\{ 1 + 10(1 - F) + 20(1 - F)^2 + 10(1 - F)^3 \right. \\
 & \left. + (1 - F)^4 \right\} / \left\{ 1 + 15(1 - F) + 50(1 - F)^2 + 50(1 - F)^3 \right. \\
 & \left. + 15(1 - F)^4 + (1 - F)^5 \right\}; \tag{11}
 \end{aligned}$$

$$\begin{aligned}
 Z_7(1, \dots, 1) = & F \left\{ 1 + 15(1 - F) + 50(1 - F)^2 + 50(1 - F)^3 \right. \\
 & \left. + 15(1 - F)^4 + (1 - F)^5 \right\} / \left\{ 1 + 21(1 - F) + 105(1 - F)^2 \right. \\
 & \left. + 175(1 - F)^3 + 105(1 - F)^4 + 21(1 - F)^5 + (1 - F)^6 \right\}; \tag{12}
 \end{aligned}$$

$$\begin{aligned}
 T_6(1, \dots, 1) = & \left\{ 6 + 75(1 - F) + 200(1 - F)^2 + 150(1 - F)^3 \right. \\
 & \left. + 30(1 - F)^4 + (1 - F)^5 \right\} / F \left\{ 1 + 10(1 - F) + 20(1 - F)^2 \right. \\
 & \left. + 10(1 - F)^3 + (1 - F)^4 \right\};
 \end{aligned}$$

$$\begin{aligned}
 T_7(1, \dots, 1) = & \left\{ 1 + 126(1 - F) + 525(1 - F)^2 + 700(1 - F)^3 \right. \\
 & \left. + 315(1 - F)^4 + 42(1 - F)^5 + (1 - F)^6 \right\} / F \left\{ 1 + 15(1 - F) \right. \\
 & \left. + 50(1 - F)^2 + 50(1 - F)^3 + 15(1 - F)^4 + (1 - F)^5 \right\}
 \end{aligned}$$

for $F_d = F$, $d = \overline{1, D}$ and

$$Z_6(1, \dots, 1) = F \frac{1 + 4(1 - F)}{1 + 5(1 - F)}; \tag{13}$$

$$Z_7(1, \dots, 1) = F \frac{1 + 5(1 - F)}{1 + 6(1 - F)}; \tag{14}$$

$$T_6(1, \dots, 1) = \frac{6 + 5(1 - F) + 10(1 - F^2)}{F \{ 1 + 4(1 - F) \}};$$

$$T_7(1, \dots, 1) = \frac{7 + 6(1 - F) + 15(1 - F^2)}{F \{ 1 + 5(1 - F) \}}$$

in case $F_d = 1$, $d = \overline{2, D - 1}$, $F_1 = F_D = f$. View relations for throughput (2), (3), (10), (13) and (14) allows to generalize them by a single entry for the path of arbitrary length D , containing $D - 2$ consecutive transit deterministic links of data transfer, located between two non-deterministic sections of the retransmission with a single volume of the drive in transit nodes:

$$Z_D(1, \dots, 1) = F \frac{1 + (D - 2)(1 - F)}{1 + (D - 1)(1 - F)}.$$

In addition, considering the type of dependencies (6) and (8) the throughput of the path consisting of deterministic channels between two non-deterministic sections of the retransmission, with an arbitrary number of buffers in transit nodes, will be rewritten as follows:

$$Z_D(K_1, \dots, K_{D-1}) = F \frac{\sum_{d=1}^{D-1} K_d - F(D-2)}{1 + \sum_{d=1}^{D-1} K_d - F(D-1)}.$$

Hence it is not difficult to conclude that when building network data transmission paths consisting of a large number of retransmission sites, reliable communication channels should be evenly distributed between links with a high level of distortion. Thus, these retransmission sites will play the role of additional buffers between unreliable links and reduce the negative locking factor of the buffer memory Table 5.

From the form of the dependencies for the probabilities of states of Markov chain and throughput of the statistically homogeneous data transmission path (4), (7), (9), (11) and (12), we can construct the lower bound of throughput indicator $Z_D^*(1, \dots, 1) \leq Z_D(1, \dots, 1)$:

$$Z_D^*(1, \dots, 1) = F \frac{\sum_{d=0}^{D-2} (1-F)^d \binom{D-2}{d} \left\{ 1 + \binom{D-2}{d} \right\}}{\sum_{d=0}^{D-1} (1-F)^d \binom{D-1}{d} \left\{ 1 + \binom{D-1}{d} \right\}}. \tag{15}$$

For $D \leq 4$ this estimate coincides with the exponent $Z_D^*(1, \dots, 1)$. Numerical studies confirm that for $D \geq 5$ the dependence (15) approximates well from below the throughput of the multilinked data transmission path, which decreases monotonically with increasing of its length for $F \leq 1$. A similar upper bound

Table 5. The distribution of the throughput values and its estimates on the reliability of the packet transmission for paths with buffer pools of unit length

F	$Z_5(1, \cdot, 1)$	$Z_6(1, \cdot, 1)$	$Z_7(1, \cdot, 1)$	$Z_5^*(1, \cdot, 1)$	$Z_6^*(1, \cdot, 1)$	$Z_7^*(1, \cdot, 1)$
0.1	0.035	0.033	0.032	0.034	0.032	0.030
0.2	0.074	0.070	0.068	0.072	0.067	0.064
0.3	0.118	0.113	0.109	0.116	0.108	0.102
0.4	0.169	0.161	0.156	0.166	0.154	0.146
0.5	0.228	0.218	0.211	0.224	0.209	0.198
0.6	0.299	0.286	0.276	0.294	0.275	0.261
0.7	0.387	0.370	0.358	0.382	0.358	0.341
0.8	0.504	0.483	0.468	0.499	0.471	0.449
0.9	0.676	0.651	0.633	0.673	0.643	0.618

for the average end-to-end delay $T_D^*(1, \dots, 1) \geq T_D(1, \dots, 1)$ has the form:

$$T_D^*(1, \dots, 1) = \frac{\sum_{d=0}^{D-1} (1-F)^{D-1-d} (d+1) \binom{D-1}{d} \left\{ 1 + \binom{D-1}{d} \right\}}{F \sum_{d=0}^{D-2} (1-F)^d \binom{D-2}{d} \left\{ 1 + \binom{D-2}{d} \right\}}.$$

5 Conclusion

Discrete models of a multi-link data transmission path are proposed, which differ by taking into account the blocking factor of the limited buffer memory of transit nodes. The proposed models allow analyzing the effect of storage capacity on transport protocols performance indicators. The invariance of the throughput index to the order of transit nodes with buffer storage of different capacities along a statistically homogeneous data path is found, and the dependence of the average end-to-end delay of the packet on this order is insignificant. The expediency of the uniform distribution of the buffer space among the transit nodes along the multi-link path, ensuring the best performance of the transport connection, is established. It is shown that when constructing network data transmission paths consisting of a large number of retransmission site, reliable communication channels should be evenly distributed between links with a high level of distortion. In this way, these retransmission sites act as additional buffers between unreliable links and reduce the negative blocking factor of the buffer memory. An analytical estimate of the lower limit of throughput and an upper estimate of the average end-to-end delay of the multi-link data path corresponding to the minimum number of buffers in the transit nodes are obtained.

References

1. Basharin, G.P., Bocharov, P.P., Kogan, Y.A.: Analysis of Queues in Computer Networks. Theory and Methods of Calculation. Nauka, Moscow (1989)
2. Boguslavskii, L.B.: Data Flow Control in Computer Networks. Energoatomizdat, Moscow (1984)
3. Zhozhikashvili, V.A., Vishnevsky, V.M.: Networks Queuing. Theory and Application to Computer Networks. Radio and Communication, Moscow (1988)
4. Vishnevsky, V.M.: Theoretical Bases of Designing of Computer Networks. Technosphere, Moscow (2003)
5. Ivnitsky, V.A.: Theory of Queuing Networks. Phys.-Mat. lit, Moscow (2004)
6. Walrand, G.: Introduction to the Theory of Queuing Networks. Mir, Moscow (1993)
7. Moiseev, A.N., Nazarov, A.A.: Infinitely Linear Systems and Queuing Networks. Publishing House NTL, Tomsk (2015)
8. Zorkaltsev, A.V.: Analysis of local flow control in packet switching node. J. Autom. Control Comput. Eng. **4**, 20–27 (1992)
9. Zorkaltsev, A.V.: Analysis of procedures flow control in a switching node of the network with virtual channels. J. Autom. Control Comput. Eng. **4**, 35–42 (1993)

10. Mikheev, P.A.: Analyzing sharing strategies for finite buffer memory in a router among outgoing channels. *J. Autom. Remote Control* **75**, 1814–1825 (2014)
11. Suschenko, S.P.: The influence of buffer overfilling on the speed of synchronous data-transmission control procedures. *J. Autom. Remote Control* **60**, 1460–1468 (1999)
12. Suschenko, S.P.: On the effect of locking the buffer memory to the operational characteristics of a link data transmission. *J. Autom. Control Comput. Eng.* **6**, 27–34 (1985)
13. Mikheev, P.A., Suschenko, S.P.: *Mathematical Models of Networks of Access Level*. Nauka, Novosibirsk (2015)
14. Ivanovsky, V.B.: On properties of output flows in digital service systems. *J. Autom. Remote Control* **45**, 1413–1419 (1984)



Estimation of Prioritized Disciplines Efficiency Based on the Metamodel of Multi-flows Queueing Systems

V. N. Zadorozhnyi¹, T. R. Zakharenkova^{1(✉)}, and D. A. Tulubaev²

¹ Omsk State Technical University, Omsk, Russia
zwn2015@yandex.ru, ZakharenkovaTatiana@gmail.com

² OOO Dalnefteprovod, Khabarovsk, Russia

Abstract. The metamodel of control systems, representing a generalized queueing system with a large number of request classes and the classes parameters as random variables, is proposed. Crucial measurable control object characteristics determining the feasibility of prioritized service disciplines development and implementation are indicated.

Keywords: Complex control object · Queueing system · Metamodel
Service discipline · Optimal priority assignment

1 Introduction

At the present time Queueing Theory, as theory adequately describing the problems of multiple-access to limited resources and developing new methods for solving these problems, holds an important position in operational research [1–4] and has numerous applications. Particularly, Queueing Theory has proved its practical value and it continues to be highly demanded in the field of computer networks design at the system level [5–7].

In large-scale control objects (CO) such as electrical power systems, airdromes, main oil pipelines, large plants and organizations, it is natural to regard the cumulative signal (request) flow processed by control system as Poisson flow, which is confirmed by the measurements [8]. However, the assumption about exponential service time distribution (where the coefficient of variation (c.v.) is equal to 1) no longer seems so natural. A complex large-scale object gives rise to a set of request classes varying in labor-intensity of their services, arrival rate, losses caused by service delay, and etc. Different request classes can be described by various distributions of service time: it may be close to a constant, i.e. with the c.v. being close to zero or highly exceeding 1 due to splittings contained in a processing algorithm. Therefore, as far as queueing theory is concerned, we will study the systems with a big amount of Poisson request flows (classes) that can be significantly diverse in terms of service time properties or other parameters. If some request classes arrive in queueing system (QS), one should take into

consideration the constraints on request waiting times given for the classes and find service disciplines that allow us to consider the constraints in the best way.

In addition, for service disciplines to be optimized by existing methods one must specify the numerical parameters of QS: a mean and c.v. for service time of different class requests, cost of waiting times constraints violation, flow intensities, etc. This leads to widespread occurrence of prioritized disciplines in digital control system, but not in organizational and technical systems. It results in widespread lack of such effective resource as queuing control optimization and makes the problem of optimization technique development based on qualitative data for complex objects the most urgent one. To solve this problem, the paper describes and analyzes the metamodel of prioritized QS with a large number of request classes. Moreover, the definition is given for crucial qualitative characteristics of CO allowing the potential of service disciplines [1–5, 7, 9, 10] to be estimated and the corresponding steps for their implementation to be undertaken.

2 Standard Cost Model for the Quality of Service

Let us consider QS with n Poisson request flows (classes) having $\lambda_1, \dots, \lambda_n$ intensities. Suppose the request service time of the k^{th} class is characterized by average request service time b_k and a second moment $b_k^{(2)}$, and a penalty on request waiting time unit is c_k nominal units ($k = 1, \dots, n$). The service efficiency is estimated by an average penalty per time unit

$$F = \sum_k \lambda_k c_k w_k, \tag{1}$$

where w_k is steady-state average waiting time of the k^{th} class requests.

If we consider non-prioritized service, time w_k for any k is the same and defined by the Pollaczek-Khinchine formula:

$$w_k = W = \frac{AB^{(2)}}{2(1-R)} = \frac{AB^2(1+V^2)}{2(1-R)} = \frac{A^{-1}R^2(1+V^2)}{2(1-R)}, k = 1, \dots, n, \tag{2}$$

where $A = \lambda_1 + \dots + \lambda_n$ is intensity of cumulative requests flow,
 $R = AB = \rho_1 + \dots + \rho_n < 1$ is an aggregated load coefficient.
 $\rho_k = \lambda_k b_k$ is a system load coefficient of the k^{th} class requests,
 $B = A^{-1} \sum_k \lambda_k b_k$ is average request service time (general),
 $B^{(2)} = A^{-1} \sum_k \lambda_k b_k^{(2)}$ a second moment of request service time,
 V is a c.v. of request service time. It can be derived from the formula

$$V^2 = \frac{A^{-1} \sum_k \lambda_k b_k^{(2)}}{\left(A^{-1} \sum_k \lambda_k b_k \right)^2} - 1. \tag{3}$$

If we look at non-prioritized service, the penalty will be given by the equation

$$F = F_0 = \sum_k \lambda_k c_k w_k = W \sum_k \lambda_k c_k = \frac{\Lambda^{-1} R^2 (1 + V^2)}{2(1 - R)} \sum_{k=1}^n \lambda_k c_k. \tag{4}$$

While using prioritized servicing, each k^{th} flow is given the priority $p_k \in \{1, 2, \dots, n\}$ (different flows have different priorities). The priority p_k of the k^{th} flow will be assumed higher than priority p_i flow, if $p_k > p_i$. The w_k time is given by the following formula for the discipline with relative priorities [7, 10]

$$w_k = w_{1,k} = \frac{\sum_{i=1}^n \lambda_i b_i^{(2)}}{2(1 - R_{k-1})(1 - R_k)} = \frac{\Lambda B^{(2)}}{2(1 - R_{k-1})(1 - R_k)}, \tag{5}$$

where

$$R_k = \sum_{i \in P(k)} \rho_i, \quad R_{k-1} = R_k - \rho_k$$

and the set of indices $P(k)$ includes the numbers of flows with priorities no less than p_k . Using the absolute priority discipline (with afterservicing of interrupted requests):

$$w_k = w_{2,k} = \frac{R_{k-1} b_k}{1 - R_{k-1}} + \frac{\sum_{i \in P(k)} \lambda_i b_i^{(2)}}{2(1 - R_{k-1})(1 - R_k)}. \tag{6}$$

The penalty $F = F_1$ at relative priorities and the penalty $F = F_2$ at absolute priorities are defined by formula (1) when w_k are given by formulas (5) and (6) correspondingly. To construct techniques allowing comparison of prioritized discipline efficiency by means of qualitative data, further, we will formulate and explore a corresponding QS metamodel.

3 Metamodel of Prioritized QS

3.1. Let us define a prioritized QS as the triplet

$$S = \langle \Omega, \alpha, \gamma \rangle, \tag{7}$$

where Ω are parameters of CO, α is server processing speed, $\gamma \in \{0, 1, 2\}$ is an index of service discipline ($\gamma = 0$ indicates non-prioritized service, $\gamma = 1$ is for relative priorities discipline and $\gamma = 2$ is for absolute priorities discipline).

We define parameters Ω as the quintuplet

$$\Omega = \langle n, \mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{V}, \mathbf{C} \rangle, \tag{8}$$

where n is the number of request classes,

$\mathbf{\Lambda} = (\lambda_1, \dots, \lambda_n)$ are intensities of requests flows corresponding to $1, \dots, n$, classes,

$\Psi = (\psi_1, \dots, \psi_n)$ are average service labor-intensities (volumes) for requests of classes $1, \dots, n$,

$\mathbf{V} = (v_1, \dots, v_n)$ are the c.v. for volumes of requests belonging to $1, \dots, n$, classes,

$\mathbf{C} = (c_1, \dots, c_n)$ are penalties of the classes per request waiting time unit.

An arrival flow of any class is Poisson one by default.

Obviously, with processing speed α given in (7), the volume vector Ψ determines all $b_k = \psi_k/\alpha$ and, along with vector \mathbf{V} , all $b_k^{(2)} = (1 + v_k^2)\psi_k^2/\alpha^2$. Processing speed α must lie within the range of $\alpha > \alpha_{\min}$, where α_{\min} depends on the condition of steady-state regime existence $R = \sum_k \lambda_k b_k \leq 1$, from which

$$\sum_k \lambda_k \psi_k / \alpha \leq 1, \alpha \geq \sum_k \lambda_k \psi_k = \alpha_{\min} \text{ are follow. If } \alpha > \alpha_{\min}, \text{ then } R < 1.$$

In order to reveal the general patterns of prioritized service unrelated to exact values of $\Lambda, \Psi, \mathbf{V}$ and \mathbf{C} parameters, these parameters will be considered as random vectors.

At the first approximation, all scalar components of the same vector can be considered as having the same probability distribution and being independent continuous nonnegative random variables (r.v.). For instance, in this case, vector Λ consists of n independent r.v. $\lambda_1, \dots, \lambda_n$ described by common probability density function (pdf) $f_\lambda(t)$. Similarly, vectors Ψ, \mathbf{V} and \mathbf{C} can be specified by pdf $f_\psi(t), f_v(t)$ and $f_c(t)$ respectively. In a number of promising cases, one can introduce dependencies between the parameters of the same request class. We impose the condition $0 \leq v_k \leq 2$ on the v_k components of vector \mathbf{V} . As a typical pdf $f_v(t)$, one will use the triangular one in the interval $(0, 2)$ having the mode in the $t = 1$ point.

We will look for estimates of prioritized service efficiency (when $\gamma > 0$) under an optimal prioritization condition providing the minimum penalty for waiting. Characteristics ξ_1 and ξ_2 for efficiency of using relative and absolute (with afterservicing) priorities are denoted by

$$\xi_1 = F_0/F_1 \text{ and } \xi_2 = F_0/F_2, \tag{9}$$

As functions of random vectors, characteristics ξ_1 and ξ_2 represent r.v. To explore the mathematical expectations (m.e.) of $\bar{\xi}_1, \bar{\xi}_2$ and other characteristics, we will use the simulation along with analytical methods. In the latter case, when n is stated, one realization of $\Lambda, \Psi, \mathbf{V}$ and \mathbf{C} parameters defines the corresponding realization of characteristics $W, w_{1,k}, w_{2,k}, F_0, F_1, F_2, \xi_1$ and ξ_2 which represent the functions of processing speed α in the range $\alpha > \alpha_{\min}$, i.e., for all $R < 1$. Averaging the values of the characteristics $W, w_{1,k}, w_{2,k}, F_0, F_1, F_2, \xi_1$ and ξ_2 over the set of realizations of $\Lambda, \Psi, \mathbf{V}$ and \mathbf{C} parameters, we get estimates of their m.e. $\bar{W} = \bar{W}(R), \dots, \bar{\xi}_1 = \bar{\xi}_1(R)$ and $\bar{\xi}_2 = \bar{\xi}_2(R)$.

The parameter n in (8) represents a variable that may take sufficiently large values in complex CO.

3.2. From the definition of ξ_1 , taking into account formulas (1), (2) and (5), we conclude that it is independent of \mathbf{V} :

$$\xi_1 = \frac{F_0}{F_1} = \frac{W \sum_k \lambda_k c_k}{\sum_k \lambda_k c_k w_{1,k}} = \frac{\frac{AB^{(2)}}{2(1-R)} \sum_{k=1}^n \lambda_k c_k}{\sum_{k=1}^n \lambda_k c_k \frac{AB^{(2)}}{2(1-R_{k-1})(1-R_k)}} = \frac{\sum_{k=1}^n \lambda_k c_k}{\sum_{k=1}^n \frac{(1-R)\lambda_k c_k}{(1-R_{k-1})(1-R_k)}} \tag{10}$$

Analysis of the latter expression in (10) shows that for any constant $h > 0$ at fixed R (defined by corresponding value α) the replacement of vector $\mathbf{C} = (c_1, \dots, c_n)$ by vector $h\mathbf{C} = (hc_1, \dots, hc_n)$ does not lead to changes of characteristic ξ_1 . Thus, ξ_1 is invariant to scale transformations of the vector \mathbf{C} . Consequently, in terms of ξ_1 all pdf $f_c(t)$, coinciding up to the accuracy of their scale transformations, are equivalent. Similar statement holds for vectors $\mathbf{\Lambda}$, $\mathbf{\Psi}$ and their pdf $f_\lambda(t)$, $f_\psi(t)$. One of the possible interpretations of r.v. $\mathbf{\Lambda}$, $\mathbf{\Psi}$, \mathbf{C} scale transformations and corresponding scale transformations of pdf $f_\lambda(t)$, $f_\psi(t)$, $f_c(t)$ is the changing of units for measuring time, volume and cost. Without doubt, it has no effect on dimensionless characteristic ξ_1 .

The analysis of ξ_2 leads to a similar conclusion apart from it depending on \mathbf{V} (which is also dimensionless). Further we consider the invariance of ξ_1 and ξ_2 with respect to scale transformations of pdf $f_\lambda(t)$, $f_\psi(t)$ and $f_c(t)$ by normalizing (whenever possible) the pdf under the condition of $E(\lambda) = E(\psi) = E(c) = 1$, when their m.e. are equal to one (such independent normalizing of λ , ψ and c , generally, takes place when they are statistically independent). Expression for ξ_2 has the following form

$$\xi_2 = \frac{\frac{\Lambda^{-1}R^2(1+V^2)}{2(1-R)} \sum_k \lambda_k c_k}{\sum_{k=1}^n \lambda_k c_k \left[\frac{R_{k-1}\psi_k/\alpha}{1-R_{k-1}} + \frac{\sum_{i \in P(k)} \lambda_i \psi_i^2(1+v_i^2)/\alpha^2}{2(1-R_{k-1})(1-R_k)} \right]} \tag{11}$$

The proposed metamodel of multiflow QS allows us to find the key parameters of complex CO and to optimize algorithms of real-time operations control in incomplete information terms. The metamodel is based on the cost model of request service [6, 7, 9].

The triangular pdf $f_v(t)$ in $(0 \leq t \leq 2)$ having the mode in the $t = 1$ point is usually used in metamodel.

The form of sums in (3), (10) and (11), similar to the form of moment estimation expressions that are defined by vectors (as by samples) $\mathbf{\Lambda}$, $\mathbf{\Psi}$, \mathbf{V} and \mathbf{C} , enable us to relate characteristics ξ_1 and ξ_2 with moments of λ , ψ , v and c random variables. Such relations open new and surprising opportunities for finding key parameters of CO that allow us to evaluate the feasibility of using prioritized disciplines under conditions of absence of complete information that is necessary for applying the classical results of prioritized discipline theory [2, 3, 5, 6, 10].

The feasibilities of using the proposed metamodel for obtaining nontrivial analytical results will be demonstrated by following example detecting service

features when the number n of request classes is large. Let all c_k from the formula of penalty for non-prioritized systems $F_0 = W \sum_k \lambda_k c_k$ are equal to one. Then, considering (2), and also the definition of intensity Λ and Little's formula [7], we obtain $F_0 = W\Lambda = L = (1 + V^2)R^2/(1 - R)/2$, where L – the average length of aggregated request queue. From this it follows that, values of F_0 and L with fixed R can be arbitrary large if the c.v. V is arbitrary large.

4 Region with Wide Variations of Average Labor-Intensities

Let us demonstrate the technique for detecting CO key parameters by means of the metamodel. If all $c_k = 1$ in the expression (4) of penalty F_0 , then

$$F_0 = \frac{\Lambda^{-1}R^2(1 + V^2)}{2(1 - R)} \sum_{k=1}^n \lambda_k = \frac{\Lambda^{-1}R^2(1 + V^2)}{2(1 - R)} \Lambda = \frac{R^2(1 + V^2)}{2(1 - R)}. \tag{12}$$

Further, in terms of the metamodel it is easy to establish that penalty F_0 for any fixed load R may be arbitrary large although all c.v. v_k are bounded by narrow limits (for example, $0 < v_k < 2, k = 1, \dots, n$).

Actually, considering the components of vectors $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ as independent samples from $f_\lambda(t)$ and $f_\psi(t)$ distributions when $n \rightarrow \infty$, taking into account (3), we get

$$V^2 + 1 = \frac{\Lambda^{-1} \sum_k \lambda_k b_k^{(2)}}{\left(\Lambda^{-1} \sum_k \lambda_k b_k\right)^2} \geq \frac{\Lambda^{-1} \sum_k \lambda_k b_k^2}{\left(\Lambda^{-1} \sum_k \lambda_k b_k\right)^2} = \frac{\frac{\Lambda}{n} \frac{1}{n} \sum_k \lambda_k \psi_k^2}{\left(\frac{1}{n} \sum_k \lambda_k \psi_k\right)^2} \rightarrow \frac{E(\lambda)E(\lambda\psi^2)}{E^2(\lambda\psi)},$$

from this it follows that the probability of holding inequality:

$$V^2 + 1 \geq \frac{E(\lambda)E(\lambda\psi^2)}{E^2(\lambda\psi)}, \text{ or } V^2 \geq \frac{E(\lambda)E(\lambda\psi^2)}{E^2(\lambda\psi)} - 1 \tag{13}$$

converges to one with growing n . When λ and ψ are independent it can be simplified:

$$V^2 \geq \frac{E(\lambda)E(\lambda)E(\psi^2)}{E^2(\lambda)E^2(\psi)} - 1, V^2 \geq \frac{E(\psi^2)}{E^2(\psi)}, V^2 \geq \frac{E(\psi^2) - E^2(\psi)}{E^2(\psi)} = v_\psi^2,$$

i.e., inequality (13), certain in the limit when $n \rightarrow \infty$, takes the form:

$$V^2 \geq v_\psi^2, \tag{14}$$

where v_ψ is c.v. of distribution $f_\psi(t)$ of average request volumes. From (14) and (12) when $n \rightarrow \infty$ it follows that if $v_\psi^2 \rightarrow \infty$, then $V^2 \rightarrow \infty$ and $F_0 \rightarrow \infty$.

The ranges of the values for vectors $\mathbf{\Lambda}$, $\mathbf{\Psi}$ and \mathbf{C} , where $F_0 \rightarrow \infty$ when $n \rightarrow \infty$, will be referred to as the critical region of QS parameters. The critical region $v_\psi^2 \rightarrow \infty$ established a moment before will be called a region with wide variations of average labor-intensities.

5 Region with Inverse Power-Law Dependence λ and ψ

Now let us consider the case when λ and ψ are dependent. Let the components of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ vectors be known as related by the formula $\lambda_k \psi_k = A$, where $A > 0$ is some constant (it is the situation when ρ_k are equal). Then $\lambda = A/\psi$ in (13), and with great probability for large n the following inequality will be fulfilled

$$V^2 \geq \frac{E(A/\psi)E((A/\psi)\psi^2)}{E^2((A/\psi)\psi)} - 1 = \frac{A^2E(\psi^{-1})E(\psi)}{A^2} - 1 = E(\psi^{-1}) - 1. \tag{15}$$

Here the normalizing condition $E(\psi) = 1$ is taken into account. But $E(\psi^{-1}) = \int_0^\infty t^{-1} f_\psi(t) dt$, and, consequently, if pdf is continuous on the right and $f_\psi(t) = f_\psi(0) > 0$ with $t = 0$, then $E(\psi^{-1}) = \infty$. Therefore, (15) implies that relation $\lambda_k \psi_k = A$ leads to unlimited increase of V with growing n , and, thus, to unlimited increase of F_0 , for a wide variety of distributions $f_\psi(t)$. Accordingly, the relation $\lambda_k \psi_k = A$ determines one more critical region.

In reality, this region is far more wider and involves the relations in the form of $\lambda_k = A\psi_k^{-\beta}$ with $\beta \in (1, 3)$. Indeed, when $\lambda_k = A\psi_k^{-\beta}$ the inequality (13) can be expressed in the form

$$V^2 \geq \frac{E(A\psi^{-\beta})E(A\psi^{-\beta}\psi^2)}{E^2(A\psi^{-\beta}\psi)} - 1 = \frac{E(\psi^{-\beta})E(\psi^{2-\beta})}{E^2(\psi^{1-\beta})} - 1 = G(\beta) - 1. \tag{16}$$

Considering for simplicity only the case of distributions $f_\psi(t)$ specified in finite interval $0 \leq t \leq g$ and assuming pdf $f_\psi(t) = f_\psi(0) > 0$ with $t = 0$ as continuous on the right, from (16) we derive:

- when $\beta \leq 0$ all m.e. $E(\psi^{-\beta})$, $E(\psi^{1-\beta})$ and $E(\psi^{2-\beta})$ are positive and finite, $G(\beta)$ is finite, and, therefore the estimate (16) does not cause the unlimited increase of V^2 when $n \rightarrow \infty$;
- when $0 < \beta < 1$ the m.e. $E(\psi^{1-\beta})$ and $E(\psi^{2-\beta})$ are positive and finite, $E(\psi^{-\beta}) = \int_0^g t^{-\beta} f_\psi(t) dt = \lim_{\varepsilon \downarrow 0} \int_\varepsilon^g t^{-\beta} f_\psi(t) dt$ is also positive and finite, therefore the estimate (16) here does not cause the unlimited increase of V^2 when $n \rightarrow \infty$; however, when $\beta \uparrow 1$ we get $E(\psi^{-\beta}) \rightarrow \infty$, $G(\beta) \rightarrow \infty$, i.e., when $n \rightarrow \infty$ and $\beta \uparrow 1$ the estimate (16) causes the unlimited increase of V^2 ;
- when $1 \leq \beta < 2$ the m.e. $E(\psi^{1-\beta})$ and $E(\psi^{2-\beta})$ are positive and finite, however, $E(\psi^{-\beta}) = \lim_{\varepsilon \downarrow 0} \int_\varepsilon^g t^{-\beta} f_\psi(t) dt = \infty$, therefore $G(\beta) \rightarrow \infty$, and, thus, the estimate (16) causes the unlimited increase of V^2 when $n \rightarrow \infty$;
- when $2 \leq \beta < 3$ the m.e. $E(\psi^{2-\beta})$ is positive and finite, but $E(\psi^{-\beta}) = \infty$ and $E(\psi^{1-\beta}) = \infty$. Eliminating the ∞/∞ indefinite form that arises in (16) for $G(\beta)$, we have $G(\beta) = \infty$, and, consequently, the estimate (16) causes the unlimited increase of V^2 when $n \rightarrow \infty$;
- when $\beta \geq 3$ the m.e. $E(\psi^{-\beta})$, $E(\psi^{1-\beta})$ and $E(\psi^{2-\beta})$ are equal to infinity and eliminating the indefinite form in (16) gives the finite value for $G(\beta)$.

Therefore, considering the critical region, i.e., the region with inverse power-law dependence λ and ψ , involves the relations in the form of $\lambda_k = A\psi_k^{-\beta}$ with $\beta \in [1, 3)$. In practice, according to simulation modeling, similar inverse power-law dependence with a high probability already results in large values of V^2 and F_0 at the number of flows n within ten, even though the dependence describes (approximately) the relation just between parameters λ_k and ψ_k .

6 Region of Stochastic Inverse Proportionality λ and ψ

Under sufficiently general conditions, the conclusion about the unlimited increase of V^2 with growing n , is also extended to the cases of stochastic relations $\lambda_k = x/\psi_k$ where independent r.v. $x > 0$.

Let, for example, r.v. x and ψ be independent and uniformly distributed in $(0, 2)$ then pdf $f_\lambda(t)$ of r.v. $\lambda = x/\psi$ has the following form:

$$f_\lambda(t) = \int_0^\infty u f_x(tu) f_\psi(u) du = \begin{cases} \frac{1}{2} \int_0^2 u f_\psi(u) du, & \text{if } t \leq 1 \\ \frac{1}{2} \int_0^{1/t} u f_x(tu) du, & \text{if } t > 1, \end{cases} = \begin{cases} \left. \frac{u^2}{8} \right|_0^2 = \frac{1}{2}, & t \leq 1, \\ \left. \frac{u^2}{8} \right|_0^{1/t} = \frac{1}{2t^2}, & t > 1. \end{cases}$$

Power-law asymptotics of pdf $f_\lambda(t)$ with index of power “2” causes the infinite m.e. $E(\lambda)$. Inserting $E(\lambda) = \infty$, $E(\lambda\psi^2) = E(\lambda\psi\psi) = E(x\psi) = E(x)E(\psi) = 1$ and $E^2(\lambda\psi) = E^2(x) = 1$ into inequality (13) we obtain inequality

$$V^2 \geq \infty,$$

resulting in unlimited increase of V^2 and F_0 when $n \rightarrow \infty$. When independent x and ψ have exponential distribution with parameters μ_1 and μ_2 , we obtain a similar result. In this case

$$\begin{aligned} f_\lambda(t) &= \int_0^\infty u f_x(tu) f_\psi(u) du = \int_0^\infty u \mu_1 e^{-\mu_1 tu} \mu_2 e^{-\mu_2 u} du \\ &= \frac{\mu_1 \mu_2}{\mu_1 t + \mu_2} \int_0^\infty u (\mu_1 t + \mu_2) e^{-(\mu_1 t + \mu_2) u} du \\ &= \frac{\mu_1 \mu_2}{\mu_1 t + \mu_2} \int_0^\infty u (\mu_1 t + \mu_2) e^{-(\mu_1 t + \mu_2) u} du = \frac{\mu_1 \mu_2}{(\mu_1 t + \mu_2)^2}. \end{aligned}$$

Here we used the integration by parts formula. The integration bounds having been substituted, the indefinite form $0 \times \infty$ was eliminated. Again, we obtained for r.v. $\lambda = x/\psi$ the pdf with power-law tail, with $E(\lambda) = \infty$, and again $V^2 \rightarrow \infty$, $F_0 \rightarrow \infty$ when $n \rightarrow \infty$.

7 Region with Correlating λ and c Having Large c.v

Using the metamodel, we will show the existence of one more critical region of parameters. Let the growth factors of c.v. V for F_0 in (4) with fixed R be absent with growing n (suppose, for example, $V = \text{const}$ for any n). Then from (4) we have

$$\begin{aligned}
 F_0 &= \frac{\Lambda^{-1}R^2(1+V^2)}{2(1-R)} \sum_{k=1}^n \lambda_k c_k = \frac{R^2(1+V^2)}{2(1-R)} \frac{1}{\Lambda/n} \frac{1}{n} \sum_{k=1}^n \lambda_k c_k \\
 &\rightarrow \frac{R^2(1+V^2)}{2(1-R)} \frac{E(\lambda c)}{E(\lambda)} = \frac{R^2(1+V^2)}{2(1-R)} E(\lambda c), \tag{17}
 \end{aligned}$$

(where arrow for r.v. means convergence in probability). If λ and c are independent, then $E(\lambda c) = E(\lambda)E(c) = 1$. In this case, from (17) we state that the penalty F_0 with growing n converges in probability to a constant defined by fixed R and V . But for dependent λ and c , from the following formula

$$r(\lambda, c) = \frac{E(\lambda c) - E(\lambda)E(c)}{\sigma_\lambda \sigma_c} = \frac{E(\lambda c) - 1}{\sigma_\lambda \sigma_c} = \frac{E(\lambda c) - 1}{v_\lambda v_c}, \tag{18}$$

we have

$$E(\lambda c) = 1 + r(\lambda, c)v_\lambda v_c, \tag{19}$$

where $r(\lambda, c)$ is the correlation coefficient of λ and c ;
 σ_λ, σ_c are mean squared deviations of λ and c ;
 v_λ, v_c are variation coefficients of λ and c .

From (19) and (17) it follows that penalty F_0 can be growing unlimitedly with growing n when λ and c are positively correlated and v_λ and/or v_c are equal to infinity. For example, if $\lambda = Ac$ we have $r(\lambda, c) = 1$ and $v_\lambda = v_c$ (here $A > 0$ is some constant). In practice, v_λ and v_c may be arbitrary large, and then with growing n the F_0 will also be increasing due to the growth of multiplier $\Lambda^{-1} \sum_{k=1}^n \lambda_k c_k$ that converges to $M(\lambda c) = 1 + r(\lambda, c)v_\lambda v_c$.

Notice that the positive correlation between flow intensity and penalty per waiting time unit are quite natural.

8 Analysis of c/b Rule Efficiency in Absolute Prioritizing. Technique of the Best Transpositions

It is known that relative prioritizing is optimal when priorities p_k of the k^{th} flow are increasing with the order of c_k/b_k characteristic growth ("rule c/b "). The c/b rule which is also called as $c\mu$ -rule was proposed firstly in [9]. In the case of absolute priorities, the c/b rule is optimal only when the service times have the exponential distributions, but it can be used with general distributions of service times as a fairly good heuristics.

To estimate the heuristics c/b acceptability, we compare it to a more precise absolute prioritizing rule. Realization of the precise rule (calculation of the

penalty F_2 for all $n!$ appointments of priorities and choosing the best appointment) for n , which is several orders of magnitudes, is near-impossible. Therefore, further, an approximate optimization technique, i.e. the best transpositions technique (TT) having the effectively constrained enumeration of priority appointment, is proposed and applied. Consider the step-by-step description of the technique.

- Step 1.** For given $\Lambda, \Psi, \mathbf{V}, \mathbf{C}$ when $\alpha = 2(\lambda_1\psi_1 + \dots + \lambda_n\psi_n)$, i.e., when $R = 0.5$ all $b_k = \psi_k/\alpha$ and $b_k^{(2)} = (1 + v_k^2)\psi_k^2/\alpha^2$ are calculated. The absolute priorities are appointed according to c/b rule and the penalty F_2 is calculated.
- Step 2.** Among all pairs of n flows, one must find such pair (k, i) that the exchanging (transposition) of priorities p_k and p_i in this pair will lead to the greatest penalty reduction.
- Step 3.** If not a single transposition leading to penalty F_2 reduction has been found, proceed to step 4. Otherwise, perform the found priority transposition, and calculate the penalty F_2 for the found appointment, then go to step 2.
- Step 4.** Obtained appointment is taken as the result of optimization.

Estimation of c/b efficiency consist of the multiple random realization of parameters $\Lambda, \Psi, \mathbf{V}$ and \mathbf{C} , optimization of priority appointment for each realization with TT and comparison of obtained penalty F_2 values with those obtained with the c/b heuristics.

The Results of Comparing TT and the Heuristics c/b . Parameters $\Lambda, \Psi, \mathbf{V}, \mathbf{C}$ were repeatedly generated from different sets of pdf $f_\lambda(t), f_\psi(t), f_v(t), f_c(t)$ with different n , the rule c/b ing compared to TT. We get the following results.

1. The c/b heuristics on average gives the result that can be improved (i.e. the penalty F_2 can be reduced) just by (0.1–1.5)% with TT.
2. In particular cases, the result of using c/b euristics can be improved by 24% and more by means of TT.
3. In any case, priority distribution obtained by c/b rule with TT is not deteriorated, therefore, TT can be applied in all cases of absolute priority distributions.
4. When the number of flows n is around 15–20, the TT almost always improves the results of using c/b heuristics.
5. The characteristics ξ_1 and ξ_2 of penalty reduction in critical regions of CO parameters increase manifold.

The latter result suggests that question about feasibility of prioritized service in critical regions of CO parameters in many practical cases should have a positive answer.

9 Analysis of Feasibility of Prioritized Disciplines in Critical Regions of Control Object Parameters

Using the metamodel, hundreds of thousands of multi-flow QS were generated and the characteristics ξ_1 and ξ_2 of prioritized service efficiency with optimal

prioritizing were defined. Consider the most important results obtained in the current research.

1. The penalties F_0, F_1, F_2 increase unlimitedly with growing n (with fixed $R, f_\lambda, f_\psi, f_v, f_c$) in critical regions of parameters, whereas in non-critical regions the growth of penalties is bounded from above.
2. At the intersections of critical regions, the rate of F_0, F_1 and F_2 penalties growth with growing n may be increased by several orders of magnitude.
3. Characteristics ξ_1 and ξ_2 increase with growing n , but are bounded from above (both in non-critical region, and critical regions of CO parameters).
4. Applying the prioritized disciplines when $R \rightarrow 0$ is not relevant, when $R \rightarrow 1$ it is marginally useful.

Statistical experiments show that determined key parameters $\mathbf{\Lambda}$ of CO, usually, affect queuing control efficiency far greater than the form of distributions for parameter vectors does. In view of this, the experimental data given below are for the exponential pdf f_λ, f_ψ, f_c and the triangular f_v one.

The asymptotic properties found above for $n \rightarrow \infty$ began to reveal themselves even at small n . As an example, the average values of L (average length of aggregated request queue), V^2 and ξ_1, ξ_2 for independent components of $\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{V}$ and \mathbf{C} vectors and for two kinds of relation between λ_k and ψ_k are given in Table 1. These results are obtained under optimal prioritizing condition.

Table 1. The dependence of efficiency parameters of prioritized disciplines on n

n	Independent parameters				$\lambda_k = A/\psi_k$				$\lambda_k = A/(\psi_k)^{1.2}$			
	\bar{L}	\bar{V}^2	$\bar{\xi}_1$	$\bar{\xi}_2$	\bar{L}	\bar{V}^2	$\bar{\xi}_1$	$\bar{\xi}_2$	\bar{L}	\bar{V}^2	$\bar{\xi}_1$	$\bar{\xi}_2$
2	0.79	2.18	1.23	3.31	3.22	11.9	1.30	3.79	2.28	8.11	1.27	1.82
10	0.97	2.87	1.50	4.70	5.50	21.0	1.60	11.1	11.8	46.5	1.57	5.62
20	1.01	3.04	1.54	4.90	6.79	26.1	1.65	14.5	16.4	64.7	1.63	8.98
50	1.05	3.19	1.57	4.98	7.13	27.5	1.71	15.2	22.7	89.7	1.71	16.4
100	1.06	3.25	1.59	5.21	7.65	29.6	1.76	17.5	28.8	114.3	1.78	32.3

Conditions, under which the numerical results are obtained, are stated in Table 1 and illustrated by Figs. 1, 2 and 3. Straight lines on the Figs. 1, 2 and 3 are trend lines derived from least square method. Figures 2 and 3 also illustrate the equations of power-law regression. The straightness of lines for relevant power-law equations can be explained by logarithmic scale of both coordinate axes. Figure 4 shows changes (with growing n) of characteristics ξ_2 for efficiency absolute prioritizing discipline, which correspond to the conditions illustrated above. Therefore, considering vectors of flow parameters as samples from distributions of r.v., we will find arithmetic mean of the parameters, their squares, products for parameters of different vectors, obtaining thereby the values of relevant sample moments. For example, the numerator of (3) contains arithmetic

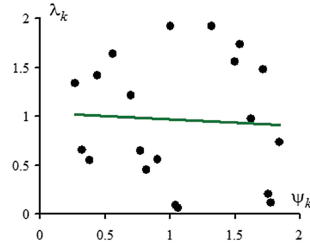


Fig. 1. Elements of intensities vector Λ and labor-intensities vector Ψ are almost independent (the first set of columns in Table 1)

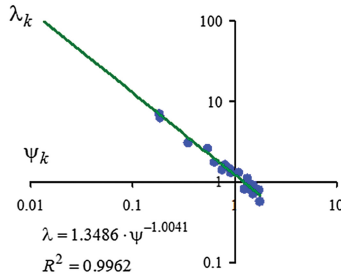


Fig. 2. Elements of intensities vector Λ are inversely proportional to the elements of labor-intensities vector Ψ (the second set of columns in Table 1)

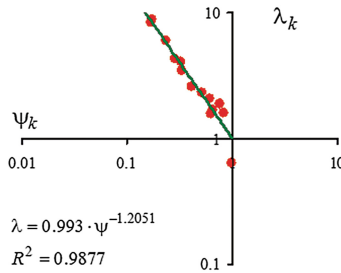


Fig. 3. Elements of intensities vector Λ and labor-intensities vector Ψ are connected with power-law dependence $\lambda_k = A/(\psi_k)^{1.2}$ (the third set of columns in Table 1)

mean of all b_k^2 and the denominator contains arithmetic mean of all b_k . The derived sample moments are the approximate values of expressions included in the right parts of (3), (10) and other formulas determining the feasibility of applying the prioritized service disciplines. Considering the parameters of multi-flow system as r.v. we simplify the calculation of characteristics for efficiency and at early stages of CO designing we can make a reasoned decision on the appropriateness of applying various prioritized disciplines. At the same time, it is not necessary to make an exact measurement of all parameters for a large number of request flows, it is sufficient to obtain just some representative sample

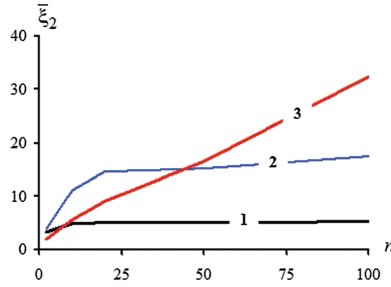


Fig. 4. Dependence of the characteristics ξ_2 for efficiency absolute prioritizing discipline on n (the curves 1–3 are obtained for the cases illustrated by Figs. 1, 2 and 3)

of parameters. To define weather the sample belongs to one of the considered in the article critical region of parameters space or not, the high precision of the of measurements is not required.

Practically-wide limiting case of the considered multi-flow system is the case when the number of flows goes to infinity so that instead of discrete probability distribution $q_k = \lambda_k/\Lambda$, which implies the belonging of some request to the k^{th} class, it is reasonable to use a continuous distribution with density $q(\lambda)$. Here, the service time density of any request is the corresponding continuous mixture of service time densities of requests from the different classes. For example, if the service time densities are exponential and shifted to the right on $K > 0$, then the density $q(\lambda)$ is exponential, but the service time of any request will have the Pareto distribution, i.e., the distribution with heavy tail. The investigation of QS with heavy tails belong to the relevant special section of Queueing Theory [12–14]. Multi-flow systems can be considered as a class of systems between classical QS and QS with heavy tails. In particular, as was shown above, considering multi-flow systems we have to deal with unlimited c.v. of service time. Such c.v. is typical for QS with heavy tails.

10 Conclusion

Operation of large-scale objects under current conditions despite the continuous increase of automation level is characterized by substantial damage caused by delayed or wrong reaction of CO in normal or emergency regimes. Practical application of the metamodel proposed in the article enables estimating the economic impact which can be obtained by optimization of service disciplines. Thus, average values of characteristics ξ_1 and ξ_2 calculated in Table 1 show how many times the damage might be reduced by using corresponding disciplines. Information aggregated in industry databases allow the metamodel to be adapted and formalized as a well-structured industry-based mathematical model.

The condition of effective application relative and absolute priorities is moderate load of server (which is not close to zero or one).

Using the metamodel one can calculate objective importance estimates for a number of qualification characteristics of operating and dispatching personnel (percentage of actions performed correctly, wrong or with delay and etc. [11]). Estimates of importance allow optimizing the scenarios of academic studies by computer simulators [8]. The results derived from the application of metamodel enable formulating the recommendations for designers of automated control systems. The optimal prioritizing technique for appointment relative and absolute priorities is developed when the distributions of service times differ from exponential, with the “rule c/b ” tested on a large amount of metamodel realizations with different values of the key parameters. It is established that in solutions with c/b rule the characteristic ξ_2 differ from the optimal one by average (0.1–1.5)%. At the same time, in particular cases, the result of using c/b uristics can be improved by 24% and more with the developed technique.

In general, critical regions identified in the article for the parameters CO are the regions with the most effective application of relative and absolute priorities. The obtained results allow the feasibility of applying the prioritized service disciplines at early stages of CO designing to be estimated with using the sample estimates of request flow intensities of different classes, average service labor-intensities (volumes) and their variation coefficients, and sample estimates of penalties per waiting time unit in these classes.

References

1. Rykov, V.: Controllable queueing systems: from the very beginning up to nowadays. *Control. Queueing Syst. RT&A* **12**, 31–38 (2017). 2(45)
2. Nazarov, A.A.: *Controlled Queueing Systems and Their Optimization*. Tomsk Univ. Publ., Tomsk (1984). (in Russian)
3. Moder, J.J., Elmaghraby, S.E.: *Handbook of Operations Research (in 2 vol.)*. vol. 1. *Foundations and Fundamentals Hardcover* (1978)
4. Wentzel, E.S.: *Operations research. Tasks, principles, methodology*. Drofa (2004). (in Russian)
5. Vishnevskiy, V.M.: *Theoretical bases of designing computer networks*. Technosphere, Moscow (2003). (in Russian)
6. Maiorov, S.A., et al.: *Fundamentals of Computing Systems Theory*. Higher School, Moscow (1978). (in Russian)
7. Kleinrock, L.: *Queueing Systems: V. II Computer Applications*. Wiley Interscience, New York (1976)
8. Merkuriev, G.V.: *Operational-Dispatching Management of Power Systems*. Center for the Training of Energy Personnel, St. Petersburg (2002). (in Russian)
9. Bronstein, O.I., Rykov, V.V.: On optimal priority rules in queueing systems. *Izv. AS USSR. Techn. Cibern.* **6**, 2837 (1965). (in Russian)
10. Jaiswall, N.: *Queues with Priority*. Mir, Moskow (1973)
11. Budovsky, V.P.: Providing the reliable operators work subjects of operational dispatch control in emergency situations in the power system. In: *Operational Management in the Energy Sector. Personnel Training and Maintenance of their Qualification*. 4, p. 1121 (2006). (in Russian)
12. Zwart, A.P.: *Queueing Systems with Heavy Tails*. Eindhoven University of Technology (2001)

13. Zadorozhnyi, V.N., Zakharenkova, T.R.: Methods of simulation queueing systems with heavy tails. In: Dudin, A., Gortsev, A., Nazarov, A., Yakupov, R. (eds.) ITMM 2016. CCIS, vol. 638, pp. 382–396. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44615-8_33
14. Zadorozhnyi, V.N., Zakharenkova, T.R.: Minimization of packet loss probability in network with fractal traffic. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) ITMM 2017. CCIS, vol. 800, pp. 168–183. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_14



Analysis of an Infinite-Server Queue $MAP_k|G_k|\infty$ in Random Environment with k Markov Arrival Streams and Random Volume of Customers

K. Kerobyan¹, R. Kerobyan², and K. Enakoutsa¹(✉)

¹ Department of Mathematics, California State University, Northridge, 18111 Nordroff Street, Northridge, CA 91330, USA

² Department of Computer Science, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
{khanik.kerobyan,koffi.enakoutsa}@csun.edu

Abstract. In this paper the model of an infinite-server $MAP_k|G_k|\infty$ queue in random environment with catastrophes is considered. The transient and limiting probability generating functions (PGF) of joint distributions of number of busy servers and numbers of served customers, the joint distributions of total resource in the model and total served resource are found. All results are obtained by using collective marks method (CMM) and renewal processes.

Keywords: Markov arrival process · Infinite-server queue model
Random environment · Catastrophe · Resource vector

1 Introduction

The distinguishing characteristic of New Generation Networks (NGN) is the integration of heterogeneous resources, applications, technologies, customers and data into one united information infrastructure which is ubiquitous and accessible anytime and anywhere. This integration process includes all layer of NGN and makes its QoS metrics guaranteeing more challenging [1]. To solve the NGN optimal design and performance providing problems the methods of statistical simulation and modeling are widely used. However the application of these methods and tools even for several elements of NGN (for example a protocol, a server, or a canal) is complicated because of networks statistical processes (traffic and service) nature [2]. As shown by large number of measurements the traffic of modern IP networks can be characterized by the heterogeneity, the non-stationarity, the burstiness, the short-range and the long-range dependence. These factors make the modeling and performance evaluation of modern networks more challenging [3]. Network traffics in queueing theory are generally described finite Markov Processes traffic models: Markov Arrival Process

(MAP), Batch Markov Arrival Process (BMAP), Marked Markov Arrival Process (MMAP) and their generalizations [4]. The MAP arrivals properties and their applications are presented in [5, 6] and are not duplicated here. To evaluate the network canals performance main parameters: capacity, delay and packet loss probability, the infinity server queue models are widely used.

An infinite-server model $M|SM|\infty$ with a Poisson arrival process and with semi-Markovian (SM) service times. The transient and asymptotic results for PGF of queue-length process is obtained by means of CMM [7]. The infinite-server queue $BM_k|G_k|\infty$ with k correlated heterogeneous customers in a batch is studied in [8]. In steady state, the joint PGF of queue length of customers is derived by using CMM and conditional expectations. The generalization of the model for the queue $BMAP_k|G_k|\infty$ with structured batch arrival of k types of customers is considered in [9]. In steady state, the differential equations for PGF of queue length and its moments are obtained. The first and second order asymptotes of queue length for the models $MAP|G|\infty$, $MMPP|G|\infty$, $G|G|\infty$ based on supplementary variable method are studied in [10].

To evaluate the impact of network environment on networks performance metrics the infinite-server models in the random environment (RE) are applied. The queue size distribution of the model $M|G|\infty$ with semi-Markov (SM) environment under asymptotic condition of high arrival rate and frequent environment transitions is studied in [11]. By the method of supplementary variable and the original method of dynamic screening first and second order asymptotes of queue size distribution are obtained. The queue $M|G|\infty$ in random environment with clearing mechanism is studied in [12]. The environmental clearing process is modeled by an m -state irreducible SMP. The transient and steady-state queue length distributions by using renewal arguments are obtained. The $MMAP_k|G_k|\infty$ queue in SM environment and catastrophes is studied in [13]. The PGFs of joint distributions of queue size and number of served customers by using renewal arguments and differential equations are found.

In many applications of queue models such as computer and communication networks, transportation systems customers characterize by the vector of requesting resources which components can be random quantities. Despite the importance of the queueing models and their applications, there are very few works devoted to research of queue resource models, see e.g. [13–19].

The main methods to study the infinite-server queues are: supplementary variables method [20], the method which based on properties of exponential distribution and conditional expectations [21], and Collective Marks Method (CMM) [22, 23]. The last method is also is called “supplementary event or catastrophes” method [22] and has been used successfully for queue models with priorities [23]. CMM have been used in [8] for infinite-server model with Poisson arrival of batches. In [7] the method is mentioned but does not used to obtain some results.

In this paper we consider some generalizations of [12, 13, 15] results for infinite-server $MAP_k|G_k|\infty$ queue. The PGFs of joint distributions of number of busy servers and numbers of served customers, the joint distributions of total

resource in the model and total served resource are found. All results are obtained by using CMM and renewal processes methods.

2 Model Description

We consider an infinite server $MAP_k|G_k|\infty$ queue model in RE with K types of customers and catastrophes. The RE operates according to stationary, irreducible semi-Markov process (SMP) $\xi(t)$, $t \leq 0$ with finite state space $S = \{1, 2, \dots, k\}$. The SMP is given by the vector of initial distribution $p^0 = \{p_0^i, i \in S\}$ and SM matrix $Q(t) = \|Q_{ij}(t)\|$, $t \geq 0, 1, j \in S$. Different type of customers arrive according to homogeneous MAP which is given by the characteristic matrices $\{D_{0r}, D_{1r}, 1 \leq r \leq K\}$. Here D_{0r} is a non-singular matrix with negative diagonal elements and D_{1r} is non-negative matrix corresponding to the type r of customers. The phase process (PP) $J_r(t)$ of MAP is an irreducible Markov process (MP) with generating matrix D_r and finite set of states E_r . D_r is a matrix of $m_r \times m_r$ size.

$$D_r = D_{0r} + D_{1r}, D_r e_r = 0, \pi_r D_r = 0, \pi_r e_r = 1, 1 \leq r \leq K,$$

where e_r is the unit column vector, π_r is the vector of stationary distribution $\pi_r = (\pi_1, \dots, \pi_{m_r})$ of PP $J_r(t)$.

Instead of independent MAPs we will consider the superposed arrival process. It is known [6] that superposition of finite number of MAPs is a MAP as well. Let m be $\prod_{r=1}^K m_r$. To distinguish arrivals of one type from the others, we introduce the following $m \times m$ matrices:

$$D_0 = D_{01} \oplus D_{02} \oplus \dots \oplus D_{0K}, D_r = I_1 \otimes \dots \otimes I_{r-1} \otimes D_{1r} \otimes I_{r+1} \otimes \dots \otimes I_K, \\ r = 1, \dots, K,$$

$$D = D_0 + D_1, D e = 0, \pi D = 0, \pi e = 1, \pi = \pi_1 \otimes \pi_2 \otimes \dots \otimes \pi_K,$$

where \otimes (resp. \oplus) denotes the Kronecker product (resp. the Kronecker sum) and I_r denotes the identity matrix of order m_r , π is a vector of size m .

For superposed MAP (SMAP) the stationary arrival rate of customers is given by

$$\lambda = \sum_{r=1}^K \pi D_r e = \sum_{r=1}^K \lambda_r, \lambda_r = \pi D_r e$$

where λ_r is the stationary arrival rate of type r customers.

The service of arriving customers begins immediately. Let the random variable (r.v.) τ_r be a service time of type r customers, and $\tau = (\tau_1, \tau_2, \dots, \tau_k)$ is a vector of service times. The components of τ are independent, identically distributed (i.i.d.) r.v.s which depend on type of the customer and state of environmental SMP. R.v. τ_r has $B_r(t) = P(\tau_r < t)$ general distribution and finite mean

value $\bar{\tau}_r$, $1 \leq r \leq K$. Each arriving and departing r type customer is characterized by k -dimensional random volume (resource) vectors $\xi_r = (\xi_{1r}, \dots, \xi_{kr})$ and $\sigma_r = (\sigma_{1r}, \dots, \sigma_{kr})$ resp., which have non-negative components $1 \leq r \leq K$. Let $F_r(\mathbf{x}) = P(\xi_{1r} \leq x_1, \dots, \xi_{kr} \leq x_k)$ and $G_r(\mathbf{x}) = P(\sigma_{1r} \leq x_1, \dots, \sigma_{kr} \leq x_k)$ be the joint distributions of resource vectors ξ_r and σ_r , where $\mathbf{x} = (\xi_1, \dots, \xi_k)$. We assume that the service time vector τ and the resource vectors $\xi_r = (\xi_{1r}, \dots, \xi_{kr})$, $\sigma_r = (\sigma_{1r}, \dots, \sigma_{kr})$ are mutually independent.

When SMP $\xi(t)$, $t \leq 0$ jumps from state i to the state r all customers in the model are instantly flashed out and the model jumps into empty state. Let consider the related with SMAP counting processes (CP) $N(t), N_s(t), M(t) : N(t) = (N_1(t), N_2(t), \dots, N_K(t)), N_s(t) = (N_{1s}(t), N_{2s}(t), \dots, N_{Ks}(t)), M(t) = (M_1(t), M_2(t), \dots, M_K(t))$, where $N_r(t)$ and $M_r(t)$ are the number of type r customers arriving and serving in time interval $[0, t)$, and $N_{rs}(t)$ is the number of customers being in service at moment t . Let $\beta(t) = (\beta_1(t), \dots, \beta_K(t))$ and $\alpha(t) = (\alpha_1(t), \dots, \alpha_K(t))$ be the vectors of total resource served during interval $[0, t)$ and accumulated in the model at moment t , resp. The components of $\beta(t)$, $\alpha(t)$, $N_s(t)$ and $M_s(t)$ vectors are defined as:

$$\beta_r(t) = \sum_{i=1}^{M_r(t)} \sigma_{ri}, \quad \alpha_r(t) = \sum_{i=1}^{N_r(t)} \zeta_{ri}.$$

Suppose that at initial time $t = 0$ the model is empty, $N(0) = 0, M(0) = 0$.

3 The Counting Process

Let consider the CP $\{N(t), J(t), t \leq 0\}$ with the matrix $P(n, t)$, $n = (n_1, \dots, n_k)$ of transition probabilities: $P(n, t) = \|p_{ij}(n, t)\|$, $p_{ij}(n, t) = P(N(t) = n, J(t) = j | J(0) = i, 1 \leq i, j \leq m$.

Let define the following generating functions (GF) $D(z), P(z, t)$,

$$P(z, t) = \sum_{n \geq 0} z^n P(n, t), \quad D(z) = D_0 + \sum_{h \in C^0} z_r D_r, \quad |z_r| \leq 1, \quad 0 \leq r \leq K, |z| \leq 1,$$

where $z = (z_1, z_2, \dots, z_k)$ and $z^n = (z_1^{n_1}, z_2^{n_2}, \dots, z_k^{n_k})$.

Theorem 1. The PGF $P(z, t)$ of counting process $\{N(t), J(t), t \geq 0\}$ satisfies the basic differential equation

$$\frac{\partial}{\partial t} P(z, t) = D(z)P(z, t), \quad |z| \leq 1, \tag{1}$$

with initial conditions $P(z, 0) = 1$. The solution of differential equation Eq.(1) is given by

$$P(z, t) = e^{D(z,t)t}. \tag{2}$$

Proof. The transition probabilities $\{P(n, t), n \leq 0\}$ of CP $N(t)$ satisfy the following Kolmogorov backward differential equations

$$\frac{d}{dt}P(n, t) = P(n, t)D_0 + \sum_{r=1}^K D_r P(n - e_r, t), \quad n \geq 0,$$

with initial condition $P_i(n, 0) = 0, n > 0, P_i(0, 0) = 1, i = 1, 2, \dots, k$, where $e_r = (0, \dots, 0, 1, 0, \dots, 0)$ is a vector with 1 in r^{th} position. Pre-multiplying each equation by corresponding z^n after summation we get differential equations for PDF $P(z, t)$. The solution of this equation in matrix exponential form is given by Eq. (2).

4 Thinning MAP

Let consider the following Bernoulli thinning process (TP) of MAP. Each type r customer which arrives at the time t can join the main stream by probability $p_r(t)$ and can be ignored by probability $1 - p_r(t)$.

Lemma. The thinned process is a MAP which counting process has matrix GF $D_T(z, t)$ and PGF $P_T(z, t)$, which are defined as follow

$$D_T(z, x) = D_0 + \sum_{r=1}^K D_r [1 - p_r(x) + z_r p_r(x)],$$

$$P_T(z, t) = e^{\int_0^t D_T(z, x) dx}.$$

The first part of lemma is a consequence of more general result for Markov-additive processes of arrivals [21]. The resulting thinned MAP characteristic matrices are

$$D_{T0} = D_0 + \sum_{r=1}^K D_r [1 - p_r(x)], \quad D_{Tr} = D_r p_r(x), \quad r = 1, 2, \dots, K.$$

The subject of our interest is the joint distribution

$$P(n, m, x, y, t) = P(N_s(t) = n, M(t) = m, \alpha(t) \leq x, \beta(t) \leq y).$$

5 Model Analysis

Let suppose that the environmental SMP $\xi(t), t \geq 0$ is in state $i \in S$ and consider the dynamic of the the model during time interval $[u, t)$. Each type r customer arriving at moment u will be in service at moment t by probability $1 - B_{ri}(t - u)$ and will finish its service before moment t by probability $B_{ri}(t - u)$.

Let $A_{jk}^i(n, m, x, y, u, t)$ be the joint probability that n customers are in service at moment t , and m customers are already served in $[u, t)$, total resources in the

model at time t is $\alpha(t) \leq x$ and the total served resource during interval $[u, t]$ is $\beta(t) \leq y$, PP $J(u)$ is in phase $i \in E$ under condition that at initial moment $t = 0$ the model was empty, and PP $J(0)$ was in phase $k \in E$.

Let denote by $\tilde{A}^i(z_1, z_2, s_1, s_2, u, t)$ the matrix which elements are Laplace–Stieltjes transformation (LST) and z transformation of $A_{jk}^i(z_1, z_2, s_1, s_2, u, t)$ and by $\tilde{F}_{ri}(s_1), \tilde{G}_{ri}(s_2)$ denote the LST of $F_{ri}(x)$ and $G_{ri}(y)$. For homogeneous model we have $\tilde{A}^i(z_1, z_2, s_1, s_2, t) = \tilde{A}^i(z_1, z_2, s_1, s_2, u, t)$, e.g. see [22].

$$\tilde{F}_{ri}(s_1) = \int_0^\infty e^{-s_1x} dF_{ri}(x), \quad \tilde{G}_{ri}(s_2) = \int_0^\infty e^{-s_2y} dG_{ri}(y),$$

$$\tilde{A}^i(z_1, z_2, s_1, s_2, t) = \sum_{n=0}^\infty \sum_{m=0}^\infty z_1^n z_2^m \int_0^\infty \int_0^\infty e^{-s_1x - s_2y} A^i(n, m, dx, dy, t),$$

with $|z_1| \leq 1, |z_2| \leq 1$. By using CMM we can prove the following result.

Theorem 2. The PGF of the model $\text{MAPr|Gr|}\infty \tilde{A}^i(z_1, z_2, s_1, s_2, t)$ satisfy the following basic differential and integral equations

$$\tilde{A}^i(z_1, z_2, s_1, s_2, t) = \exp\left\{\int_0^t [D_0(i) + \tilde{S}_i(z_1, z_2, s_1, s_2, u)] du\right\}, \quad |z_1| \leq 1, |z_2| \leq 1,$$

where

$$\tilde{S}_i(z_1, z_2, s_1, s_2, t) = \sum_{r=1}^K D_{ri} \left[z_{2r} \tilde{G}_{ri}(s_2) B_{ri}(t) + z_{1r} \tilde{F}_{ri}(s_1) (1 - B_{ri}(t)) \right].$$

Proof. The queueing process in the model can be considered as a CP of some special TP. According to Theorem 1 the CP of that TP has matrix exponential form for every state of environmental SMP. Let define the rate of that matrix exponential function by using CMM. Suppose that a customer of type r arrives at moment u with rate D_r . This customer will be served up to moment t with probability $B_r(t - u)$ or will be in the model at moment t with probability $1 - B_r(t - u)$. Let mark each served type r customer red by $z_{2r} \tilde{G}_r(s_2)$ or blue by $1 - z_{1r} \tilde{F}_r(s_1)$ with probabilities. Alike, we mark each serving in the model type r customer red by $z_{1r} \tilde{F}_r(s_1)$ or blue by $1 - z_{1r} \tilde{F}_r(s_1)$ with probabilities. Then $D_r [z_{2r} \tilde{G}_r(s_2) B_r(t - u) + z_{1r} \tilde{F}_r(s_1) (1 - B_r(t - u))]$ is the rate of red type r customers arriving at moment u and the common rate of red (all types) customers arriving at moment u is $\tilde{S}_i(z_1, z_2, s_1, s_2, u)$. Finally, the common rate of red (all types) customers arriving in $[0, t]$ is

$$\int_0^t [D_0(i) + \tilde{S}_i(z_1, z_2, s_1, s_2, u)] du.$$

Recall that $D_0(i)t$ is no customer arrivals rate in $[0, t]$ interval.

Theorem 3. The PGF $\tilde{A}^i(z_1, z_2, s_1, s_2, t)$ satisfy the following basic differential and integral equations

$$\begin{aligned} \tilde{A}^i(z_1, z_2, s_1, s_2, t) &= e^{D_0(i)t} \\ &+ \int_0^t e^{D_0(i)u} \tilde{S}_i(z_1, z_2, s_1, s_2, u) \tilde{A}^i(z_1, z_2, s_1, s_2, t-u) du. \end{aligned} \quad (3)$$

$$\frac{\partial}{\partial t} \tilde{A}^i(z_1, z_2, s_1, s_2, t) = [D_0(i) + \tilde{S}_i(z_1, z_2, s_1, s_2, t)] \tilde{A}^i(z_1, z_2, s_1, s_2, t), \quad i \in S, \quad (4)$$

with initial conditions $\tilde{A}^i(z_1, z_2, s_1, s_2, 0) = I$.

When $z_2 = 1, s_2 = 0$ from Eqs. (3) and (4) we obtain the PGF of joint distribution of number of busy servers and total accumulated resources in the model $\tilde{P}(z, s, t, i) = \tilde{A}^i(z_1, \mathbf{1}, s_1, \mathbf{0}, t)$.

6 $MAP_k|G_k|\infty$ Model with Catastrophes

Let consider the general homogeneous Markovian model under influence of SMP generated catastrophes. As in [14], after every transition of environmental SMP the model jumps into the special state, let say 0-state, and then works from that state. When the SMP is in i^{th} state all parameters of the model are related to that state: DF of inter-arrival time of customers, DF and rates of service time of customers, their resource vectors.

Let $\tilde{P}(n, m, s_1, s_2, t, i)$ and $\tilde{\tilde{P}}(n, m, s_1, s_2, t, i)$ defined the LST of probabilities of having in the model $n = (n_1, n_2, \dots, n_k)$ customers at moment t and having $m = (m_1, m_2, \dots, m_k)$ served customers in the interval of time $[0, t)$ when environmental SMP is in state i , for the models without catastrophes and with catastrophes, resp.

The following theorem gives the connection between these two models probabilities.

Theorem 4. The $\tilde{P}(n, m, s_1, s_2, t, i)$ satisfies the following integral equation

$$\begin{aligned} \tilde{P}(n, m, s_1, s_2, t, i) &= (1 - F_i(t)) \tilde{\tilde{P}}(n, m, s_1, s_2, t, i) \\ &+ \sum_{j \in S} \int_0^t \tilde{P}(n, m, s_1, s_2, t-u, j) dQ_{ij}(u), \quad i \in S. \end{aligned} \quad (5)$$

The solution of equations Eq. (4) can be found

$$\begin{aligned} P(n, m, s_1, s_2, t, i) &= (\bar{F}_i(t)) \tilde{\tilde{P}}(n, m, s_1, s_2, t, i) \\ &+ \sum_{j \in S} \int_0^t \bar{F}_j(t-u) \tilde{\tilde{P}}(n, m, s_1, s_2, t-u, j) dH_{ij}(u), \end{aligned} \quad (6)$$

where $\bar{F}(t) = 1 - F(t)$ and $F(t) = \{F_i(t), i \in S\}$ is a sojourn time distribution vector of SMP: $F_i(t) = \sum_{j \in S} Q_{ij}(t), i \in S$.

$H(t) = \|H_{ij}(t)\|$ is a renewal matrix of SMP which components satisfy the following equations

$$H_{ij}(t) = 1 - F_i(t) + \sum_{k \in S} \int_0^t H_{kj}(t - u) dQ_{ik}(u), i, j \in S.$$

The proof of Eqs. (4), (6) can be done by using standard renewal arguments (see for example [22]).

Let $\tilde{P}(z_1, z_2, s_1, s_2, t, i)$ and $\tilde{\tilde{P}}(z_1, z_2, s_1, s_2, t, i)$ be the PGFs of $\tilde{P}(n, m, s_1, s_2, t, i)$ and $\tilde{\tilde{P}}(n, m, s_1, s_2, t, i)$, respectively. Then we get from Eqs. (4), (6)

$$\begin{aligned} \tilde{P}(z_1, z_2, s_1, s_2, t, i) &= (1 - F_i(t))\tilde{\tilde{P}}(z_1, z_2, s_1, s_2, t, i) \\ &+ \sum_{j \in S} \int_0^t \tilde{P}(z_1, z_2, s_1, s_2, t - u, j) dQ_{ij}(u), i \in S, \end{aligned} \tag{7}$$

which solution is

$$\begin{aligned} \tilde{P}(z_1, z_2, s_1, s_2, t, i) &= \bar{F}_i(t)\tilde{\tilde{P}}(z_1, z_2, s_1, s_2, t, i) \\ &+ \sum_{j \in S} \int_0^t \bar{F}_j(t - u)\tilde{\tilde{P}}(z_1, z_2, s_1, s_2, t - u, j) dH_{ij}(u). \end{aligned} \tag{8}$$

The proof can be done by using CMM or standard renewal argument (see for example [23]). Let consider the proof by CMM. We mark each type r customer in the model red or blue color with probabilities z_r and $1 - z_r$ resp. Then left side of Eq. (7) is a probability of event “in the model with catastrophes there are no blue customers at moment and the SMP is in state”. This event can be realized either “the SMP is in state i at moment t given that at initial time $t = 0$ it was in state i and in the model without catastrophes there are no blue customers in interval $[0, t)$ ” and in the model there are no blue customers at moment or “SMP jumps from state i into state j in $[u, du)$ ”, $u \leq t$ (with probability $dQ_{ij}(u)$), the model jumps into 0-state and from that state “in the model are not blue customers in interval” (with probability $P(z, t - u, j)$). In the proof of Eq. (8) we have to paraphrase second term of right side: “SMP jumps into state ji in $[u, du)$ ”, $u \leq t$ (with probability $dH_i(u)$), the model jumps into 0 state and “SMP will stay in the state j during interval of time $[t - u, t)$ and in the model without catastrophes starting from 0-state (empty state) are not blue customers in interval of time $[t - u, t)$ ” (with probability $\bar{F}_j(t - u)\tilde{\tilde{P}}(z_1, z_2, s_1, s_2, t - u, j)$.”

Theorem 5. The distributions of $\tilde{P}(n, m, s_1, s_2, t, i)$ and $\tilde{P}(z_1, z_2, s_1, s_2, t, i)$ have the limits

$$\begin{aligned} \tilde{P}(n, m, s_1, s_2) &= \lim_{t \rightarrow \infty} \sum_{i \in S} p_i^0 \tilde{P}(n, m, s_1, s_2, t, i) \\ &= \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \int_0^\infty (1 - F_j(u)) \tilde{\tilde{P}}(n, m, s_1, s_2, u, j) du, \quad i \in S, \end{aligned}$$

$$\begin{aligned} \tilde{P}(z_1, z_2, s_1, s_2) &= \lim_{t \rightarrow \infty} \sum_{i \in S} p_i^0 \tilde{P}(z_1, z_2, s_1, s_2, t, i) \\ &= \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \int_0^\infty (1 - F_j(u)) \tilde{\tilde{P}}(z_1, z_2, s_1, s_2, u, j) du, \quad i \in S, \end{aligned}$$

where

$$\begin{aligned} \bar{\eta}_i &= \int_0^\infty (1 - F_i(u)) du, \quad q_i = \frac{\bar{\eta}_i \rho_i}{\sum_{r \in S} \bar{\eta}_r \rho_r}, \quad \sum_{r \in S} q_r = 1, \quad \rho_i = \sum_{r \in S} p_{ri} \rho_r, \\ &\sum_{r \in S} \rho_r = 1, \quad p_{ri} = Q_{ri}(\infty), \quad r, i \in S. \end{aligned}$$

Let $\tilde{f}(s)$ denote the Laplace Transformation of a function $f(x)$, $\tilde{f}(s) = \int_0^\infty e^{-su} f(u) du$.

Corollary. When $F_i(t) = 1 - e^{-v_i t}$, $i \in S$ then for LT of $\tilde{P}(n, m, s_1, s_2, s, i)$, PGF $\tilde{P}(z_1, z_2, s_1, s_2, s, i)$ and their limiting values we get

$$\begin{aligned} \tilde{P}(n, m, s_1, s_2, s, i) &= \tilde{\tilde{P}}(n, m, s_1, s_2, s + v_i, i) \\ &+ \frac{1}{s} \sum_{j \in S} \tilde{\tilde{P}}(n, m, s_1, s_2, s + v_j, j) \tilde{H}_{ij}(s), \quad i \in S, \end{aligned}$$

$$\begin{aligned} \tilde{P}(z_1, z_2, s_1, s_2, s, i) &= \tilde{\tilde{P}}(z_1, z_2, s_1, s_2, s + v_i, i) \\ &+ \frac{1}{s} \sum_{j \in S} \tilde{\tilde{P}}(z_1, z_2, s_1, s_2, s + v_j, j) \tilde{H}_{ij}(s), \quad i \in S, \end{aligned}$$

$$\tilde{P}(z_1, z_2, s_1, s_2) = \sum_{i \in S} q_i v_i \tilde{\tilde{P}}(z_1, z_2, s_1, s_2, v_i, i),$$

$$\tilde{P}(n, m, s_1, s_2) = \sum_{i \in S} q_i v_i \tilde{\tilde{P}}(n, m, s_1, s_2, v_i, i).$$

Let consider the infinite-server models $MAP_k|G_k|\infty$ with catastrophes.

Assume that $\tilde{P}(z_1, s_1, t)$ is a LST of PGF joint distribution of number of busy servers and total accumulated resources in the model at moment t given that at $t = 0$ SMP is in state i for the model with catastrophes and $\tilde{P}(z_1, s_1)$ is its limit value.

Theorem 6. The PGFs $\tilde{P}(z_1, s_1, t)$ and its limit value $\tilde{P}(z_1, s_1)$ are given by

$$\tilde{P}(z_1, s_1, t) = \sum_{i \in S} p_i^0 \tilde{P}(z_1, s_1, t, i),$$

where $\{p_i^0, i \in S\}$ is a vector of initial distribution of SMP, and $\tilde{P}(z_1, s_1, t, i)$ satisfies the following integral equations

$$\tilde{P}(z_1, s_1, t, i) = \bar{F}_i(t) e^{\int_0^t [D_0(i) + \tilde{S}_i(z_1, s_1, u)] du} + \sum_{r \in S} \int_0^t \tilde{P}(z_1, s_1, t-u, r) dQ_{ir}(u), \quad i \in S,$$

which solutions are

$$\begin{aligned} \tilde{P}(z_1, s_1, t, i) &= \bar{F}_i(t) e^{\int_0^t [D_0(i) + \tilde{S}_i(z_1, s_1, u)] du} \\ &+ \sum_{r \in S} \int_0^t \bar{F}_r(t-u) e^{\int_0^{t-u} [D_0(r) + \tilde{S}_r(z_1, s_1, x)] dx} dH_{ir}(u), \end{aligned} \tag{9}$$

$$\begin{aligned} \tilde{P}(z_1, s_1) &= \lim_{t \rightarrow \infty} \sum_{i \in S} p_i^0 \tilde{P}(z_1, s_1, t, i) \\ &= \sum_{i \in S} \frac{q_i}{\bar{\eta}_i} \int_0^\infty \bar{F}_i(x) e^{\int_0^x [D_0(i) + \tilde{S}_i(z_1, s_1, u)] du} dx, \quad i \in S. \end{aligned}$$

Corollary. From above model when catastrophes occur according to Poisson distribution with parameter ν then by Eq.(9) we get the results for homogeneous model [24].

For example, for $\tilde{P}(z_1, s_1)$ and $\tilde{P}(z_1, s_1, s)$ we obtain

$$\tilde{P}(z_1, s_1, s) = \tilde{\tilde{P}}(z_1, s_1, s + \nu) \left[1 + \frac{\nu}{s} \right], \quad \tilde{P}(z_1, s_1) = \nu \tilde{\tilde{P}}(z_1, s_1, \nu), \tag{10}$$

where

$$\begin{aligned} \tilde{\tilde{P}}(z_1, s_1, \nu) &= \int_0^\infty e^{-\nu x} e^{\int_0^x [D_0(i) + \tilde{S}_i(z_1, s_1, u)] du} dx, \\ \tilde{\tilde{P}}(z_1, s_1, s + \nu) &= \int_0^\infty e^{-(\nu+s)x} e^{\int_0^x [D_0(i) + \tilde{S}_i(z_1, s_1, u)] du} dx. \end{aligned}$$

This result can be interpreted by the CMM. Let consider flow of event A which has exponentially distributed inter event time with parameter s . First, let Eq. (10) write in the form $s\tilde{P}(z_1, s_1, s) = \tilde{\tilde{P}}(z_1, s_1, s + \nu)[s + \nu]$, then it can be interpreted as follow: $s\tilde{P}(z_1, s_1, s)$ is a probability of the event event A appears when in the model with catastrophes no blue customers at moment but it is the same as happens the event sum of the event A and catastrophes appears when in the model without catastrophes no blue customers at moment. The probability of this event is $\tilde{\tilde{P}}(z_1, s_1, s + \nu)[s + \nu]$.

Remark. Let consider the model $MAPr|Gr|\infty$ in which after transition of environmental SMP and flashing out of all customers the model jumps into the recovery station. The recovery time ϑ of the model has general distribution $U(t) = P(\vartheta \leq t)$ with finite mean value $\bar{\vartheta}_1$. To define the model performance metrics we can use formulated for above $MAPr|Gr|\infty$ model theorems. Let $T(t) = ||T_{ij}(t)||$ is a SM matrix of new model SMP with state space S . The elements of $T(t)$ SM matrix $T_{ij}(t)$ are define as convolution of $Q_{ij}(t)$ and DF $U(t)$: $T_{ij}(t) = Q_{ij}(t) * U(t)$, $i, j \in S$. For example the transient probabilities of the model $\tilde{P}(n, s_1, t, i)$ and $\tilde{P}(z_1, s_1, t, i)$ satisfy the following integral equations

$$\tilde{P}((n, s_1, t, i) = (1 - T_i(t))\tilde{\tilde{P}}((n, s_1, t, i) + \sum_{j \in S} \int_0^t \tilde{P}((n, s_1, t - u, j) dT_{ij}(u), \quad i \in S,$$

$$\tilde{P}(z_1, s_1, t, i) = (1 - T_i(t))\tilde{\tilde{P}}(z_1, s_1, t, i) + \sum_{j \in S} \int_0^t \tilde{P}(z_1, s_1, t - u, j) dT_{ij}(u), \quad i \in S,$$

where $T_i(t) = \sum_{j \in S} T_{ij}(t)$, $i \in S$.

The corresponding limiting values are given by

$$\tilde{P}(n, s_1) = \lim_{t \rightarrow \infty} \sum_{i \in S} p_i^0 \tilde{P}(n, s_1, t, i) = \sum_{j \in S} \frac{q_j^*}{\bar{\eta}_j + \bar{\vartheta}_1} \int_0^\infty (1 - F_j(u)) \tilde{\tilde{P}}(n, s_1, u, j) du,$$

$$\begin{aligned} \tilde{P}(z_1, s_1) &= \lim_{t \rightarrow \infty} \sum_{i \in S} p_i^0 \tilde{P}(z_1, s_1, t, i) \\ &= \sum_{j \in S} \frac{q_j^*}{\bar{\eta}_j + \bar{\vartheta}_1} \int_0^\infty (1 - F_j(u)) \tilde{\tilde{P}}(z_1, s_1, u, j) du, \quad i \in S \end{aligned}$$

where $q_i^* = \frac{(\bar{\eta}_i + \bar{\vartheta}_1)\rho_i}{\sum_{r \in S} (\bar{\eta}_r + \bar{\vartheta}_1)\rho_r}$, $\sum_{r \in S} q_r^* = 1$, $\frac{q_j^*}{\bar{\eta}_j + \bar{\vartheta}_1}$.

7 Performance Measures of the Model

Let $\omega_{1r}(t)$, ω_{1r} and $Var_{1r}(t)$, Var_{1r} denote the transient and steady state mean and variance of queue length of type r customers. Then for $\omega_{1r}(t)$ and ω_{1r} we get

$$\omega_{1r}(t) = \lim_{s_1 \rightarrow 0} \tilde{\omega}_{1r}(s_1, t), \quad \tilde{\omega}_{1r}(s_1, t) = \left. \frac{\partial \tilde{P}(z_1, s_1, t)}{\partial z_{1r}} \right|_{z_{1r}=1, z_{11}=z_{12}=\dots=z_{1k}=1},$$

$$\omega_{1r} = \lim_{s_1 \rightarrow 0} \tilde{\omega}_{1r}(s_1), \quad \tilde{\omega}_{1r}(s_1) = \left. \frac{\partial \tilde{P}(z_1, s_1)}{\partial z_{1r}} \right|_{z_{1r}=1, z_{11}=z_{12}=\dots=z_{1k}=1},$$

$$\tilde{\omega}_{1r}(s_1, t) = \sum_{i \in S} p_i^0 (1 - F_i(t)) \tilde{\omega}_{1r}(s_1, t, i)$$

$$+ \sum_{j \in S} \int_0^t (1 - F_j(t - u)) \tilde{\omega}_{1r}(s_1, t - u, j) dH_j(u),$$

$$\tilde{\omega}_{1r}(s_1) = \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \int_0^\infty (1 - F_j(u)) \tilde{\omega}_{1r}(s_1, u, j) du,$$

where $\tilde{\omega}_{1r}(s_1, t, i)$ is a transient mean queue length of r type customers of the model without catastrophes when the SMP is in state i . The transient and steady state variance of queue length for type r customers $Var_{1r}(t)$ and Var_{1r} we get

$$Var_{1r}(t) = \omega_{2r}(t) + \omega_{1r}(t)(1 - \omega_{1r}(t)), \quad Var_{1r} = \omega_{2r} + \omega_{1r}(1 - \omega_{1r})$$

$$\omega_{2r}(t) = \lim_{s_1 \rightarrow 0} \tilde{\omega}_{2r}(s_1, t), \quad \tilde{\omega}_{2r}(s_1, t) = \left. \frac{\partial^2 \tilde{P}(z_1, s_1, t)}{\partial z_{1r}^2} \right|_{z_{1r}=1, z_{11}=z_{12}=\dots=z_{1k}=1},$$

$$\omega_{2r} = \lim_{s_1 \rightarrow 0} \tilde{\omega}_{2r}(s_1), \quad \tilde{\omega}_{2r}(s_1) = \left. \frac{\partial^2 \tilde{P}(z_1, s_1)}{\partial z_{1r}^2} \right|_{z_{1r}=1, z_{11}=z_{12}=\dots=z_{1k}=1},$$

$$\tilde{\omega}_{2r}(s_1, t) = \sum_{i \in S} p_i^0 (1 - F_i(t)) \tilde{\omega}_{2r}(s_1, t, i)$$

$$+ \sum_{j \in S} \int_0^t (1 - F_j(t - u)) \tilde{\omega}_{2r}(s_1, t - u, j) dH_j(u),$$

$$\tilde{\omega}_{2r}(s_1) = \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \int_0^\infty (1 - F_j(u)) \tilde{\omega}_{2r}(s_1, u, j) du.$$

Let $\delta_r(t)$, δ_r , $r = 1, 2, \dots, K$, be the transient and steady-state mean values of accumulated type r resources in the model and δ be a total accumulated resources in the model.

$$\delta_r(t) = \pi \delta_r(t) e, \quad \delta_r(t) = \lim_{s_1 \rightarrow 0} \left. \frac{\partial \tilde{P}(z_1, s_1, t)}{\partial s_{1r}} \right|_{z_{11}=z_{12}=\dots=z_{1k}=1},$$

$$\delta_r(t) = \sum_{i \in S} p_i^0 (1 - F_i(t)) \bar{\delta}_r(t, i) + \sum_{j \in S} \int_0^t (1 - F_j(t - u)) \bar{\delta}_r(t - u, j) dH_j(u),$$

$$\delta_r = \lim_{t \rightarrow \infty} \delta_r(t), \quad \delta_r = \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \lambda_{jr} \bar{c}_{1r}(j) \int_0^\infty \int_0^u (1 - F_j(u))(1 - B_{jr}(x)) dx du,$$

where $\bar{\delta}_r(t, j) = \lambda_{jr} \bar{c}_{1r}(j) \int_0^t (1 - B_{jr}(x)) dx$, $\bar{c}_{1r}(j)$ is the mean value of DF $C_{jr}(t)$.

$$\delta = \sum_{r=1}^K \delta_r = \sum_{r=1}^K \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \lambda_{jr} \bar{c}_{1r}(j) \int_0^\infty \int_0^u (1 - F_j(u))(1 - B_{jr}(x)) dx du.$$

If L_{losr} denote the steady state mean number of destroyed type r customers, then

$$L_{losr} = \lim_{s_1 \rightarrow 0} \pi \tilde{L}_{losr}(s_1) e,$$

where $\tilde{L}_{losr}(s_1) = \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \int_0^\infty \tilde{\omega}_{1r}(s_1, u, j) dF_j(u)$,

$$L_{losr} = \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \lambda_{jr} \int_0^\infty \int_0^u (1 - B_{jr}(x)) dx dF_j(u).$$

If L_{los} is the steady-state total mean number of destroyed customers of all types, then

$$L_{los} = \sum_{r=1}^K L_{losr} = \sum_{r=1}^K \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \lambda_{jr} \int_0^\infty \int_0^u (1 - B_{jr}(x)) dx dF_j(u).$$

Let L_{qr} and L_q be the steady state mean number of type r and all types customers in the model. Then

$$L_{qr} = \pi \tilde{\omega}_{1r} e = \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \int_0^\infty (1 - F_j(u)) \pi \tilde{\omega}_{1r}(u, j) e du =$$

$$\sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \lambda_{jr} \int_0^\infty \int_0^u (1 - F_j(u))(1 - B_{jr}(x)) dx du,$$

$$L_q = \sum_{r=1}^K L_{qr} = \sum_{r=1}^K \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \lambda_{jr} \int_0^\infty \int_0^u (1 - F_j(u))(1 - B_{jr}(x)) dx du.$$

Suppose that MAP is defined by following matrices $D_0(i) = -\alpha_i I$, $D_r(i) = \alpha_{ir} I$, $r = 1, 2, \dots, K$, $i \in S$, where I is an identity matrix. Then for $\bar{P}(n, t, i)$, $P(n, t)$, $P(n)$, $\omega_{1r}(t, i)$ and L_{losr} we obtain

$$\tilde{\bar{P}}(z, s, t, i) = e^{-\int_0^t \sum_{r=1}^k \lambda_{ir}(1 - B_{ir}(x))(1 - z_{ir} \bar{C}_{ir}(s)) dx}, \quad \bar{P}(n, t, i) = \prod_{r=1}^K \frac{a_{ir}(t)^{n_r}}{n_r!} e^{-a_i(t)}$$

$$P(n, t) = \sum_{j \in S} p_j^0 (1 - F_j(t)) \prod_{r=1}^K \frac{a_{ir}(t)^{n_r}}{n_r!} e^{-a_i(t)} + \sum_{j \in S} \int_0^t (1 - F_j(t - u)) \prod_{r=1}^K \frac{a_{ir}(t - u)^{n_r}}{n_r!} e^{-a_i(t - u)} dH_j(u),$$

$$P(n) = \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \int_0^\infty (1 - F_j(u)) \prod_{r=1}^K \frac{a_{jr}(u)^{n_r}}{n_r!} e^{-a_j(u)} du,$$

$$\omega_1(t) = \sum_{j \in S} p_j^0 (1 - F_j(t)) a_j(t) + \sum_{j \in S} \int_0^t (1 - F_j(t - u)) a_j(t - u) dH_j(u),$$

$$\omega_1 = \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \int_0^\infty (1 - F_j(u)) a_j(u) du,$$

$$L_{los} = \sum_{j \in S} \frac{q_j}{\bar{\eta}_j} \int_0^\infty a_j(u) dF_j(u),$$

where $a_{ir}(t) = \alpha_{ir} \int_0^t (1 - B_{ir}(u)) du$, $i \in S$.

8 Conclusion

In this paper we consider the infinite-server $MAP_k|G_k|\infty$ queue in random environment with resource vectors of customers, subject to catastrophes. The transient and steady state PGFs of joint distributions of number of different types customers at moment and number of served different types of customers in interval $[0, t)$, the joint distributions of total resource in the model and total served resource are found. All results are obtained by using CMM and renewal processes methods. The obtained results may be applied for evaluating the performance metrics, as well as for finding the optimal strategies of managing resources for a wide class of computer systems and networks, whereas the queue $MAP_k|G_k|\infty$ may be used as a model.

References

1. Kouvatsos, D.: Network Performance Engineering: A handbook on Convergent Multi-service Networks and Next Generation Internet. Springer, Heidelberg (2011). <https://doi.org/10.1007/978-3-642-02742-0>
2. Tripathi, S.K., Sharma, P., Raghavan, S.V.: Challenges in design of next generation networks. In: Nejat, I.A., Topuz, E. (eds.) Modeling and Simulation Tools for Emerging Telecommunication Networks, pp. 19–42. Springer, Boston (2006). https://doi.org/10.1007/0-387-34167-6_2
3. Ma, S., Ji, C.: Modeling heterogeneous network traffic in wavelet domain. IEEE/ACM Trans. Netw. **9**(5), 125–136 (2001)
4. Paxson, V., Floyd, S.: Wide-area traffic: the failure of Poisson modeling. In: Proceedings of the ACM, pp. 257–268 (1994)
5. Artalejo, J.R., Gomez-Corral, A., He, Q.M.: Markovian arrivals in stochastic modeling: a survey and some new results. SORT **34**(2), 101–144 (2010)
6. He, Q.: Fundamentals of Matrix-analytic Methods. Springer, New York (2014). <https://doi.org/10.1007/978-1-4614-7330-5>
7. Neuts, M.F., Chen, S.Z.: The infinite-server Queue with Poisson arrivals and Semi-Markovian services. Oper. Res, **20**(2), 425–433 (1972)
8. Tong, D.C.: On the $BM/G/\infty$ queue with heterogeneous customers in a batch. J. Appl. Prob. **31**(1), 280–286 (1994)
9. Masuyama, H.: Studies on algorithmic analysis of queues with batch Markovian arrival streams. Ph.D. thesis, Kyoto University (2003)
10. Moiseev, A., Nazarov, A.: Infinite-server Queueing Systems and Networks. Publ. NTL, Tomsk (2015)
11. Linton, D., Purdue, P.: An $M/G/\infty$ queue with m customer types subject to periodic clearing. Opsearch **16**, 80–88 (1979)
12. Nazarov, A., Baymeeva, G.: The $M/G/\infty$ queue in random environment. In: Proceedings of 13th International Conference, ITMM, pp. 312–324 (2014)
13. Kerobyan, K., Enakoutsu, K., Kerobyan, R.: An infinite-server queueing model $MMAP_k|G_k|\infty$ with marked MAP arrival, semi-Markov random environment and subject to catastrophes. In: Proceedings of 8th International Conference, CSIT (2018)
14. Tikhonenko, O.M.: Distribution of the total message flow in group arrival queueing system. Avtomatika i Telemekhanika **11**, 111–120 (1987)
15. Tikhonenko, O.M.: Generalized erlang problem for service systems with finite total capacity. Probl. Inf. Transm. **41**(3), 243–253 (2005)
16. Lisovskaya, E., Moiseeva, S., Pagano, M., Potatueva, V.: Study of the $MMPP/GI/\infty$ queueing system with random customers capacities. Informatika i ee prilozheniya **11**(4), 109–117 (2017)
17. Lisovskaya, E., Moiseeva, S.: Asymptotical analysis of a non-Markovian queueing system with renewal input process and random capacity of customers. Proc. TSU **39**, 30–38 (2017)
18. Kerobyan, Kh., Kerobyan, R.: Transient analysis of infinite-server queue $MMAP_k|G_k|\infty$ with marked MAP arrival and disasters. In: Proceedings of 7th International Conference HET-NETs 2013, November 2013, Ilkley, UK, pp. 11–13 (2013)
19. Naumov, V., Samuylov, K.: On the modeling of queue systems with multiple resources. Proc. RUDN. **3**, 60–63 (2014)
20. Nazarov, A., Terpugov, A.: Queueing Theory. Tomsk State University, NTL (2004)

21. Latouche, G., Ramaswami, V.: Introduction to matrix analytic methods in stochastic modeling. SIAM (1999)
22. Runnenberg, J.: On the use of Collective marks in queueing theory, pp. 339–348. In: Proceedings of the Symp. Cong. UNC, Hill (1965)
23. Gnedenko, B.V., Danielyan, E.A.: Priority service systems. M.: MSU (1973)
24. Breuer, L.: From Markov Jump Processes to Spatial Queues. Springer, NY (2003). <https://doi.org/10.1007/978-94-010-0239-4>



Optimization of Two-Level Discount Values Using Queueing Tandem Model with Feedback

Maria Shklennik^(✉), Svetlana Moiseeva, and Alexander Moiseev

Tomsk State University, Tomsk, Russia

shklennikm@yandex.ru, smoiseeva@mail.ru, moiseev.tsu@gmail.com

Abstract. The trading company model with two levels of discount is considered in the paper. The problem of choosing the optimal discount values is solved. The mathematical model of the company is formulated in the form of an infinite-server queueing tandem with feedback at the second stage. The analytical form of generating function of multi-dimensional joint probability distribution of the number of purchases is obtained. Analytical expressions are found for the mean and variance of the company's profit. Optimal discount values are obtained for the case when the probabilities of repeated purchases linearly depend on the value of discounts.

Keywords: Discount · Infinite-server queue · Feedback · Optimization

1 Introduction

The goal of any trading company is to receive the maximal profit. To achieve this, companies use various methods of sales stimulation. There are a large number of such methods. One of them is a stimulation by reducing the price or in other words, giving a discount. Techniques for implementing discounts are different: sales, discount cards, special promotions. An important issue for the company is the problem of assigning the discount value in order to stimulate the market of the company's products and receive the maximal profit [1–5].

The work of a trading company looks as a sequence of transactions with customers. These transactions are performed at different time moments and the mathematical models of transaction flows are widely used for their analysis [6–8]. The sequences of transactions can be considered as random processes and we may apply the methods of the theory of random processes [9] and the queueing theory [10] to their description and study. We can consider an unlimited number of potential customers of the company. In addition, each customer has the opportunity to reapply to this company. Due to these features, we can use queueing systems with an infinite number of servers and feedback [11–16] as mathematical models of considered processes.

In this paper, we consider the influence of discounts' values on the profit of the trading company that uses two levels of discounts. We solve the problem of

finding such values of discount parameters that ensure maximum profit for the trading company.

The paper is organized as follows. The problem statement and the mathematical model are formulated in Sect. 2. We construct the system of Kolmogorov differential equations for the probabilities of the system states in Sect. 3. Further, the generating function of the process is introduced and the analytic expression is obtained for it. Expressions for the main average characteristics affecting the profit of the trading company are obtained in Sect. 4. Solution of the problem of finding optimal values for discounts of both discount levels for given dependencies of repeated customer flows on these values is presented in Sect. 5. An example of calculating these values for specific parameter values from practice is presented in this section too. In Conclusion, we discuss main results and further studies.

2 Problem Statement and Mathematical Model

Consider the trading company that uses two levels of cumulative discount to stimulate sales of its products. That is, the customer who applied to the company for the first time receives a discount card with a discount $(1 - \delta_1)$ for the second purchase. We denote by r_1 the probability that the customer will return to the store for the second purchase. Obviously, r_1 depends on many factors, including the value of provided discount. If the customer buys goods during the second visit his or her discount card is given a new status and the discount becomes equal to $(1 - \delta_2)$. Also, it is obvious that the condition $\delta_2 < \delta_1$ should take place. The probability of subsequent visits to the store r_2 depends on the value of δ_2 . In this paper, we assume that the discount value does not change after the subsequent purchases. In addition, we suppose that after each purchase, a customer does not need goods of this trading company during some period. The goal of the study is to determine the most profitable conditions for the trading company to carry out this kind of marketing policy. In other words, the problem is to find such values of δ_1 and δ_2 that give the maximal profit.

We use the following queueing system (Fig. 1) as a mathematical model for solving the problem. The system is a tandem with two stages and an infinite number of servers at each stage. Incoming flow of customers (new customers who make purchases for the first time) is a stationary Poisson process with rate λ . A new customer arrives at the first stage of the tandem. He or she stays there for a random time distributed exponentially with parameter μ_1 . This period models the time interval when the customer does not need the goods of the trading company. After the end of this period, the customer may leave the system with probability $(1 - r_1)$ or may go to the second stage of the tandem (return for the second purchase) with probability r_1 . The customer stays at the second stage during a random period distributed exponentially with parameter μ_2 . After that, he or she may go to the second stage again with probability r_2 or may leave the system with probability $(1 - r_2)$. Thus, in terms of queueing theory, we construct a mathematical model of the company in the form of a two-stage infinite-server queueing tandem with feedback at the second stage.

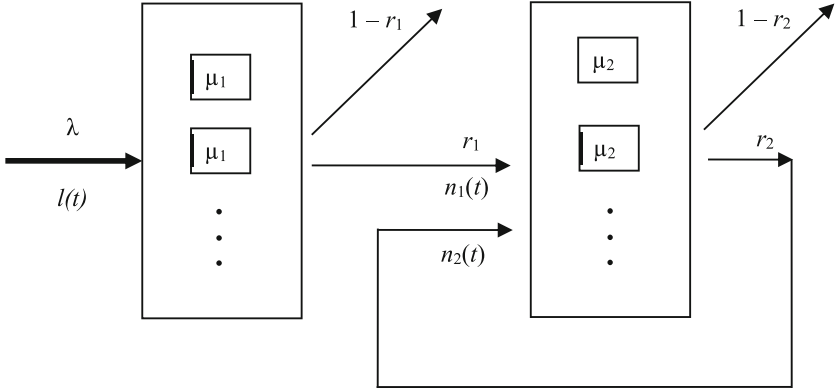


Fig. 1. A model of the company in the form of the queueing system.

Let the initial time moment be $t_0 = 0$. We introduce the following notation:

$i_1(t)$ is the number of customers at the instant t at the first stage of the system,

$i_2(t)$ is the number of customers at the instant t at the second stage of the system,

$l(t)$ is the total number of customers arriving at the first stage from the incoming flow up to the instant t (the total number of primary purchases),

$n_1(t)$ is the total number of customers arriving at the second stage of the system directly from the first stage up to the instant t (the total number of secondary purchases),

$n_2(t)$ is the total number of customers that arrive at the second stage after the second stage up to the instant t (the total number of the third and subsequent purchases).

A profit of the trading company is formed as a result of the purchases of customers. Let us consider the profit from each sale as random variables with identical distributions, then the total number of purchases $l(t) + n_1(t) + n_2(t)$ is crucial for calculating of the total profit from all sales up to the time moment t . Unfortunately, neither the process $l(t) + n_1(t) + n_2(t)$ nor the process $\{l(t), n_1(t), n_2(t)\}$ are Markovian. Therefore, we study the five-dimensional Markov process $\{l(t), n_1(t), n_2(t), i_1(t), i_2(t)\}$.

3 Kolmogorov Differential Equations and Generating Function of the Process

In this section, we study the process $\{l(t), n_1(t), n_2(t), i_1(t), i_2(t)\}$ by constructing and solving Kolmogorov differential equations for its generating function

$$G(z, y_1, y_2, x_1, x_2, t) = \sum_{l=0}^{\infty} \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \sum_{i_1=0}^{\infty} \sum_{i_2=0}^{\infty} z^l y_1^{n_1} y_2^{n_2} x_1^{i_1} x_2^{i_2} P(l, n_1, n_2, i_1, i_2, t). \tag{1}$$

Here $P(l, n_1, n_2, i_1, i_2, t) = \mathbb{P}\{l(t) = l, n_1(t) = n_1, n_2(t) = n_2, i_1(t) = i_1, i_2(t) = i_2\}$. Using an analytical expression for the generating function, you can obtain any characteristics of the process. At the end of the section, we derive the expression for the generation function of the process $\{l(t), n_1(t), n_2(t)\}$ which is under the main interest of the study.

First we solve the auxiliary problem of deriving an expression for the generating function of the process $\{i_1(t), i_2(t)\}$. For the probability distribution of this process $P(i_1, i_2, t) = \mathbb{P}\{i_1(t) = i_1, i_2(t) = i_2\}$, we can write the following equations:

$$\begin{aligned}
 P(i_1, i_2, t + \Delta t) &= P(i_1, i_2, t)(1 - \lambda\Delta t)(1 - i_1\mu_1\Delta t)(1 - \mu_2\Delta t) \\
 &+ P(i_1 - 1, i_2, t)\lambda\Delta t + P(i_1 + 1, i_2 - 1, t)(i_1 + 1)\mu_1\Delta t r_1 + \\
 &P(i_1 + 1, i_2, t)(i_1 + 1)\mu_1\Delta t(1 - r_1) + P(i_1, i_2, t)i_2\mu_2\Delta t r_2 \\
 &+ P(i_1, i_2 + 1, t)(i_2 + 1)\mu_2\Delta t(1 - r_2).
 \end{aligned}$$

for $i_1, i_2 = 0, 1, 2, \dots$. Then we derive the following system of Kolmogorov differential equations:

$$\begin{aligned}
 \frac{\partial P(i_1, i_2, t)}{\partial t} + (\lambda + i_1\mu_1 + i_2\mu_2 - i_2\mu_2r_2)P(i_1, i_2, t) &= \lambda P(i_1 - 1, i_2, t) \\
 + (i_1 + 1)\mu_1r_1P(i_1 + 1, i_2 - 1, t) + (i_1 + 1)\mu_1(1 - r_1)P(i_1 + 1, i_2, t) & \\
 + (i_2 + 1)\mu_2(1 - r_2)P(i_1, i_2 + 1, t) & \tag{2}
 \end{aligned}$$

The initial condition is as follows:

$$P(i_1, i_2, 0) = \begin{cases} 1 & \text{if } i_1 = i_2 = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Consider the generating function of the process $\{i_1(t), i_2(t)\}$

$$g(x_1, x_2, t) = \sum_{i_1=0}^{\infty} \sum_{i_2=0}^{\infty} x_1^{i_1} x_2^{i_2} P(i_1, i_2, t). \tag{3}$$

Substituting this expression into Eq. (2), we obtain the following partial differential equation of the first order for the function $g(x_1, x_2, t)$:

$$\begin{aligned}
 \frac{\partial g(x_1, x_2, t)}{\partial t} + \mu_1(x_1 - 1 + r_1(1 - x_2)) \frac{\partial g(x_1, x_2, t)}{\partial x_1} & \\
 + \mu_2(1 - r_2)(x_2 - 1) \frac{\partial g(x_1, x_2, t)}{\partial x_2} = \lambda(x_1 - 1)g(x_1, x_2, t) & \tag{4}
 \end{aligned}$$

with the initial condition

$$g(x_1, x_2, 0) = 1. \tag{5}$$

Using the method of characteristics [17], we can reduce Eq. (4) to the following system of differential equations:

$$\frac{dt}{1} = \frac{dx_1}{\mu_1[x_1 - 1 + r_1(1 - x_2)]} = \frac{dx_2}{\mu_2(1 - r_2)(x_2 - 1)} = \frac{dg(x_1, x_2, t)}{\lambda(x_1 - 1)g(x_1, x_2, t)}.$$

This system has three independent first integrals, and the general solution of the system has the form

$$F(C_1, C_2, C_3) = 0.$$

Integrating the equality $\frac{dt}{1} = \frac{dx_2}{\mu_2(1-r_2)(x_2-1)}$, we derive one integral of the system:

$$x_2 - 1 = C_1 e^{\mu_2(1-r_2)t}, \tag{6}$$

or

$$C_1 = (x_2 - 1)e^{-\mu_2(1-r_2)t}. \tag{7}$$

Substituting (6) into the equation $\frac{dt}{1} = \frac{dx_1}{\mu_1[x_1 - 1 + r_1(1 - x_2)]}$ and integrating it, we derive

$$x_1 - 1 = C_2 e^{\mu_1 t} + \frac{C_1 \mu_1 r_1}{\mu_1 - \mu_2(1 - r_2)} e^{\mu_2(1-r_2)t}. \tag{8}$$

Taking into account (7), we obtain the second integral of the system:

$$C_2 = \left[x_1 - 1 - \frac{\mu_1 r_1 (x_2 - 1)}{\mu_1 - \mu_2(1 - r_2)} \right] e^{-\mu_1 t}. \tag{9}$$

Finally, integrating the equality $\frac{dt}{1} = \frac{dg(x_1, x_2, t)}{\lambda(x_1 - 1)g(x_1, x_2, t)}$ and taking into account expression (8), we obtain

$$g(x_1, x_2, t) = C_3 \exp \left[\frac{\lambda C_2}{\mu_1} e^{\mu_1 t} + \frac{C_1 \mu_1 r_1}{\mu_1 - \mu_2(1 - r_2)} \cdot \frac{e^{\mu_2(1-r_2)t}}{\mu_2(1 - r_2)} \right].$$

Hence, the general solution of Eq. (4) has the form

$$g(x_1, x_2, t) = \Phi(C_1, C_2) \exp \left[\frac{\lambda}{\mu_1} (x_1 - 1) + \frac{\lambda r_1}{\mu_2(1 - r_2)} (x_2 - 1) \right], \tag{10}$$

where $\Phi(C_1, C_2)$ is some differentiable function, C_1 and C_2 are determined by expressions (7) and (9).

Function $\Phi(C_1, C_2)$ can be found using initial condition (5). Substituting $t = 0$ into (10), we obtain

$$\begin{aligned} g(x_1, x_2, 0) &= \Phi \left(x_2 - 1; x_1 - 1 - \frac{\mu_1 r_1 (x_2 - 1)}{\mu_1 - \mu_2(1 - r_2)} \right) \\ &\times \exp \left[\frac{\lambda}{\mu_1} (x_1 - 1) + \frac{\lambda r_1}{\mu_2(1 - r_2)} (x_2 - 1) \right] = 1. \end{aligned}$$

Denoting $U = x_2 - 1$ and $V = x_1 - 1 - \frac{\mu_1 r_1 (x_2 - 1)}{\mu_1 - \mu_2(1 - r_2)}$, we derive the expression

$$\Phi(U, V) = \exp \left[-\frac{\lambda}{\mu_1} V - \frac{\lambda r_1 \mu_1 U}{\mu_2(1 - r_2) (\mu_1 - \mu_2(1 - r_2))} \right].$$

Then, taking into account initial condition (5) and after performing all the necessary transformations, we obtain the following solution of Eq. (4)

$$g(x_1, x_2, t) = \exp \left[\frac{\lambda}{\mu_1} (x_1 - 1) (1 - e^{-\mu_1 t}) + \lambda r_1 (x_2 - 1) \right. \\ \left. \times \left(\frac{1}{\mu_2(1 - r_2)} - \frac{e^{-\mu_1 t}}{\mu_2(1 - r_2) - \mu_1} + \frac{\mu_1 e^{-\mu_2(1-r_2)t}}{\mu_2(1 - r_2)(\mu_2(1 - r_2) - \mu_1)} \right) \right]. \tag{11}$$

If we let $t \rightarrow \infty$, we obtain the following expression for the generating function of the stationary distribution:

$$g(x_1, x_2) = \exp \left[\frac{\lambda}{\mu_1} (x_1 - 1) + \frac{\lambda r_1}{\mu_2(1 - r_2)} (x_2 - 1) \right]. \tag{12}$$

Let us return back to the five-dimensional process $\{l(t), n_1(t), n_2(t), i_1(t), i_2(t)\}$. For its probability distribution $P(l, n_1, n_2, i_1, i_2, t)$ we can write down the system of Kolmogorov differential equations

$$\frac{\partial P(l, n_1, n_2, i_1, i_2, t)}{\partial t} + (\lambda + i_1 \mu_1 + i_2 \mu_2) P(l, n_1, n_2, i_1, i_2, t) \\ = \lambda P(l - 1, n_1, n_2, i_1 - 1, i_2, t) + (i_1 + 1) \mu_1 r_1 P(l, n_1 - 1, n_2, i_1 + 1, i_2 - 1, t) \\ + (i_1 + 1) \mu_1 (1 - r_1) P(l, n_1, n_2, i_1 + 1, i_2, t) + i_2 \mu_2 r_2 P(l, n_1, n_2 - 1, i_1, i_2, t) \\ + (i_2 + 1) \mu_2 (1 - r_2) P(l, n_1, n_2, i_1, i_2 + 1, t) \tag{13}$$

for $l, n_1, n_2, i_1, i_2 = 0, 1, 2, \dots$. The initial condition for this system is as follows:

$$P(l, n_1, n_2, i_1, i_2, 0) = \begin{cases} q(i_1, i_2) & \text{if } n_1 = n_2 = l = 0, \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

where $q(i_1, i_2)$ is the joint probability distribution of the number of customers at the stages of the system at the initial time moment. For purposes of our study, we use the joint probability distribution of the number of customers at the stages of the system in the steady state regime. Its generating function has the form (12).

Consider the generating function (1) of the distribution $P(l, n_1, n_2, i_1, i_2, t)$. From problem (13)–(14), we obtain the following linear partial differential equation of the first order for the function $G(z, y_1, y_2, x_1, x_2, t)$:

$$\frac{\partial G(z, y_1, y_2, x_1, x_2, t)}{\partial t} + \mu_1 (x_1 - r_1 x_2 y_1 - 1 + r_1) \frac{\partial G(z, y_1, y_2, x_1, x_2, t)}{\partial x_1} \\ + \mu_2 (x_2 - r_2 x_2 y_2 - 1 + r_2) \frac{\partial G(z, y_1, y_2, x_1, x_2, t)}{\partial x_2} \\ = \lambda (x_1 z - 1) G(z, y_1, y_2, x_1, x_2, t) \tag{15}$$

with the initial condition

$$G(z, y_1, y_2, x_1, x_2, 0) = g(x_1, x_2). \tag{16}$$

Solving the problem (15)–(16), we obtain the following expression for the generating function $G(z, y_1, y_2, x_1, x_2, t)$ of the multidimensional Markov process $\{l(t), n_1(t), n_2(t), i_1(t), i_2(t)\}$:

$$\begin{aligned}
 G(z, y_1, y_2, x_1, x_2, t) = & \exp \left\{ \frac{\lambda}{\mu_1} (x_1 - 1) [z(1 - e^{-\mu_1 t}) + e^{-\mu_1 t}] \right. \\
 & + \lambda r_1 (x_2 - 1) \left[\frac{e^{-\mu_2(1-r_2y_2)t}}{\mu_2(1-r_2)} + \frac{y_1(z-1)}{\mu_2(1-r_2y_2) - \mu_1} (e^{-\mu_2(1-r_2y_2)t} - e^{-\mu_1 t}) \right. \\
 & \left. \left. - \frac{y_1 z}{\mu_2(1-r_2y_2)} (1 - e^{-\mu_2(1-r_2y_2)t}) \right] + \lambda \frac{r_1}{\mu_1} (z-1) \left(1 - \frac{1-r_2}{1-r_2y_2} y_1 \right) \right. \\
 & \left. \times (1 - e^{-\mu_1 t}) + \lambda t \left[z \left(1 - r_1 \left(1 - \frac{1-r_2}{1-r_2y_2} y_1 \right) \right) - 1 \right] \right\}.
 \end{aligned} \tag{17}$$

Then the expression for the generating function of the three-dimensional process $\{l(t), n_1(t), n_2(t)\}$ is as follows:

$$\begin{aligned}
 G(z, y_1, y_2, t) = G(z, y_1, y_2, 1, 1, t) = & \exp \left\{ \lambda t \left[z - 1 - r_1 z \left(1 - \frac{1-r_2}{1-r_2y_2} y_1 \right) \right] \right. \\
 & \left. + \lambda \frac{r_1}{\mu_1} (z-1) \left(1 - \frac{1-r_2}{1-r_2y_2} y_1 \right) (1 - e^{-\mu_1 t}) + \frac{\lambda r_1 r_2 (y_2 - 1)}{1 - r_2 y_2} \right. \\
 & \left. \times \left[\frac{1 - e^{-\mu_2(1-r_2y_2)t}}{\mu_2(1-r_2)} - \frac{y_1(z-1)}{\mu_2(1-r_2y_2) - \mu_1} - \frac{y_1 z (1 - e^{-\mu_2(1-r_2y_2)t})}{\mu_2(1-r_2y_2)} \right] \right\}.
 \end{aligned} \tag{18}$$

4 Numerical Characteristics of Company’s Profit

For the correctness of the further presentation from the view of the problem domain, we assume that parameters $(1 - \delta_1)$ and $(1 - \delta_2)$ are discounts on the profit from one sale of a product, and not on the price of the product. If we obtain values of these parameters, then the discount on the price can be calculated in easy way.

Suppose that the company receives a profit in the amount ξ from the first customer’s purchase (this is a profit without any discounts). Let ξ be a random variable with finite moments $\mathbb{E}[\xi] = a_1$ and $\mathbb{E}[\xi^2] = a_2$. Taking into account that the customer has discount $(1 - \delta_1)$ after his or her first purchase, the company’s profit from the second purchase of this customer is equal to $\xi\delta_1$. Similarly, from the third and subsequent purchases of the same customer, the company receives profit $\xi\delta_2$. Hence, the total profit of the company from all customers’ purchases up to the time moment t is equal to

$$S(t) = \sum_{k=0}^{l(t)} \xi_k + \sum_{m=0}^{n_1(t)} \xi_m^{(1)} \cdot \delta_1 + \sum_{p=0}^{n_2(t)} \xi_p^{(2)} \cdot \delta_2.$$

Here ξ_k , $\xi_m^{(1)}$ and $\xi_p^{(2)}$ are implementations of the random variable ξ for primary, repeated and subsequent customers’ purchases respectively.

Consider the function $H(\alpha, t) = \mathbb{E} [e^{-\alpha S(t)}]$. We can write

$$\begin{aligned}
 H(\alpha, t) &= \mathbb{E} \left[\exp \left\{ -\alpha \sum_{k=0}^{l(t)} \xi_k - \alpha \sum_{m=0}^{n_1(t)} \xi_m^{(1)} \cdot \delta_1 - \alpha \sum_{\nu=0}^{n_2(t)} \xi_\nu^{(2)} \cdot \delta_2 \right\} \right] \\
 &= \sum_{l=0}^{\infty} \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} (\mathbb{E} [e^{-\alpha \xi}])^l \cdot (\mathbb{E} [e^{-\alpha \delta_1 \xi}])^{n_1} \cdot (\mathbb{E} [e^{-\alpha \delta_2 \xi}])^{n_2} P(l, n_1, n_2, t),
 \end{aligned}$$

where $P(l, n_1, n_2, t) = \mathbb{P}\{l(t) = l, n_1(t) = n_1, n_2(t) = n_2\}$. Denoting $\mathbb{E} [e^{-\alpha \xi}] = \varphi(\alpha)$, $\mathbb{E} [e^{-\alpha \delta_1 \xi}] = \psi_1(\alpha)$, $\mathbb{E} [e^{-\alpha \delta_2 \xi}] = \psi_2(\alpha)$, we obtain

$$H(\alpha, t) = \sum_{l=0}^{\infty} \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} [\varphi(\alpha)]^l [\psi_1(\alpha)]^{n_1} [\psi_2(\alpha)]^{n_2} P(l, n_1, n_2, t).$$

Considering the generating function $G(z, y_1, y_2, t)$ of the process $\{l(t), n_1(t), n_2(t)\}$, we obtain

$$H(\alpha, t) = G(\varphi(\alpha), \psi_1(\alpha), \psi_2(\alpha), t).$$

From (18) it follows that

$$\begin{aligned}
 &H(\alpha, t) = G(\varphi(\alpha), \psi_1(\alpha), \psi_2(\alpha), t) \\
 &= \exp \left\{ \lambda t \left[\varphi(\alpha) - 1 - r_1 \varphi(\alpha) \left(1 - \frac{1-r_2}{1-r_2\psi_2(\alpha)} \psi_1(\alpha) \right) \right] \right. \\
 &\quad \left. + \lambda \frac{r_1}{\mu_1} (\varphi(\alpha) - 1) \left(1 - \frac{1-r_2}{1-r_2\psi_2(\alpha)} \psi_1(\alpha) \right) \left(1 - e^{-\mu_1 t} \right) + \frac{\lambda r_1 r_2 (\psi_2(\alpha) - 1)}{1 - r_2 \psi_2(\alpha)} \right. \\
 &\quad \left. \times \left[\frac{1 - e^{-\mu_2 (1-r_2\psi_2(\alpha))t}}{\mu_2(1-r_2)} - \frac{\mu_1}{\mu_2(1-r_2\psi_2(\alpha)) - \mu_1} - \frac{\psi_1(\alpha)\varphi(\alpha)(1 - e^{-\mu_2(1-r_2\psi_2(\alpha))t})}{\mu_2(1-r_2\psi_2(\alpha))} \right] \right\}. \tag{19}
 \end{aligned}$$

Since $\left. \frac{\partial H(\alpha, t)}{\partial \alpha} \right|_{\alpha=0} = -\mathbb{E}[S(t)]$, then differentiating (19) with respect to α , calculating the resulting expression for $\alpha = 0$ and taking into account that $\varphi(0) = \psi_1(0) = \psi_2(0) = 1$, $\varphi'(0) = -a_1$, $\psi_1'(0) = -a_1 \cdot \delta_1$, $\psi_2'(0) = -a_1 \cdot \delta_2$, we obtain the mean of the total profit of the trading company up to the time moment t :

$$\mathbb{E}[S(t)] = a_1 \lambda t \cdot \left(1 + r_1 \delta_1 + \frac{r_1 r_2}{1 - r_2} \delta_2 \right). \tag{20}$$

The variance of the company’s total profit can be calculated using the expression $\text{Var}[S(t)] = \mathbb{E} [S^2(t)] - (\mathbb{E}[S(t)])^2$.

Taking into account that $\left. \frac{\partial^2 H(\alpha, t)}{\partial \alpha^2} \right|_{\alpha=0} = \mathbb{E} [S(t)^2]$ and $\varphi''(0) = a_2$, $\psi_1''(0) = a_2 \cdot \delta_1^2$, $\psi_2''(0) = a_2 \cdot \delta_2^2$, we obtain

$$\begin{aligned}
 \text{Var}[S(t)] &= a_2 \lambda t \left(1 + r_1 \delta_1^2 + \frac{r_1 r_2}{1 - r_2} \delta_2^2 \right) \\
 &\quad + a_1^2 \left\{ \frac{2\lambda t r_1}{1 - r_2} \left[\delta_1 + r_2 \delta_1 \delta_2 + r_2 \delta_2 + \frac{r_2 \delta_2^2}{(1 - r_2)^2} \right] \right. \\
 &\quad + \frac{2\lambda r_1 r_2 \delta_2}{1 - r_2} \left[\frac{e^{-\mu_1 t} - e^{-\mu_2(1-r_2)t}}{\mu_2(1-r_2) - \mu_1} - \frac{1 - e^{-\mu_2(1-r_2)t}}{\mu_2(1-r_2)} \left(1 + \delta_1 + \frac{r_2 \delta_2}{1 - r_2} \right) \right] \\
 &\quad \left. + \frac{2\lambda r_1}{\mu_1(1 - r_2)} (r_2 \delta_1 - r_2 \delta_2 - \delta_1) (1 - e^{-\mu_1 t}) \right\}. \tag{21}
 \end{aligned}$$

Thus, expressions (20) and (21) allow direct calculations of the mean and variance of the trading company’s profit over a time interval of length t .

5 Optimal Values of Discount

We formulate an optimization problem as obtaining a maximum of the average value of the total profit of the trading company:

$$\mathbb{E}[S(t)] \rightarrow \max.$$

Analyzing expression (20), we can conclude that the company’s profit has the maximal value when the function $f(\delta_1, \delta_2) = r_1\delta_1 + \frac{r_1r_2}{1-r_2}\delta_2$ reaches its maximum.

Obviously, in real life, the likelihood that a customer will reapply to the trading company depends on the offered discounts. Therefore, parameters r_1 and r_2 depend on parameters δ_1 and δ_2 :

$$r_1 = r_1(\delta_1), \quad r_2 = r_2(\delta_2).$$

Using the necessary extremum condition for the function of two variables $f(\delta_1, \delta_2)$, we obtain the following system of equations for determining δ_1 and δ_2 :

$$\begin{cases} r_1 + r_1' \left(\delta_1 + \frac{r_2}{1-r_2} \delta_2 \right) = 0, \\ r_2 + \frac{1}{1-r_2} r_2' \delta_2 = 0. \end{cases} \tag{22}$$

To solve this system, we should make certain assumptions about the dependencies $r_1(\delta_1)$ and $r_2(\delta_2)$. Let us suppose that the dependencies are linear and have the form

$$\begin{cases} r_1(\delta_1) = r_0^{(1)} + \left(r_1^{(1)} - r_0^{(1)} \right) (1 - \delta_1), \\ r_2(\delta_2) = r_0^{(2)} + \left(r_1^{(2)} - r_0^{(2)} \right) (1 - \delta_2). \end{cases} \tag{23}$$

Here $r_0^{(1)}$ is the probability of the customer’s repeated purchase if $\delta_1 = 1$ (without any discount after the first purchase); $r_1^{(1)}$ is the probability of the customer’s repeated purchase if $\delta_1 = 0$ (after the first purchase, discount is equal to 100%); values $r_0^{(2)}$ and $r_1^{(2)}$ are the similar probabilities of the customer’s third and subsequent purchases if $\delta_2 = 1$ and $\delta_2 = 0$ respectively.

Substituting expressions (23) into system (22), we obtain

$$\begin{cases} r_0^{(1)} + \left(r_1^{(1)} - r_0^{(1)} \right) (1 - \delta_1) + \left(r_1^{(1)} - r_0^{(1)} \right) \left(\delta_1 + \frac{r_0^{(2)} + \left(r_1^{(2)} - r_0^{(2)} \right) (1 - \delta_2)}{1 - r_0^{(2)} - \left(r_1^{(2)} - r_0^{(2)} \right) (1 - \delta_2)} \delta_2 \right) = 0, \\ r_0^{(2)} + \left(r_1^{(2)} - r_0^{(2)} \right) (1 - \delta_2) - \frac{\left(r_1^{(2)} - r_0^{(2)} \right) \delta_2}{1 - r_0^{(2)} - \left(r_1^{(2)} - r_0^{(2)} \right) (1 - \delta_2)} = 0. \end{cases} \tag{24}$$

Solving this system of equations and taking into account the sufficient extremum condition, we derive that the function $f(\delta_1, \delta_2)$ reaches its maximum at the point

$$\begin{cases} \delta_1 = \frac{r_1^{(1)}}{2(r_1^{(1)} - r_0^{(1)})} - \frac{(1 - \sqrt{1 - r_1^{(2)}})^2}{2(r_1^{(2)} - r_0^{(2)})}, \\ \delta_2 = \frac{\sqrt{1 - r_1^{(2)}}(1 - \sqrt{1 - r_1^{(2)}})}{(r_1^{(2)} - r_0^{(2)})}. \end{cases} \tag{25}$$

In addition, we should take into account that

$$0 \leq \delta_2 < \delta_1 \leq 1. \tag{26}$$

Thus, the function $f(\delta_1, \delta_2)$ reaches its maximum value at the point (25) when conditions (26) hold.

Example

Let us consider the following example. Let the flow of new customers have the intensity $\lambda = 20$ (customers per day), the average profit from one sale without discounts is equal to $a_1 = 10$ (in some monetary units – MU), the observation time is $t = 30$ days, the probabilities of repeated purchases are determined by formulas (23) with the following values of parameters: $r_0^{(1)} = 0.40$, $r_1^{(1)} = 0.65$, $r_0^{(2)} = 0.45$, $r_1^{(2)} = 0.95$.

Using formulas (25), we calculate: $\delta_1 = 0.697$, $\delta_2 = 0.347$. Therefore, to obtain the highest average profit from sales, you should set the discount on the profit from one sale equal to 30.3% after the first customer purchase, and 65.3% after the second one. If we suppose that the company sells a product with price 130 MU where 30 MU is a clear profit of the company, then we should set the price equal to 120.92 MU for secondary purchases (7% discount on the price) and 110.42 MU for the third and subsequent purchases (15.1% discount on the price). The average total profit over the considered period t will be equal to 11,636 MU.

6 Conclusion and Discussion

We have considered the problem of the discount values’ influence on the profit of a trading company when a two-level discounting system is used. A mathematical model of this trading company has been constructed in the form of a two-stage queueing tandem with an unlimited number of servers and a feedback at the second stage. Analytical expressions are found for the mean and variance of the company’s total profit. Also, analytical expressions are found for the optimal size of discounts for obtaining the maximal profit in the case of linear dependences of the probabilities of repeated purchases on the discounts values.

Let us discuss the possible ways of future research. Firstly, in the paper we have considered only the stationary regime of the company’s functioning, that is,

the company has been working for a long time. In particular, it is assumed that the initial condition corresponds to the stationary distribution of the number of customers at the stages of the system (Sect. 3). In addition, in the model, the intensity of new customers does not change during time. This will not be the case for the companies that are just starting their work or are working in unsteady conditions (for example, selling seasonal stock).

Further, formulas (23) look natural and simple, but cover only the linear dependencies of the probabilities of repeated purchases on the size of discounts. In addition, even in this simple case, determining the values of parameters $r_0^{(1)}$, $r_1^{(1)}$, $r_0^{(2)}$, $r_1^{(2)}$ for a real life example is quite difficult.

The development of the model is possible in the direction of increasing the number of the discounts levels, as well as taking into account the number of purchases that is necessary to move to the next level of discounting. Also, further studies may be related to the analysis of possible risks by studying the influence of discounts values on the total profit variance (21).

So, the paper may be also considered as a base for the further studies in the optimization problems of discounts' assigning using the queueing theory approach.

References

1. Nagle, T.T., Holden, R.K.: The Strategy and Tactics of Pricing: A Guide to Profitable Decision Making. Prentice Hall, New York (2001)
2. Higgins, R.C.: Analysis for Financial Management. McGraw-Hill, New York (2001)
3. Kim, K.H., Hwang, H.: An incremental discount pricing schedule with multiple customers and single price break. *Eur. J. Oper. Res.* **35**(1), 71–79 (1988)
4. Goel, S., Gupta, Y.P., Bector, C.R.: Quantity discount model to increase vendor's profits and decrease buyer's costs. *Int. J. Syst. Sci.* **20**(11), 2341–2346 (1988)
5. Chakravarty, A.K., Martin, G.E.: An optimal joint buyer-seller discount pricing model. *Comput. Oper. Res.* **15**(3), 271–281 (1988)
6. Malliaris, A.G., Brock, W.A.: Stochastic Methods in Economics and Finance. North Holland, Elsevier (1982)
7. Ziemba, W.T., Vickson, R.G.: Stochastic Optimization Models in Finance. Academic Press, New York (1975)
8. Harvey, C.M., Osterbal, L.P.: Discounting models for outcomes over continuous time. *J. Math. Econ.* **48**(5), 284–294 (2012)
9. Cox, D.R., Miller, H.D.: The Theory of Stochastic Processes. Chapman and Hall, London (1965)
10. Allen, A.: Probability, Statistics, and Queueing Theory, 2nd edn. Academic Press, New York (2014)
11. Takács, L.: A single-server queue with feedback. *Bell Labs Tech. J.* **42**(2), 505–519 (1963)
12. Foley, R.D., Disneyl, R.L.: Queues with delayed feedback. *Adv. Appl. Probab.* **15**(1), 162–182 (1983)
13. Pekoz, E.A., Joglekar, N.J.: Poisson traffic flow in a general feedback. *J. Appl. Probab.* **39**(3), 630–636 (2002)

14. Melikov, A.Z., Ponomarenko, L.A., Kuliyeva, K.N.: Calculation of the characteristics of multichannel queuing system with pure losses and feedback. *J. Autom. Inf. Sci.* **47**(5), 19–29 (2015)
15. Moiseeva, S., Zadiranova, L.: Feedback in Infinite-server queuing systems. In: Vishnevsky, V., Kozyrev, D. (eds.) *DCCN 2015. CCIS*, vol. 601, pp. 370–377. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30843-2_38
16. Melikov, A., Zadiranova, L., Moiseev, A.: Two asymptotic conditions in queue with MMPP arrivals and feedback. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) *DCCN 2016. CCIS*, vol. 678, pp. 231–240. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-51917-3_21
17. Polyanin, A.D.: *Handbook of Linear Partial Differential Equations for Engineers and Scientists*. Chapman & Hall/CRC Press, Boca Raton (2002)



Performance Analysis of an M/G/1 Retrial Queueing System Under LCFS-PR Discipline with General Retrial and Setup Times

B. Krishna Kumar¹(✉), R. Sankar¹, and R. Rukmani²

¹ Department of Mathematics, Anna University, Chennai 600 025, India
drbkkumar@hotmail.com

² Department of Mathematics, Pachaiyappa's College, Chennai 600 030, India

Abstract. This paper deals with the analysis of an M/G/1 retrial queueing system with general retrial and setup times. The customers are served under the preemptive resume priority last-come, first-service (LCFS-PR) discipline and only the customer at the head of the orbit queue is allowed to access the server. The necessary and sufficient condition for the system to be stable is investigated. Using the generating functions technique, the joint steady-state distribution of the server states and the number of customers in the orbit are obtained along with some interesting and important performance measures. Finally, some numerical examples to illustrate the effect of system parameters on several performance characteristics are carried out.

Keywords: Retrial queue · Setup time · LCFS-PR · Ergodicity
Steady-state distribution

1 Introduction

In recent years, there has been a keen interest in the study of retrial queueing systems. In view of the network complexity and increasing the amount of incoming flows, the retrial phenomenon may have a significant impact on the computer network performance. Retrial queueing systems are described by the feature that the customers (or data, packet) who find the server busy, do not wait in an ordinary queue. Instead of that they join a pool of unsatisfied customers, called orbit/retrial group, trying to obtain service after a random amount of time in a random order. Apart from theoretical interests, retrial queues have been successfully and widely applied in designing of telephone switching systems, telematic devices, call centers and several computer network systems. A comprehensive overview of the literature and discussion of practical situations where retrial queues arise can be found in the classical bibliographie by Artalejo [2], in the survey article by Falin [8] and in the monographs by Falin and Templeton [9] and Artalejo and Gomez-Corral [3].

© Springer Nature Switzerland AG 2018

A. Dudin et al. (Eds.): ITMM 2018/WRQ 2018, CCIS 912, pp. 333–347, 2018.

https://doi.org/10.1007/978-3-319-97595-5_26

Recently several researchers have investigated intensively the retrial queueing systems where the retrial time has a general distribution and only the customer at the head of the orbit queue is allowed to retry for service. Due to applications in teletraffic theory, computer communication protocols, and local area network systems, retrial queues with general and non-exponential retrial time distributions have drawn more attention in recent years. Choi et al. [5] have discussed an M/M/1 retrial queue where the retrial times are generally distributed. Later on, Gomez-Corral [12] investigated an M/G/1 retrial queue with general retrial time distribution where only the customer at the head of the orbit queue is allowed to retry for service. Since the work of Gomez-Corral, many retrial queueing systems have been studied and his analysis has been extended to the systems with additional characteristics. For related literature on the general retrial time distribution with additional features and applications, one can refer to Atencia and Moreno [4], Ke and Chang [15], Krishna Kumar et al. [16], Gao and Wang [10] and Li and Zhang [14] and references therein.

In the complex communication and control systems with waiting lines, it is attractive to always focus the service resource on the job that has waited least. This leads to a LCFS (last-come, first-service) queue discipline which can be implemented preemptively, where a newly arrived job immediately enters service and the one in service goes on standby. Later, the interrupted job has to be restored and processed from scratch, so that the previous service effort expended on that job is lost. Further, it is assumed that every time the job is restarted, the remaining service is identical to its initial service requirement. This kind of service rule, the so-called “last-come, first-service preemptive-repeat identical” (LCFS-PR) service discipline, has been investigated in telematic devices and communication packet switching systems (Conti et al. [6] and Harchol-Balter [13]).

Many queueing systems containing “mechanical parts” need a setup period to serve the customers. The server’s setup period corresponding to the preparatory work of the server which is needed before starting each busy period. Potential applications of such kinds of queueing models can be found in switched virtual connection (SVC)- based virtual local area network (LAN), in internet protocol (IP) over asynchronous transfer mode (ATM) networks and in power saving data centers (Sakai et al. [19], Niu and Takahashi [17] and Gebrehiwot et al. [11]).

In this paper, we develop an analysis of the preemptive priority LCFS single server retrial queueing system with general distributions for both retrial time and setup time. Our system is more suitable for the control switching nodes in the information systems (Agalarov [1] and Zayats et al. [22]). In a communication switching node retrial transmission system, when a message (data, packets) is being processed, a new information coming to the server is more actual than the one on service. In this situation, the message is moved to another place (orbit), if the contained information is valuable and that can be processed later on or if the information is not valuable any more, it is deleted and in both cases the message under service is interrupted. Further the server requires the warm-up time (setup time) before starting of each busy period. We can view such system as the LCFS-PR retrial queueing system with setup time.

2 Model Description of the System

We consider a single server retrial queueing system with setup time and LCFS-PR discipline. Customers arrive at the server according to a Poisson process with intensity $\lambda > 0$. The server provides service to all arriving customers and the service times of customers are independent and identical random variables $B_i \sim B$ with cumulative distribution function (cdf) $B(x)$, probability density function (pdf) $b(x)$ and $E(B)$ and $E(B^2)$ as the finite first two moments. In addition, we assume that $B(0) = 0$, i.e., the service times cannot be zero and $B^*(s) = \int_0^\infty e^{-sx} dB(x)$ is the Laplace-stieltjes transform (LST) of $B(x)$.

If an arriving primary customer who finds the server free upon arrival immediately begins its service. Otherwise, the arriving primary customer either displaces the customer that is currently being served to the orbit with probability θ , ($0 < \theta < 1$), or the primary customer expels the customer who is currently being served out of the system with complementary probability $1 - \theta$ and begins its own service immediately. From the description, it is clear that the preemption of the services of customers occur only when the server is busy with the customers.

After a service completion without preemption, if there is at least one customer in the system, the server will stay idle until either a customer from the orbit or a new primary customer to arrive. On the other hand, if there is no customer in the orbit after a service completion, i.e., the system becomes empty, the server will be turned off at once. A new arriving primary customer can reactivate the off server and occupy it. The off server needs some setup period to turn on (i.e., the setup time is activated upon an arrival of a new primary customer) in order to serve the waiting customer at the server. The time of setup is a random variable 'S' with cdf $S(x)$, pdf $s(x)$, LST $S^*(s)$ and finite first two moments $E(S)$ and $E(S^2)$. During the setup period, arriving new primary customers enter the orbit and behave the same as other customers in the orbit.

Further, it is assumed that, due to preemption or duration of the setup time, the customers who enter the orbit form a virtual queue in accordance with an FCFS queueing discipline. Thus it is clear that at any service completion instant without preemption, the server becomes free and in such a case, a possible new primary arrival and the one (if any) at the head of the orbit queue compete for service. The attempt time of the retrial customer, say H , is a generally distributed random variable with cdf $H(x)$, pdf $h(x)$, LST $H^*(s)$ and the finite first two moments $E(H)$ and $E(H^2)$. All the stochastic processes involved in the system under study are assumed to be independent of each other.

Based on the model description, the stochastic behaviour of the system can be described by $\{N(t); t \geq 0\} = \{(C(t), X(t), \xi(t)); t \geq 0\}$ where $C(t)$ denotes the server state (0, 1 or 2, according as the server is idle, the server is busy or the server is in setup process, respectively), $X(t)$ corresponds to the number of customers in the orbit and $\xi(t)$ represents the supplementary variable at any time t . If $C(t) = 0$ and $X(t) > 0$, $\xi(t)$ represents the elapsed retrial time, if $C(t) = 1$ and $X(t) \geq 0$, $\xi(t)$ denotes the elapsed service time and if $C(t) = 2$ and $X(t) \geq 0$, $\xi(t)$ stands for the elapsed setup time.

Thus, the process $\{N(t); t \geq 0\} = \{(C(t), X(t), \xi(t)); t \geq 0\}$ is a Markov process. In what follows, we neglect $\xi(t)$ and consider only the pair $(C(t), X(t))$ whose state space is $E = \{0, 1, 2\} \times \mathbb{Z}_+$. In addition, we define the functions $\beta(x)$, $\eta(x)$ and $\alpha(x)$, respectively, as the conditional completion rates/hazard rates at times x for service, for retrial attempt and for setup process; i.e., $\beta(x) = \frac{b(x)}{1-B(x)}$, $\eta(x) = \frac{h(x)}{1-H(x)}$ and $\alpha(x) = \frac{s(x)}{1-S(x)}$.

2.1 Ergodicity Condition

In order to establish a criterion for the existence of the limit distribution, we first analyze the ergodicity of the system by using embedded Markov chain at departure epochs of customers immediately after completion of services without preemption. Since the stability condition of the system concerns only the situation where the number of customers in the orbit is large enough, the effect of setup time can be ignored.

To this end, let η_k be the time epoch of the k^{th} customer's departure after a non-interrupted service completion when the server is free. The next customer begins its service at epoch $\xi_k = \eta_{k-1} + T_k$ where T_k denotes the random time during which the server is free. Define $X_k = X(\eta_k+)$ to be the number of customers in the orbit immediately after instant η_k . Thus the random sequence $\{X_k; k \in \mathbb{N}\}$ forms a discrete time Markov chain which is embedded in our continuous time retrial queueing system. It can be seen that $\{X_k; k \in \mathbb{N}\}$ is irreducible and aperiodic on state set \mathbb{Z}_+ . Moreover, for the random variable X_k , the following state equation is valid:

$$X_k = X_{k-1} - B_k + A_k, \tag{1}$$

where A_k is the number of preempted customers that enter into the orbit during the time $\eta_k - \xi_k$ and $B_k = 0$, if the customer that begins its service at time ξ_k is a primary customer and $B_k = 1$, if the head of retrial customer starts its service. Thus, the conditional probability distribution of the Bernoulli random variable B_k is defined as

$$P(B_k = 0/X_{k-1} = n) = 1 - U(n)H^*(\lambda), \quad P(B_k = 1/X_{k-1} = n) = U(n)H^*(\lambda), \tag{2}$$

where $U(n)$ denotes the heaviside unit function. It is evident that, for $k \in \mathbb{N}$,

$$\eta_k - \xi_k = \sum_{j=1}^n V_j + B, \quad n = 1, 2, 3, \dots \tag{3}$$

where n is the number of customers preempted from service and V_j is the preempted service time of the j^{th} customer due to the arrival of a new primary customer before completing service and B is a non-interrupted service time. As the primary customers arrive according to a Poisson process with intensity λ , the probability that there is no interruption during a service time is

$$B^*(\lambda) = \int_0^\infty e^{-\lambda x} dB(x). \tag{4}$$

Next, we shall determine the probability distribution of the random variable A_k as follows:

By the total probability argument, we have

$$\begin{aligned}
 P(A_k = n) &= \sum_{j=0}^{\infty} P(A_k = n, M = n + j) \\
 &= \sum_{j=0}^{\infty} P(A_k = n/M = n + j)P(M = n + j) \tag{5}
 \end{aligned}$$

where the random variable M represents the total number of preemptions (interruptions) occurred to customers that are being served due to the arrivals of new primary customers. Consequently, the random variable M is geometrically distributed and its probability mass function is given by

$$P(M = m) = (1 - B^*(\lambda))^m B^*(\lambda), \quad m = 0, 1, 2, \dots \tag{6}$$

By applying the binomial probability law, the conditional probability

$$P(A_k = n/M = n + j) = \binom{n + j}{n} \theta^n (1 - \theta)^j, \quad \text{where } 0 < \theta < 1. \tag{7}$$

Taking into account of (6) and (7), (5) yields

$$P(A_k = n) = \sum_{j=0}^{\infty} \binom{n + j}{n} \theta^n (1 - \theta)^j (1 - B^*(\lambda))^{n+j} B^*(\lambda),$$

whence

$$P(A_k = n) = \frac{B^*(\lambda)[(1 - B^*(\lambda))\theta]^n}{[1 - (1 - B^*(\lambda))(1 - \theta)]^{n+1}}, \quad n = 0, 1, 2, 3, \dots \tag{8}$$

and the corresponding mean

$$E(A_k) = \sum_{n=0}^{\infty} nP(A_k = n) = \frac{\theta[1 - B^*(\lambda)]}{B^*(\lambda)}. \tag{9}$$

Theorem 1. *The inequality $\frac{\theta[1 - B^*(\lambda)]}{B^*(\lambda)} < H^*(\lambda)$ is a necessary and sufficient condition for the retrial queueing system under discussion to be stable.*

Proof. To prove the sufficient condition for ergodicity, we shall use Foster’s criterion (see Pakes [18]) which states that an irreducible and aperiodic Markov chain $\{X_k; k \in \mathbb{N}\}$ is ergodic if there exists a non-negative function $f(n), n \in \mathbb{Z}_+$ and $\epsilon > 0$ such that the mean drift $\chi_n = E[f(X_k) - f(X_{k-1})/X_{k-1} = n]$ is finite for all $n \in \mathbb{Z}_+$ and $\chi_n < -\epsilon$ for all $n \in \mathbb{Z}_+$, except perhaps for a finite number. In our case, we consider the function $f(n) = n$, so that the mean drift

$$\chi_n = E[X_k - X_{k-1}/X_{k-1} = n] = -E(B_k) + E(A_k),$$

so that

$$\chi_n = -U(n)H^*(\lambda) + \frac{\theta[1 - B^*(\lambda)]}{B^*(\lambda)} = \begin{cases} \frac{\theta[1 - B^*(\lambda)]}{B^*(\lambda)} & \text{if } n = 0 \\ -H^*(\lambda) + \frac{\theta[1 - B^*(\lambda)]}{B^*(\lambda)} & \text{if } n = 1, 2, 3, \dots \end{cases}$$

Thus, we have $|\chi_n| < \infty$ for all $n \in \mathbb{Z}_+$ and if $\frac{\theta[1 - B^*(\lambda)]}{B^*(\lambda)} < H^*(\lambda)$, then $\lim_{n \rightarrow \infty} \sup \chi_n < 0, \forall n \in \mathbb{Z}_+$. Hence the condition $\frac{\theta[1 - B^*(\lambda)]}{B^*(\lambda)H^*(\lambda)} < 1$ turns out to be sufficient condition for the ergodicity of the embedded Markov chain.

The same inequality is also necessary for ergodicity. As stated in Sennott et al. [20], we can guarantee non-ergodicity of the Markov chain $\{X_k; k \in \mathbb{N}\}$, if it satisfies Kaplan’s condition, namely $\chi_n < \infty$ for all $n \geq 0$ and there exists $n_0 \in \mathbb{N}$ such that $\chi_n \geq 0$ for all $n \geq n_0$. Notice that, in our case, Kaplan’s condition is satisfied because there is a ‘ r ’ such that $p_{ij} = 0$ for $j < i - r$ and $i > 0$, where $P = (p_{ij})$ is the transition probability matrix of $\{X_k; k \in \mathbb{N}\}$. Then $\frac{\theta[1 - B^*(\lambda)]}{B^*(\lambda)} \geq H^*(\lambda)$ implies the non-ergodicity of the Markov chain.

Since the arrival stream of the primary customers is a Poisson process, it can be shown from PASTA (Poisson Arrival See Time Averages) property (Wolff [21]) and Burke’s theorem (Cooper [7], pp. 187–188) that the limit distribution of $\{N(t); t \geq 0\} = \{(C(t), X(t)); t \geq 0\}$ exists and positive if and only if $\frac{\theta[1 - B^*(\lambda)]}{B^*(\lambda)} < H^*(\lambda)$. □

Remark 1. It must be pointed out that the stability condition does not depend on the setup time distribution as stated earlier.

3 Steady-State Distribution

We now investigate an analytical solution for the joint stationary distribution of the number of customers in the orbit and the state of the server of our queueing system in terms of generating functions.

For the process $\{N(t); t \geq 0\}$, we define the joint state probability

$$P_0(t) = P \{C(t) = 0, X(t) = 0\} \quad \text{for } t \geq 0$$

and the joint state probability densities

$$\begin{aligned} P_n(x, t)dx &= P \{C(t) = 1, X(t) = n, x \leq \xi(t) < x + dx\} \text{ for } t \geq 0, x \geq 0, n \geq 0, \\ Q_n(x, t)dx &= P \{C(t) = 0, X(t) = n, x \leq \xi(t) < x + dx\} \text{ for } t \geq 0, x \geq 0, n \geq 1, \\ S_n(x, t)dx &= P \{C(t) = 2, X(t) = n, x \leq \xi(t) < x + dx\} \text{ for } t \geq 0, x \geq 0, n \geq 0. \end{aligned}$$

Fulfillment of the ergodic condition $\frac{\theta[1 - B^*(\lambda)]}{B^*(\lambda)} < H^*(\lambda)$ yields the existence of the limiting probability $P_0 = \lim_{t \rightarrow \infty} P_0(t)$ and the limiting densities $P_n(x) = \lim_{t \rightarrow \infty} P_n(x, t)$ for $x \geq 0, n \geq 0, Q_n(x) = \lim_{t \rightarrow \infty} Q_n(x, t)$ for $x \geq 0, n \geq 1$ and

$S_n(x) = \lim_{t \rightarrow \infty} S_n(x, t)$ for $x \geq 0, n \geq 0$. By the method of supplementary variable technique, we obtain the system of equilibrium equations as

$$\lambda P_0 = \int_0^\infty P_0(x)\beta(x)dx, \tag{10}$$

$$\frac{dP_n(x)}{dx} = -(\lambda + \beta(x))P_n(x) \text{ for } x > 0 \text{ and } n \geq 0, \tag{11}$$

$$\frac{dQ_n(x)}{dx} = -(\lambda + \eta(x))Q_n(x) \text{ for } x > 0 \text{ and } n \geq 1, \tag{12}$$

$$\frac{dS_0(x)}{dx} = -(\lambda + \alpha(x))S_0(x) \text{ for } x > 0, \tag{13}$$

$$\frac{dS_n(x)}{dx} = -(\lambda + \alpha(x))S_n(x) + \lambda S_{n-1}(x) \text{ for } x > 0 \text{ and } n \geq 1. \tag{14}$$

The boundary conditions of the above system of equations in stationary regime are

$$P_0(0) = \int_0^\infty Q_1(x)\eta(x)dx + \lambda(1 - \theta) \int_0^\infty P_0(x)dx + \int_0^\infty S_0(x)\alpha(x) dx, \tag{15}$$

$$P_n(0) = \lambda\theta \int_0^\infty P_{n-1}(x)dx + \lambda(1 - \theta) \int_0^\infty P_n(x)dx + \int_0^\infty Q_{n+1}(x)\eta(x)dx + \int_0^\infty S_n(x)\alpha(x)dx + \lambda \int_0^\infty Q_n(x)dx \text{ for } n \geq 1, \tag{16}$$

$$Q_n(0) = \int_0^\infty P_n(x)\beta(x) dx \text{ for } n \geq 1, \tag{17}$$

$$S_n(0) = \lambda P_0 \delta_{n0} \text{ for } n \geq 0, \tag{18}$$

where δ_{n0} is the Kronecker delta function and the normalization condition is

$$P_0 + \sum_{n=0}^\infty \int_0^\infty P_n(x)dx + \sum_{n=1}^\infty \int_0^\infty Q_n(x)dx + \sum_{n=0}^\infty \int_0^\infty S_n(x)dx = 1. \tag{19}$$

In order to solve the Eqs. (10)–(18), we introduce the following partial generating functions:

$$P(x, z) = \sum_{n=0}^\infty P_n(x)z^n, \quad Q(x, z) = \sum_{n=1}^\infty Q_n(x)z^n, \quad S(x, z) = \sum_{n=0}^\infty S_n(x)z^n, \quad |z| < 1.$$

Theorem 2. *If $\frac{\theta[1 - B^*(\lambda)]}{B^*(\lambda)} < H^*(\lambda)$, then the joint steady-state distribution of $\{N(t); t \geq 0\}$ is obtained as*

$$P(x, z) = \frac{P_0 \lambda \{z S^*(\lambda(1-z)) - [H^*(\lambda) + z(1-H^*(\lambda))]\} e^{-\lambda x - \int_0^x \beta(u) du}}{z - z[1 - B^*(\lambda)](1 - \theta + \theta z) - B^*(\lambda)[H^*(\lambda) + z(1 - H^*(\lambda))]}, \tag{20}$$

$$Q(x, z) = \frac{P_0 \lambda z \{(1 - \theta + \theta z)(1 - B^*(\lambda)) - [1 - B^*(\lambda)S^*(\lambda(1-z))]\} e^{-\lambda x - \int_0^x \eta(u) du}}{z - z[1 - B^*(\lambda)](1 - \theta + \theta z) - B^*(\lambda)[H^*(\lambda) + z(1 - H^*(\lambda))]}, \tag{21}$$

$$S(x, z) = P_0 \lambda e^{-\lambda(1-z)x - \int_0^x \alpha(u) du}, \tag{22}$$

with $S^*(\lambda(1-z)) = \int_0^\infty \alpha(x) e^{-\int_0^x \alpha(u) du - \lambda(1-z)x} dx$, $H^*(\lambda) = \int_0^\infty \eta(x) e^{-\int_0^x \eta(u) du - \lambda x} dx$ and $B^*(\lambda) = \int_0^\infty \beta(x) e^{-\int_0^x \beta(u) du - \lambda x} dx$. Here the only unknown probability P_0 of the state $(0, 0)$ can be determined from the normalization condition (19).

Proof. The proof follows by some calculus and mathematical manipulations and thus we omit the details. □

We now define the partial probability generating functions for the number of customers in the orbit by suppressing the supplementary variables under the ergodicity as

$$P(z) = \int_0^\infty P(x, z) dx, \quad Q(z) = \int_0^\infty Q(x, z) dx \quad \text{and} \quad S(z) = \int_0^\infty S(x, z) dx. \tag{23}$$

Note that $P(z)$ is the partial probability generating function of the orbit size when the server is busy, $Q(z)$ is the partial probability generating function of the orbit size when the server is free during the retrial time, $S(z)$ is the partial probability generating function of the number of customers in the orbit when the server is in the setup process, and P_0 is the stationary probability that the server is idle and no customer in the system.

Theorem 3. *Under the steady state condition, $\frac{\theta[1 - B^*(\lambda)]}{B^*(\lambda)} < H^*(\lambda)$, the partial probability generating functions are derived as*

$$P(z) = \frac{P_0 [1 - B^*(\lambda)] \{z S^*(\lambda(1-z)) - [H^*(\lambda) + z(1 - H^*(\lambda))]\}}{(1-z) \{ \theta z [1 - B^*(\lambda)] - B^*(\lambda) H^*(\lambda) \}}, \tag{24}$$

$$Q(z) = \frac{P_0 z [1 - H^*(\lambda)] \{ (1 - \theta + \theta z)(1 - B^*(\lambda)) - [1 - B^*(\lambda) S^*(\lambda(1-z))] \}}{(1-z) \{ \theta z [1 - B^*(\lambda)] - B^*(\lambda) H^*(\lambda) \}}, \tag{25}$$

$$S(z) = \frac{P_0 [1 - S^*(\lambda(1-z))]}{(1-z)}, \tag{26}$$

where

$$P_0 = \frac{B^*(\lambda) H^*(\lambda) - \theta [1 - B^*(\lambda)]}{[\lambda E(S) + H^*(\lambda)] [(1 - \theta) + \theta B^*(\lambda)]}. \tag{27}$$

Proof. Taking into account of (23), from (20), (21) and (22), we will get (24), (25) and (26), respectively. At this point, the only unknown is P_0 which can be determined using the normalizing condition, $P_0 + P(1) + Q(1) + S(1) = 1$. Thus, setting $z = 1$ in (24)–(26) and applying L'Hospital's rule whenever necessary, after using normalization condition and rearrangement, we get (27). Hence, the partial probability generating functions are determined explicitly. \square

Corollary 1. *If $\frac{\theta[1 - B^*(\lambda)]}{B^*(\lambda)} < H^*(\lambda)$, then*

(1) *the probability that the server is busy providing service*

$$P_B = \lim_{z \rightarrow 1} P(z) = \frac{[1 - B^*(\lambda)]}{[1 - \theta + \theta B^*(\lambda)]}, \tag{28}$$

(2) *the probability that the server is free during the retrial time*

$$P_R = \lim_{z \rightarrow 1} Q(z) = \frac{[1 - H^*(\lambda)] \{ \theta(1 - B^*(\lambda)) + \lambda E(S) B^*(\lambda) \}}{[\lambda E(S) + H^*(\lambda)] [1 - \theta + \theta B^*(\lambda)]}, \tag{29}$$

(3) *the probability that the server is in the setup mode*

$$P_S = \lim_{z \rightarrow 1} S(z) = \frac{\lambda E(S) \{ B^*(\lambda) H^*(\lambda) - \theta [1 - B^*(\lambda)] \}}{[\lambda E(S) + H^*(\lambda)] [1 - \theta + \theta B^*(\lambda)]}, \tag{30}$$

(4) *the probability that the server is idle or free*

$$P_I = P_0 + P_R + P_S = \frac{B^*(\lambda) - \theta [1 - B^*(\lambda)]}{[1 - \theta + \theta B^*(\lambda)]}, \tag{31}$$

(5) *the probability that the orbit is empty, i.e., no customer in the orbit*

$$\begin{aligned} P(X = 0) &= P_0 + S(0) + P(0) + Q(0) \\ &= \left\{ \frac{B^*(\lambda) H^*(\lambda) - \theta [1 - B^*(\lambda)]}{B^*(\lambda) [(1 - \theta) + \theta B^*(\lambda)]} \right\} \left\{ \frac{1 + B^*(\lambda) (1 - S^*(\lambda))}{\lambda E(S) + H^*(\lambda)} \right\} \end{aligned} \tag{32}$$

Remark 2. It can be observed that the probability descriptors P_B and P_I are independent of both the inter-retrial time distribution and the setup time distribution.

Corollary 2. *If the condition $\frac{\theta[1 - B^*(\lambda)]}{B^*(\lambda)} < H^*(\lambda)$ is fulfilled, then*

(1) *the probability generating function, $\Phi(z)$, of the number of customers in the orbit is given as*

$$\begin{aligned} \Phi(z) &= P_0 + S(z) + P(z) + Q(z) \\ &= \frac{P_0 \left\{ \begin{aligned} &z B^*(\lambda) H^*(\lambda) + z S^*(\lambda (1 - z)) [1 - B^*(\lambda) H^*(\lambda)] \\ &- [H^*(\lambda) + z(1 - H^*(\lambda))] + \theta z(1 - z) (1 - B^*(\lambda)) H^*(\lambda) \\ &+ (1 - S^*(\lambda (1 - z))) [z \theta (1 - B^*(\lambda)) - B^*(\lambda) H^*(\lambda)] \end{aligned} \right\}}{(1 - z) [\theta z(1 - B^*(\lambda)) - B^*(\lambda) H^*(\lambda)]} \end{aligned} \tag{33}$$

where P_0 is given in (27),

(2) the probability generating function, $\Psi(z)$, of the number of customers in the system including the customer being served is obtained as

$$\begin{aligned} \Psi(z) &= P_0 + zS(z) + zP(z) + Q(z) \\ &= \left\{ \frac{[B^*(\lambda)H^*(\lambda) - \theta(1 - B^*(\lambda))]}{[B^*(\lambda)H^*(\lambda) - \theta z(1 - B^*(\lambda))]} \right\} \left\{ \frac{[B^*(\lambda) + (1 - \theta)z(1 - B^*(\lambda))]}{[1 - \theta + \theta B^*(\lambda)]} \right\} \\ &\quad \times \left\{ \frac{z(1 - S^*(\lambda(1 - z))) + (1 - z)H^*(\lambda)}{(1 - z)[\lambda E(S) + H^*(\lambda)]} \right\}. \end{aligned} \tag{34}$$

4 Cyclic Analysis and Performance Measures

We now present some key performance measures for the retrial queueing system under investigation.

1. Recall that the regeneration cycle, T , of queueing system is the time elapsed between two consecutive primary customer arrivals finding the system empty. Thus, the mean length of regeneration cycle of our retrial queueing system is

$$E(T) = \frac{1}{\lambda} = \frac{1}{\lambda} \left\{ \frac{[\lambda E(S) + H^*(\lambda)][(1 - \theta) + \theta B^*(\lambda)]}{[B^*(\lambda)H^*(\lambda) - \theta(1 - B^*(\lambda))]} \right\}. \tag{35}$$

2. The mean length, $E(T_0)$, of the system being empty during the regeneration cycle period is computed as

$$E(T_0) = P_0 E(T) = \frac{1}{\lambda}. \tag{36}$$

3. The mean length, $E(T_1)$, of the server's idle/free state during the regenerative cycle period is given by

$$E(T_1) = Q(1)E(T) = \frac{[1 - H^*(\lambda)]}{\lambda} \left\{ \frac{\theta(1 - B^*(\lambda)) + B^*(\lambda)\lambda E(S)}{[B^*(\lambda)H^*(\lambda) - \theta(1 - B^*(\lambda))]} \right\}. \tag{37}$$

4. The mean length, $E(T_2)$, of the server being busy providing the service during the regenerative cycle period is determined as

$$E(T_2) = P(1)E(T) = \frac{1}{\lambda} \left\{ \frac{(1 - B^*(\lambda))[\lambda E(S) + H^*(\lambda)]}{[B^*(\lambda)H^*(\lambda) - \theta(1 - B^*(\lambda))]} \right\}. \tag{38}$$

5. The mean length, $E(T_3)$, of the server being in the setup period during the regenerative cycle period is obtained as

$$E(T_3) = S(1)E(T) = E(S). \tag{39}$$

6. The busy period, L , of the retrial queueing system is defined as the period that starts at the epoch when an arriving primary customer finds an empty system (i.e., the server is free and no customer is in the orbit) and ends at the

departure epoch after service completion of a customer without preemption when the system becomes empty again. Thus, the mean length, $E(L)$, of the system busy period of our retrial queueing system is obtained directly by the theory of regenerative process as

$$\begin{aligned}
 E(L) &= \frac{1}{\lambda} \left[\frac{1}{P_0} - 1 \right] \\
 &= \frac{1}{\lambda} \frac{\{\lambda E(S)[1 - \theta + \theta B^*(\lambda)] + (1 - B^*(\lambda))[H^*(\lambda) + \theta(1 - H^*(\lambda))]\}}{\{B^*(\lambda)H^*(\lambda) - \theta[1 - B^*(\lambda)]\}}. \quad (40)
 \end{aligned}$$

7. The expected number, L_Q , of customers in the orbit in stationarity is calculated by differentiating (33) with respect to z and evaluating at $z = 1$. Thus

$$\begin{aligned}
 L_Q = \Phi'(1) &= \left(\frac{\frac{\lambda^2 E(S^2)}{2}}{\lambda E(S) + H^*(\lambda)} \right) + \left(\frac{1 - B^*(\lambda)H^*(\lambda)}{(1 - \theta) + \theta B^*(\lambda)} \right) \\
 &\quad \times \left\{ \frac{\lambda E(S)}{\lambda E(S) + H^*(\lambda)} + \frac{\theta(1 - B^*(\lambda))}{B^*(\lambda)H^*(\lambda) - \theta(1 - B^*(\lambda))} \right\}. \quad (41)
 \end{aligned}$$

8. The expected number, L_S , of customers in the system under steady state condition is determined from (34) as

$$\begin{aligned}
 L_S = \Psi'(1) &= \left(\frac{\lambda E(S) + \frac{\lambda^2 E(S^2)}{2}}{\lambda E(S) + H^*(\lambda)} \right) + \left(\frac{(1 - B^*(\lambda))}{(1 - \theta) + \theta B^*(\lambda)} \right) \\
 &\quad \times \left\{ 1 + \frac{\theta[1 - B^*(\lambda)H^*(\lambda)]}{[B^*(\lambda)H^*(\lambda) - \theta(1 - B^*(\lambda))]} \right\}. \quad (42)
 \end{aligned}$$

9. Let W_Q denote the average time a customer spends in the orbit under stability condition. Applying Little's law leads to

$$W_Q = \frac{L_Q}{\lambda_{orbit}}, \quad (43)$$

where $\lambda_{orbit} = \lambda\theta P(1) + \lambda S(1)$ is the arrival rate of the customer to the orbit.

10. The average time a customer spends in the system in stationary regime, by invoking again the Little's law, is calculated as

$$W_S = \frac{L_S}{\lambda_{system}}, \quad (44)$$

where $\lambda_{system} = \lambda\theta P(1) + \lambda[S(1) + Q(1) + P_0]$ is the arrival rate of the customer to the system.

5 Numerical Results and Discussions

In the sequel, we study the effect of the system parameters on the main performance measures of our retrial queueing system. Of course, in all the numerical examples below, the chosen parametric values satisfy the ergodic condition.

For computational purposes, we assume that the service time, retrial time and setup time all follow

- (i) Exponential distributions:
 $b(x) = \mu e^{-\mu x}, x > 0, h(x) = \nu e^{-\nu x}, x > 0, s(x) = \alpha e^{-\alpha x}, x > 0.$
- (ii) 2-stage Erlangian distributions:
 $b(x) = \mu^2 x e^{-\mu x}, x > 0, h(x) = \nu^2 x e^{-\nu x}, x > 0, s(x) = \alpha^2 x e^{-\alpha x}, x > 0.$
- (iii) Hyperexponential distributions:
 $b(x) = p \mu e^{-\mu x} + (1 - p) \mu^2 e^{-\mu^2 x}, x > 0,$
 $h(x) = p \nu e^{-\nu x} + (1 - p) \nu^2 e^{-\nu^2 x}, x > 0,$
 $s(x) = p \alpha e^{-\alpha x} + (1 - p) \alpha^2 e^{-\alpha^2 x}, x > 0,$ where $0 < p < 1.$

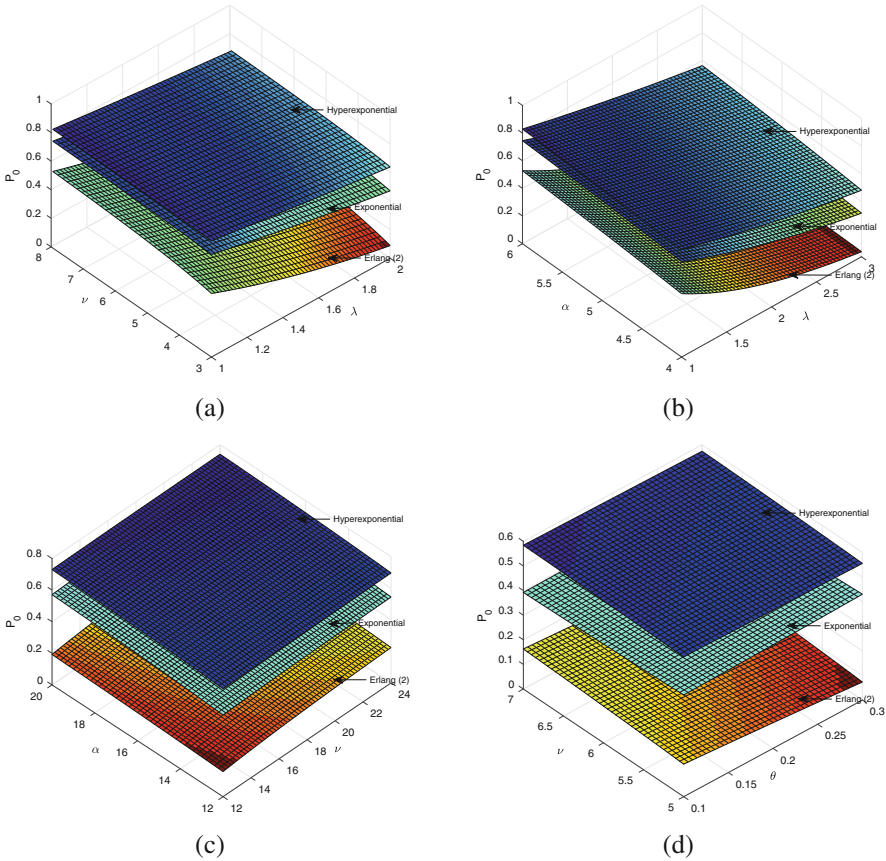


Fig. 1. (a) P_0 versus (λ, ν) for $(\alpha, \mu, \theta, p) = (6, 9, 0.5, 0.6)$, (b) P_0 versus (λ, α) for $(\nu, \mu, \theta, p) = (7, 9, 0.5, 0.6)$, (c) P_0 versus (ν, α) for $(\lambda, \mu, \theta, p) = (3, 9, 0.5, 0.6)$, (d) P_0 versus (θ, ν) for $(\lambda, \alpha, \mu, p) = (3, 6, 9, 0.6)$

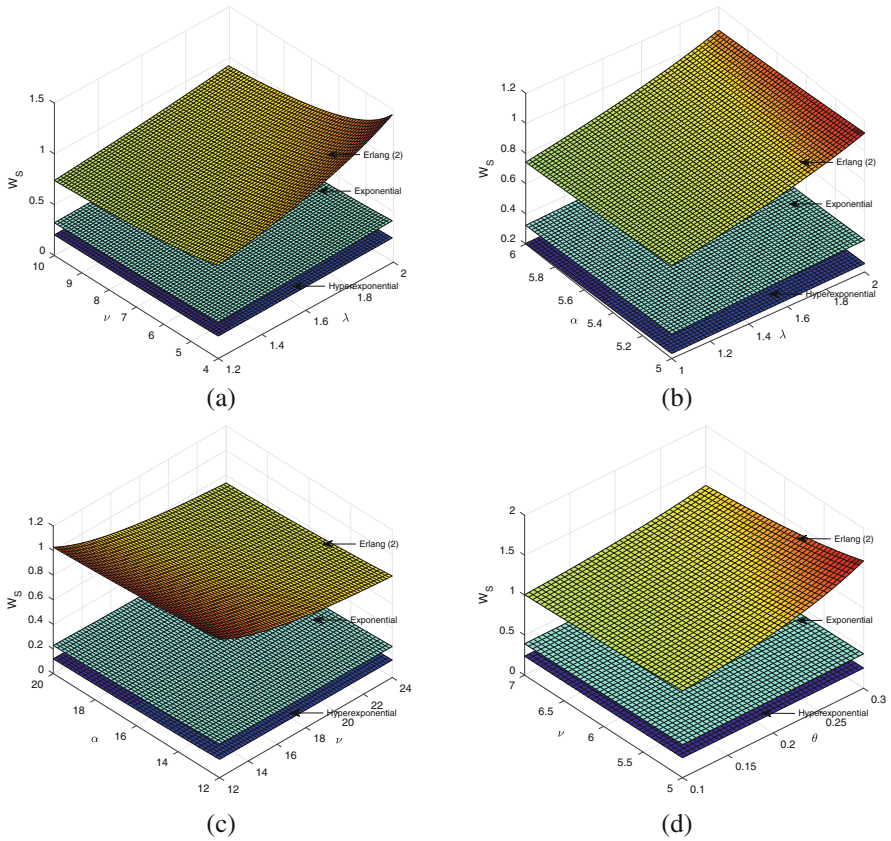


Fig. 2. (a) W_S versus (λ, ν) for $(\alpha, \mu, \theta, p) = (6, 9, 0.5, 0.6)$, (b) W_S versus (λ, α) for $(\nu, \mu, \theta, p) = (7, 9, 0.5, 0.6)$, (c) W_S versus (ν, α) for $(\lambda, \mu, \theta, p) = (3, 9, 0.5, 0.6)$, (d) W_S versus (θ, ν) for $(\lambda, \alpha, \mu, p) = (3, 6, 9, 0.6)$

Figures 1(a)–(d) display the trends of the probability, P_0 , of the system being empty. We have presented three surfaces when the service time, retrieval time and setup time all follow either exponential distributions or 2-stage Erlangian distributions or hyperexponential distributions as mentioned before. The effect of varying values of ν and λ on P_0 is displayed in Fig. 1(a) for the set of parameters $(\alpha, \mu, \theta, p) = (6, 9, 0.5, 0.6)$. We have noticed that all three surfaces of the descriptor P_0 increase for increasing values of ν whereas they decrease for increasing values of λ as is to be expected. Referring to Fig. 1(b), it can be seen that as λ increases, all surfaces of the descriptor P_0 appear to decrease gradually but they appear to increase as α increases where we fix $(\nu, \mu, \theta, p) = (7, 9, 0.5, 0.6)$. The possible explanation of α effect on P_0 is that a large α results in a short mean setup time $E(S)$ and thus the descriptor P_0 increases. Figure 1(c) depicts the influence of ν and α on P_0 for the chosen system parameters $(\lambda, \mu, \theta, p) = (3, 9, 0.5, 0.6)$. We see that all three surfaces

of P_0 are increasing apparently for increasing values of α while for increasing values of μ , they increase at the slower rate. Finally, our next numerical example is to illustrate the effect of varying parametric values of θ and ν on P_0 by fixing $(\lambda, \alpha, \mu, p) = (3, 6, 9, 0.6)$. A quick examination of Fig. 1(d) reveals that three surfaces of the descriptor P_0 decrease moderately in θ but they increase apparently with ν . In addition, in all Figs. 1(a)–(d), the surface for hyperexponential distribution case appears to be higher than the surfaces for exponential and 2-stage Erlangian distributions cases.

Next, we turn our attention to the study the behaviour of the average waiting time W_S of a customer in the system for varying values of the system parameters. The impact of λ and ν on measure W_S is reported in Fig. 2(a) for $(\alpha, \mu, \theta, p) = (6, 9, 0.5, 0.6)$. We can see that the surfaces of W_S appear to increase as λ increases, whereas they appear to decrease as ν increases. The influence of the parameters λ and α on the measure W_S is sketched in Fig. 2(b) for $(\nu, \mu, \theta, p) = (7, 9, 0.5, 0.6)$. It can be observed from Fig. 2(b) that the surfaces of W_S increase significantly with λ whereas they decrease in α at the slower rate. Next, examination of Fig. 2(c) reveals, an unsurprising result, that the surfaces of W_S decrease at the slower rate for increasing parametric values of both ν and α . Finally, the variation of W_S with respect to θ and ν is depicted in Fig. 2(d) by taking $(\lambda, \alpha, \mu, p) = (3, 6, 9, 0.6)$. As a result, all three surfaces of W_S increase with θ but decrease in ν . This is again expected. With regard to this measure, we also see, by comparison, that the surface for 2-stage Erlangian case appears to be higher than the surfaces for exponential and hyperexponential cases.

6 Conclusion

In this paper, we analyzed the preemptive priority LCFS single server retrieval queueing system with general retrieval and setup times. Using the mean drift analysis, the necessary and sufficient condition for the system to be stable is provided. Further, we obtained the joint steady-state distribution of the server states and the number of customers in the orbit along with some important system performance measures through the supplementary variable technique. Numerical illustrations are also presented to show effect of the system parameters on the system characteristics.

References

1. Agalarov, Y.M.: On a numerical method for calculating stationary characteristics of the switching node with retrieval transmissions. *Autom. Remote Control.* **72**, 88–98 (2011)
2. Artalejo, J.R.: Accessible bibliography on retrieval queues: progress in 2000–2009. *Math. Comput. Model.* **51**, 1071–1081 (2010)
3. Artalejo, J.R., Gomez-Corral, A.: *Retrial Queueing Systems: A Computational Approach*. Springer, Heidelberg (2008). <https://doi.org/10.1007/978-3-540-78725-9>
4. Atencia, I., Moreno, P.: A single-server retrieval queue with general retrieval times and Bernoulli schedule. *Appl. Math. Comput.* **162**, 855–880 (2005)

5. Choi, B.D., Park, K.K., Pearce, C.E.M.: An $M/M/1$ retrial queue with control policy and general retrial times. *Queueing Syst.* **14**, 275–292 (1993)
6. Conti, M., Gregori, E., Lenzi, L.: *Metropolitan Area Networks*. Springer, London (1997). <https://doi.org/10.1007/978-1-4471-0909-9>
7. Cooper, R.B.: *Introduction to Queueing Theory*. North-Holland, New York (1981)
8. Falin, G.I.: A survey of retrial queues. *Queueing Syst.* **7**, 127–168 (1990)
9. Falin, G.I., Templeton, J.G.C.: *Retrial Queues*. Chapman and Hall, London (1997)
10. Gao, S., Wang, J.: Performance and reliability analysis of an $M/G/1-G$ retrial queue with orbital search and non-persistent customers. *Eur. J. Oper. Res.* **236**, 561–572 (2014)
11. Gebrehiwot, M.E., Aalto, S., Lassila, P.: Energy-performance trade-off for processor sharing queues with setup delay. *Oper. Res. Lett.* **44**, 101–106 (2016)
12. Gomez-Corral, A.: Stochastic analysis of a single server retrial queue with general retrial times. *Nav. Res. Logist.* **46**, 561–581 (1999)
13. Harchol-Balter, M.: *Performance Modeling and Design of Computer System: Queueing Theory in Action*. Cambridge University Press, Cambridge (2013)
14. Li, T., Zhang, L.: An $M/G/1$ retrial G-queue with general retrial times and working breakdowns. *Math. Comput. Appl.* **22**, 1–17 (2017)
15. Ke, J.C., Chang, F.M.: $M^{[x]}/(G1, G2)/1$ retrial queue under Bernoulli vacation schedules with general repeated attempts and starting failures. *Appl. Math. Model.* **33**, 3186–3196 (2009)
16. Krishna Kumar, B., Rukmani, R., Thangaraj, V., Udo Krieger, R.: A single server retrial queue with Bernoulli feedback and collisions. *J. Stat. Theor. Pract.* **4**, 243–260 (2010)
17. Niu, Z., Takahashi, Y.: A finite-capacity queue with exhaustive vacation/close-down/setup times and Markovian arrival processes. *Queueing Syst.* **31**, 1–23 (1999)
18. Pakes, A.G.: Some conditions for ergodicity and recurrence of Markov chains. *Oper. Res.* **17**, 1058–1061 (1969)
19. Sakai, Y., Takahashi, Y., Hasegawa, T.: A composite queue with vacation/setup/close-down times for SVCC in IP over ATM networks. *J. Oper. Res. Soc. Jpn.* **41**, 68–78 (1998)
20. Sennott, L.I., Hamblet, P.A., Tweedie, R.L.: Mean drifts and the non-ergodicity of Markov chains. *Oper. Res.* **31**, 783–789 (1983)
21. Wolff, W.R.: Poisson arrivals see time averages. *Oper. Res.* **30**, 223–231 (1982)
22. Zayats, O., Korenevskaya, M., Ilyashenko, A., Lukashin, A.: Retrial queueing systems in series of space experiments “Kontur”. *Procedia Comput. Sci.* **103**, 562–568 (2017)



Method of Generating Functions for Performance Characteristic Analysis of the Polling Systems with Adaptive Polling and Gated Service

Olga V. Semenova^{1,2(✉)} and Duy T. Bui^{1,2}

¹ Institute of Control Sciences of Russian Academy of Sciences,
Profsoyuznaya Street 65, Moscow 117997, Russia

olgasmnv@gmail.com

² Moscow Institute of Physics and Technology, Institutskiy per. 9,
Dolgoprudny, Moscow Region 141701, Russia

duytan@phystech.edu

Abstract. In the paper, we consider a polling system with a cyclic adaptive polling order. The server polls the queues in a cyclic way but skips (does not visit) those that were empty when polling them in the previous cycle. We apply the generating function method to derive first and second order moments of the stationary state distribution of the queue length at the polling moments that allows calculating the mean delay in queues. Also we provide numerical results.

Keywords: Polling systems · Adaptive polling order
Gated service · Generating function method

1 Introduction

Polling systems are queuing systems with multiple queues (or multiple flows of customers) attended by a single server (or multiple servers). The server visits (polls) the queues according to a certain polling rule and serves their customers. The systematization and generalization of theoretical results obtained in the field of polling systems research prior to 1985 was carried out in the monograph by Takagi [1], the classification of polling systems is also presented in [2]. The further development of theoretical results of polling system analysis up to 1995 is presented by Borst in [3], and the papers published in 1996–2009 are reviewed in [4, 5]. The book [6] is devoted to the generalization and systematization of models and methods for studying stochastic systems with cyclic polling and their application to design the broadband wireless networks.

O. V. Semenova—The publication has been prepared with the partial support of the Russian Science Foundation and DTS (India) according to the research project No. 16-49-02021.

Depending on the number of queues in the system, the polling systems may be *discrete* (the number of waiting places is finite or countable) and *continuous* (the number of waiting places is more than countable). In the latter case, consideration is given to the systems where the customers are located on a circle or in an n -dimensional domain.

Discrete polling systems are characterized by the number of queues, their capacity (the number of the waiting places), the number of servers, the processes of customer arrival and service, durations of server switchover between the queues, and also the order and discipline of queue service. We assume that all queues are numerated from 1 to N where $N \geq 2$ is the number of queues in the system. The queue with the number $i = \overline{1, N}$ will be denoted by Q_i .

The *polling order* or visit order is the rule used by the server to choose the next queue. The polling order can be either static or dynamic. With the *static* order, the rule of choosing queues remains invariable in the entire course of system operation. With the *dynamic* order, the queue is chosen for service at certain decision-making moments on the basis of complete or partial information about system state.

The *static order* can be

1. *Cyclic* order where the server polls the queues in the order $Q_1, Q_2, \dots, Q_N, Q_1, Q_2, \dots, Q_N, \dots$. These polling systems are called the *cyclic* systems.
2. *Adaptive cyclic* order where a server polls the queues in a cyclic way but skips (does not visit) those that were empty when polling them in the previous cycle.
3. *Periodic* order where the server polls the queues in the order $Q_{T(1)}, Q_{T(2)}, \dots, Q_{T(M)}, Q_{T(1)}, Q_{T(2)}, \dots, Q_{T(M)}, \dots$ which is characterized by the so-called polling table $(T(1), T(2), \dots, T(M))$ of length M ($M \geq N$), $T(i) \in \{1, \dots, N\}$, $i = \overline{1, M}$. It is assumed that the polling table comprises the numbers of all system queues.
4. *Random* order where the queue Q_i is taken for service with the probability p_i , $i = \overline{1, N}$, $\sum_{i=1}^N p_i = 1$. Feasible is another variant of choosing the queue where after polling the queue Q_i the server switches over to Q_j with the probability p_{ij} , $i, j = \overline{1, N}$, $\sum_{j=1}^N p_{ij} = 1$, $i = \overline{1, N}$.
5. *Priority* order where the system has queues of different priorities and some queue may be served only if all higher-priority queues have no customers.

Time periods called *cycles* are specified in operation of the cyclic or periodic polling system. For the cyclic polling systems, the cycle is the time required for the server to serve the queues from Q_1 to Q_N . For the periodic polling systems, the cycle is the time required to serve queues from $Q_{T(1)}$ to $Q_{T(M)}$.

The *queue service discipline* is the number of customers served by the server in one polling. Within the queue, the customers are served in the order defined by the *customer service discipline* which most frequently lies in serving them in the arrival order. The classical service disciplines of the queue (say Q_i) are the following:

1. *Exhaustive*, where the server serves customers until the queue is emptied.
2. *Gated*, where the server serves only those customers that waited in the queue at its polling instant. If the server serves only those customers which waited in the queue by the beginning of the cycle, this discipline is called the *globally-gated* discipline.
3. *l_i -limited*, where the number of customers that can be served by the server is limited by l_i , $l_i \geq 1$. Limited disciplines can be of exhaustive or gated type. Under the limited exhaustive discipline, a server serves customers in the queue until either l_i customers are served or the queue becomes empty whichever occurs first. The limited gated discipline implies that the queue is served until either l_i customers are served or all customers which were present at the queue polling moment are served whichever occurs first. The partial case $l_i = 1$ is often called non-exhaustive service in the literature.
4. *l_i -decrementing*, where the server serves queued customers until the queue length is decremented by l_i as compared to the polling instant, $l_i \geq 1$. It can also be exhaustive or gated as described above. The case $l_i = 1$ is often called semi-exhaustive service.
5. *T-limited service* when the time the server spends in the queue is limited.
6. *Threshold service* when the server visits the queue if its length exceeds the given level (threshold).
7. The *random* discipline when the number of customers that can be served by the server in the queue Q_i is defined by the value of the discrete random variable ξ_i with the distribution law $\{a_j^i, j \geq 1\}$ which can vary with each visit to the queue. Some of the random disciplines are:
 - (a) *Binomial* discipline with the random variable ξ_i having the binomial distribution with the parameters X_i and p_i , where X_i is the number of customers queued in Q_i at the polling instant and p_i is some number, $0 < p_i \leq 1$. For this discipline, $a_j^i = C_{X_i}^j p_i^j (1 - p_i)^{X_i - j}$, $j = \overline{1, X_i}$, $a_j^i = 0$ for $j > X_i$.
 - (b) *Bernoulli* discipline where the first customer queued in Q_i is served with the probability 1 and each subsequent customer, with a given probability p_i . The server leaves the queue with the probability $1 - p_i$. For this discipline, $a_j^i = p_i^{j-1}$, $j \geq 1$.

2 Adaptive Cyclic Polling

An adaptive cyclic polling of queues was presented in paper [7]. For such a polling order, a server polls the queues in a cyclic way but skips (does not visit) those that were empty when polling them in the previous cycle. All queues skipped in the current cycle will be visited by the server in the next cycle of polling. The analysis of such a discipline, or the polling order, requires information about the state of queues in the previous cycle which considerably complicates the analysis, and in [7] we proposed only an approximate algorithm for calculating the characteristics of adaptive polling systems. Below we provide exact analysis by means of the PGF (Probability Generating Function) method, see [8–11], and for the sake of brevity, we consider the case of gated service only.

Consider a polling system with a single server attending N queues of $M/G/1$ -type with gated service. The i th queue Q_i has a Poisson input of arrivals of intensity λ_i . The service time in the queue has the distribution function $B_i(t)$ with the first and second initial moments $b_i = \int_0^\infty t dB_i(t)$ and $b_i^{(2)} = \int_0^\infty t^2 dB_i(t)$, and the Laplace-Stieltjes transform (LST) $\tilde{B}_i(s) = \int_0^\infty e^{-st} dB_i(t)$. The time the server switches to queue Q_i has the distribution function $S_i(t)$ with the first and second initial moments s_i and $s_i^{(2)}$, respectively, and the LST $\tilde{S}_i(s)$. We assume that in case the server meets N queues empty consecutively at their polling moment (starting from any queue) it stops at the N th empty queue and takes a vacation having the distribution function $H(t)$ with the first and second initial moments β and $\beta^{(2)}$ and the LST $\tilde{H}(s)$. After a vacation, the server leaves its current position and switches to the next queue. The polling procedure is repeated again. The service of queues is gated, i.e. the server serves only those customers that were present in a queue at its polling moment.

The stability condition for the polling system considered is $\rho = \sum_{i=1}^N \rho_i < 1$ where $\rho_i = \lambda_i b_i$ is the queue Q_i load.

The cycle time is supposed to be the time the server spends polling queues from Q_1 and Q_N including a vacation time (if the server stops polling). The mean cycle time is given by formula

$$C = \frac{\sum_{i=1}^N s_i u_i + \beta \prod_{i=1}^N (1 - u_i)}{1 - \rho} \tag{1}$$

where u_i is the probability that queue Q_i is polled in the cycle.

To find the probabilities $u_i, i = \overline{1, N}$, we consider the random variables $c_n^{(i)}, i = \overline{1, N}$ where $c_n^{(i)}$ is a status of queue Q_i in the n th polling cycle, $c_n^{(i)} = 1$ if the queue is polled in the cycle, and $c_n^{(i)} = 0$ if the queue is skipped. The random variable $c_n^{(i)}$ changes its values as follows:

$$c_n^{(i)} = \begin{cases} 0, & c_n^{(i)} = 1 \text{ or } N_{n-1}^{(i)} = 0, \\ 1, & \text{otherwise,} \end{cases} \quad j \geq 1.$$

where $N_n^{(i)}$ is the number of customers in the queue Q_i at its polling moment at the n th cycle.

Denote by $x_{kl}^{(i)}$ the probability of the random variable $c_n^{(i)}$ one-step transition from state k to state $l, k, l = \overline{0, 1}$,

$$x_{00}^{(i)} = 0, \quad x_{01}^{(i)} = 1, \\ x_{10}^{(i)} = P\{N_{n-1}^{(i)} = 0\} = \pi_0^{(i)}, \quad x_{11}^{(i)} = P\{N_{n-1}^{(i)} \neq 0\} = 1 - \pi_0^{(i)}$$

where $\pi_0^{(i)}$ is the probability that the queue Q_i is empty when it is polled.

The stationary state probability $u_i = \lim_{n \rightarrow \infty} P(c_n^{(i)} = 1)$ that the i th queue is polled in an arbitrary cycle can be calculated from the balance equation

$$u_i = u_i x_{01}^{(i)} + (1 - u_i) x_{11}^{(i)}$$

which results in

$$u_i = \frac{1}{1 + \pi_0^{(i)}}.$$

Note however, that for the second approach to calculating the probability u_i we need to know exactly or estimate the value $\pi_0^{(i)}$. In the present paper, we suppose that $\pi_0^{(i)} = 1 - e^{\lambda_i C}$.

$$u_i = \frac{1}{1 + e^{-\lambda_i C}}, \quad i = \overline{1, N}. \tag{2}$$

Equations (1)–(2) give the system to find the mean cycle time and the probabilities $u_i, i = \overline{1, N}$.

3 Method of Probability Generating Functions

Let X_i^j be the number of customers present in the queue Q_j when server polls the queue $Q_i, i, j = \overline{1, N}, A_i(T)$ be the number of Poisson arrivals to the queue Q_i during a random time interval of length T, B_{ik} be the service time of the k th customer in the queue Q_i, S_i be switchover time to $Q_i, i = \overline{1, N}$, and V be a vacation time.

For the gated service, the evolution of the system state is given by

$$\begin{aligned} X_{i+1}^j | M_{i+1}^{(0)} &= \begin{cases} X_i^j + A_j \left(\sum_{k=1}^{X_i^j} B_{i,k} + S_{i+1} \right), & i \neq j, \\ A_j \left(\sum_{k=1}^{X_i^j} B_{i,k} + S_{i+1} \right), & i = j, \end{cases} \\ X_{i+1}^j | M_{i+1}^{(1)} &= \begin{cases} X_{i-1}^j + A_j \left(\sum_{k=1}^{X_{i-1}^{j-1}} B_{i-1,k} + S_{i+1} \right), & i-1 \neq j, \\ A_j \left(\sum_{k=1}^{X_{i-1}^{j-1}} B_{i-1,k} + S_{i+1} \right), & i-1 = j, \end{cases} \\ X_{i+1}^j | M_{i+1}^{(2)} &= \begin{cases} X_{i-2}^j + A_j \left(\sum_{k=1}^{X_{i-2}^{j-2}} B_{i-2,k} + S_{i+1} \right), & i-2 \neq j, \\ A_j \left(\sum_{k=1}^{X_{i-2}^{j-2}} B_{i-2,k} + S_{i+1} \right), & i-2 = j, \end{cases} \\ \dots \\ X_{i+1}^j | M_{i+1}^{(N-1)} &= \begin{cases} X_{i-N+1}^j + A_j \left(\sum_{k=1}^{X_{i-N+1}^{j-N+1}} B_{i-N+1,k} + S_{i+1} \right), & i-N+1 \neq j, \\ A_j \left(\sum_{k=1}^{X_{i-N+1}^{j-N+1}} B_{i-N+1,k} + S_{i+1} \right), & i-N+1 = j, \end{cases} \\ X_{i+1}^j | M_{i+1}^{(N)} &= \begin{cases} X_{i-N}^j + A_j \left(\sum_{k=1}^{X_{i-N}^{j-N}} B_{i-N,k} + S_{i+1} + V \right), & i-N \neq j, \\ A_j \left(\sum_{k=1}^{X_{i-N}^{j-N}} B_{i-N,k} + S_{i+1} + V \right), & i-N = j, \end{cases} \end{aligned} \tag{3}$$

where $M_{i+1}^{(j)}$ is the event that the server skipped exactly j queues before polling the current queue Q_{i+1} , i.e. the previously polled queue was Q_{i-j} if $i > j$, and Q_{i-j+N} otherwise, $j = \overline{0, N}$. If $i - k < 0$ we assume that $X_{i-k}^j = X_{i-k+N}^j$.

For the fixed i , the probabilities of $M_{i+1}^{(j)}$, $j = \overline{0, N}$ are calculated as follows:

$$\begin{aligned}
 \mathbf{P}\{M_{i+1}^{(0)}\} &= u_i, \\
 \mathbf{P}\{M_{i+1}^{(1)}\} &= (1 - u_i)u_{i-1}, \dots, \\
 \mathbf{P}\{M_{i+1}^{(i-1)}\} &= (1 - u_i)(1 - u_{i-1}) \cdots (1 - u_2)u_1, \\
 \mathbf{P}\{M_{i+1}^{(i)}\} &= \prod_{k=1}^i (1 - u_k)u_N, \\
 \mathbf{P}\{M_{i+1}^{(i+1)}\} &= \prod_{k=1}^i (1 - u_k)(1 - u_N)u_{N-1}, \dots
 \end{aligned} \tag{4}$$

$$\mathbf{P}\{M_{i+1}^{(N)}\} = \prod_{k=1}^i (1 - u_k)(1 - u_N) \cdots (1 - u_{i-1}) = \prod_{k=1}^N (1 - u_k).$$

It is easy to see that $\sum_{j=0}^N \mathbf{P}\{M_{i+1}^{(j)}\} = 1$.

Let $p_i(r_1, r_2, \dots, r_N)$ be the stationary probability that Q_j has r_j customers at the polling instant of Q_i , $r_j \geq 0$, $i, j = \overline{1, N}$. Consider the probability generating functions (PGFs).

$$F_i(\mathbf{z}) = F_i(z_1, z_2, \dots, z_N) = \sum_{r_1=0}^{\infty} \sum_{r_2=0}^{\infty} \cdots \sum_{r_N=0}^{\infty} p_i(r_1, r_2, \dots, r_N) z_1^{r_1} \cdots z_N^{r_N}.$$

They can be also presented as

$$F_i(\mathbf{z}) = \mathbf{E} \left[\prod_{j=1}^N z_j^{X_i^j} \right], \quad i = \overline{1, N},$$

where \mathbf{E} is the expectation. While using (3), we get

$$\begin{aligned}
 F_i(\mathbf{z}) &= u_i \mathbf{E} \left[\prod_{j=1}^N z_j^{X_{i+1}^j} \middle| M_{i+1}^{(0)} \right] + (1 - u_i)u_{i-1} \mathbf{E} \left[\prod_{j=1}^N z_j^{X_{i+1}^j} \middle| M_{i+1}^{(1)} \right] + \dots \\
 &\quad + (1 - u_1) \cdots (1 - u_N) \mathbf{E} \left[\prod_{j=1}^N z_j^{X_{i+1}^j} \middle| M_{i+1}^{(N)} \right] \\
 &= u_i M_{i+1}^{(0)}(\mathbf{z}) + (1 - u_i)u_{i-1} M_{i+1}^{(1)}(\mathbf{z}) + \dots + (1 - u_1) \cdots (1 - u_{N-1})u_N M_{i+1}^{(N-1)}(\mathbf{z}) \\
 &\quad + (1 - u_1) \cdots (1 - u_N) M_{i+1}^{(N)}(\mathbf{z}) \tag{5}
 \end{aligned}$$

where $u_{i-N} = u_i$, $F_{i-N}(\mathbf{z}) = F_i(\mathbf{z})$, $M_{i+1}^{(l)}(\mathbf{z}) = \mathbf{E} \left[\prod_{j=1}^N z_j^{X_{i+1}^j} \middle| M_{i+1}^{(l)} \right]$, $l = \overline{0, N}$.

Here we have

$$\begin{aligned}
 M_{i+1}^{(l)}(\mathbf{z}) &= \mathbf{E} \left[\prod_{j=1}^N z_j^{X_{i+1}^j} \middle| M_{i+1}^{(l)} \right] = \mathbf{E}_{X_i} \left[\prod_{j=1}^N z_j^{X_i^j} \mathbf{E} \left[\prod_{j=1}^N z_j^{A_j \left(\sum_{k=1}^{X_i^j} B_{i,k} \right)} \middle| \mathbf{X}_i \right] \right] \\
 &\times \mathbf{E} \left[\prod_{j=1}^N z_j^{A_j(S_{i+1})} \right] \tag{6}
 \end{aligned}$$

where $\mathbf{X}_i = (X_i^1, X_i^2, \dots, X_i^N)$, $\mathbf{E}[\cdot | X_i]$ is the conditional expectation.

Now we find the expectation of the random variable $z_j^{A_j(T)}$ where T is a random variable with distribution function $D(T)$:

$$\begin{aligned}
 \mathbf{E} \left[z_j^{A_j(T)} \right] &= \int_0^\infty \mathbf{E} \left[z_j^{A_j(t)} \right] dD(t) = \int_0^\infty \sum_{k=0}^\infty P(A_j(t) = k) z_j^k dD(t) \\
 &= \int_0^\infty \sum_0^\infty \frac{(\lambda_j t)^k}{k!} e^{-\lambda_j t} z_j^k dD(t) = \int_0^\infty e^{-\lambda_j t(1-z_j)} dD(t) = \tilde{D}(\lambda_j(1-z_j)) \tag{7}
 \end{aligned}$$

where $\tilde{D}(w)$ is the LST of $D(t)$. Then we have

$$\mathbf{E} \left[z_j^{A_j(T)} \right] = \tilde{D} \left(\sum_{j=1}^N \lambda_j(1-z_j) \right). \tag{8}$$

With formulas (4) and (8), we get the functions $M_{i+1}^{(l)}(\mathbf{z})$, $l = \overline{0, N}$ as follows:

$$\begin{aligned}
 M_{i+1}^{(l)}(\mathbf{z}) &= F_{i-l} \left(z_1, z_2, \dots, z_{i-l-1}, \tilde{B}_{i-l} \left(\sum_{j=1}^N \lambda_j(1-z_j) \right), z_{i-l+1}, \dots, z_N \right) \\
 &\times \tilde{S}_{i-l+1} \left[\sum_{j=1}^N \lambda_j(1-z_j) \right], \quad l = \overline{0, N-1}, \\
 M_{i+1}^{(N)}(\mathbf{z}) &= F_{i-N} \left(z_1, z_2, \dots, z_{i-N-1}, \tilde{B}_{i-N} \left(\sum_{j=1}^N \lambda_j(1-z_j) \right), z_{i-N+1}, \dots, z_N \right) \\
 &\times \tilde{S}_{i-N+1} \left[\sum_{j=1}^N \lambda_j(1-z_j) \right] \tilde{H} \left(\sum_{j=1}^N \lambda_j(1-z_j) \right). \tag{9}
 \end{aligned}$$

The mean number of customers $f_i(j) = \mathbf{E}[X_i^j]$ in queue Q_j at a polling moment of Q_i is given by

$$f_i(j) = \mathbf{E} \left[X_i^j \right] = \left. \frac{\partial F_i(\mathbf{z})}{\partial z_j} \right|_{\mathbf{z}=\mathbf{1}} \tag{10}$$

where $\mathbf{1} = (1, 1, \dots, 1)$.

First, we differentiate equations (9) at $\mathbf{z} = \mathbf{1}$ for $l = \overline{0, N-1}$. Denote by $\nu = \sum_{j=1}^N \lambda_j(1 - z_j)$ then from (9) we have

$$\begin{aligned} \left. \frac{\partial M_{i+1}^{(l)}(\mathbf{z})}{\partial z_j} \right|_{\mathbf{z}=\mathbf{1}} &= \frac{d}{dz_j} F_{i-l} \left(z_1, \dots, z_{i-l-1}, \tilde{B}_{i-l}(\nu), z_{i-l+1}, \dots, z_N \right) \Big|_{\mathbf{z}=\mathbf{1}} \tilde{S}_{i-l+1}(0) \\ &\quad + F_{i-l} \left(z_1, \dots, z_{i-l-1}, \tilde{B}_{i-l}(\nu), z_{i-l+1}, \dots, z_N \right) \frac{d}{dz_j} j \tilde{S}_{i-l+1}(\nu) \Big|_{\mathbf{z}=\mathbf{1}}. \end{aligned}$$

Then,

$$\begin{aligned} \left. \frac{\partial M_{i+1}^{(l)}(\mathbf{z})}{\partial z_j} \right|_{\mathbf{z}=\mathbf{1}} &= \left[\frac{\partial}{\partial z_j} F_{i-l} \left(z_1, \dots, z_{i-l-1}, \tilde{B}_{i-l}(\nu), z_{i-l+1}, \dots, z_N \right) \right. \quad (11) \\ &\quad \left. + \frac{\partial}{\partial \tilde{B}_{i-l}} F_{i-l} \left(z_1, \dots, z_{i-l-1}, \tilde{B}_{i-l}(\nu), z_{i-l+1}, \dots, z_N \right) \frac{d\tilde{B}_{i-l}(\nu)}{d\nu} \frac{\partial \nu}{\partial z_j} \right] \Big|_{\mathbf{z}=\mathbf{1}} \\ &\quad \times \tilde{S}_{i-l+1}(0) \\ &\quad + F_{i-l+1} \left(z_1, \dots, z_{i-l-1}, \tilde{B}_{i-l}(\nu), z_{i-l+1}, \dots, z_N \right) \frac{d\tilde{S}_{i-l+1}(\nu)}{d\nu} \frac{\partial \nu}{\partial z_j} \Big|_{\mathbf{z}=\mathbf{1}}. \end{aligned}$$

Note that $F_i(\mathbf{1}) = 1$, $\nu(\mathbf{1}) = 0$ and $\tilde{B}_i(0) = \tilde{S}_i(0) = 1$. Also

$$\left. \frac{d\tilde{B}_i(\nu)}{d\nu} \right|_{\nu=0} = - \int_0^\infty t e^{-\nu t} dB_i(t) = -b_i, \quad \left. \frac{\partial \nu}{\partial z_i} \right|_{\mathbf{z}=\mathbf{1}} = -\lambda_i, \quad i = \overline{1, N}.$$

Then using the equations above, the relation (11) takes the form

$$\begin{aligned} \left. \frac{\partial M_{i+1}^{(l)}(\mathbf{z})}{\partial z_j} \right|_{\mathbf{z}=\mathbf{1}} &= f_i(j) + \lambda_j b_i f_i(i) + \lambda_j s_{i+1}, \quad j \neq i, \\ \left. \frac{\partial M_{i+1}^{(l)}(\mathbf{z})}{\partial z_i} \right|_{\mathbf{z}=\mathbf{1}} &= \lambda_i b_i f_i(i) + \lambda_i s_{i+1}. \end{aligned} \quad (12)$$

The case $l = N$ is considered similarly.

Differentiating equations (5) by means of (12), we get the following system of linear algebraic equations for $f_i(j)$, $i, j = \overline{1, N}$:

$$\begin{aligned} f_{i+1}(j) &= u_i [k_{i,j} f_i(j) + \lambda_j b_i f_i(i) + \lambda_j s_{i+1}] \\ &\quad + (1 - u_i) u_{i-1} [k_{i-1,j} f_{i-1}(j) + \lambda_j b_{i-1} f_{i-1}(i-1) + \lambda_j s_{i+1}] + \dots \\ &\quad + (1 - u_1) \dots (1 - u_N) [k_{i-N,j} f_{i-N}(j) + \lambda_j b_{i-N} f_{i-N}(i-N) + \lambda_j (s_{i+1} + \beta)] \end{aligned} \quad (13)$$

where $k_{i,j} = 0$ if $i = j$ and $k_{i,j} = 1$ if $i \neq j$. This system has the unique solution and can be solved numerically for N^2 unknowns.

The second order moments of the random variables $X_i^j, i, j = \overline{1, N}$ are calculated as

$$f_i(j, k) = \mathbf{E} [X_i^j X_i^k] = \left. \frac{\partial^2 F_i(\mathbf{z})}{\partial z_j \partial z_k} \right|_{\mathbf{z}=\mathbf{1}}, \tag{14}$$

$$f_i(i, i) = \mathbf{E} [X_i^i (X_i^i - 1)] = \left. \frac{\partial^2 F_i(\mathbf{z})}{\partial z_i^2} \right|_{\mathbf{z}=\mathbf{1}}. \tag{15}$$

By differentiating (5), we get

$$f_{i+1}(j, k) = \left[u_i \frac{\partial^2 M_{i+1}^{(0)}(\mathbf{z})}{\partial z_j \partial z_k} + (1 - u_i) u_{i-1} \frac{\partial^2 M_{i+1}^{(1)}(\mathbf{z})}{\partial z_j \partial z_k} + \dots \right. \\ \left. + (1 - u_1) \dots (1 - u_N) \frac{\partial^2 M_{i+1}^{(N)}(\mathbf{z})}{\partial z_j \partial z_k} \right] \Bigg|_{\mathbf{z}=\mathbf{1}} \tag{16}$$

where $\frac{\partial^2 M_{i+1}^{(l)}(\mathbf{z})}{\partial z_j \partial z_k}, l = \overline{0, N}$ are calculated as

$$\left. \frac{\partial^2 M_{i+1}^{(0)}(\mathbf{z})}{\partial z_j \partial z_k} \right|_{\mathbf{z}=\mathbf{1}} = \lambda_j \lambda_k s_{i+1}^{(2)} + s_{i+1} \lambda_k f_i(j) + s_{i+1} \lambda_j f_i(k) \\ + f_i(i) \lambda_j \lambda_k [2b_i s_{i+1} + b_i^{(2)}] + f_i(i, i) \lambda_j \lambda_k b_i^2 + f_i(i, j) b_i \lambda_k \\ + f_i(i, k) b_i \lambda_j + f_i(j, k), \quad i \neq j, i \neq k, \\ \left. \frac{\partial^2 M_{i+1}^{(0)}(\mathbf{z})}{\partial z_j \partial z_k} \right|_{\mathbf{z}=\mathbf{1}} = \lambda_i \lambda_j s_{i+1}^{(2)} + s_{i+1} \lambda_i f_i(j) + f_i(i) \lambda_i \lambda_j [2b_i s_{i+1} + b_i^{(2)}] \\ + f_i(i, j) b_i \lambda_i + \lambda_i \lambda_j b_i^2 f_i(i, i), \quad i \neq j, \\ \left. \frac{\partial^2 M_{i+1}^{(0)}(\mathbf{z})}{\partial z_j \partial z_k} \right|_{\mathbf{z}=\mathbf{1}} = \lambda_i^2 s_{i+1}^{(2)} + f_i(i) \lambda_i^2 [2b_i s_{i+1} + b_i^{(2)}] + \lambda_i^2 b_i^2 f_i(i, i), i, j, k = \overline{1, N}. \tag{17}$$

...

$$\left. \frac{\partial^2 M_{i+1}^{(N)}(\mathbf{z})}{\partial z_j \partial z_k} \right|_{\mathbf{z}=\mathbf{1}} = \lambda_j \lambda_k (s_{i+1}^{(2)} + \beta^{(2)}) + (s_{i+1} + \beta) \lambda_k f_i(j) \\ + (s_{i+1} + \beta) \lambda_j f_i(k) + f_i(i) \lambda_j \lambda_k [2b_i (s_{i+1} + \beta) + b_i^{(2)}] + f_i(i, i) \lambda_j \lambda_k b_i^2 \\ + f_i(i, j) b_i \lambda_k + f_i(i, k) b_i \lambda_j + f_i(j, k), \quad i \neq j, i \neq k,$$

$$\left. \frac{\partial^2 M_{i+1}^{(N)}(\mathbf{z})}{\partial z_j \partial z_k} \right|_{\mathbf{z}=\mathbf{1}} = \lambda_i \lambda_j (s_{i+1}^{(2)} + \beta^{(2)}) + (s_{i+1} + \beta) \lambda_i f_i(j) + f_i(i, j) b_i \lambda_i \\ + f_i(i) \lambda_i \lambda_j [2b_i (s_{i+1} + \beta) + b_i^{(2)}] + \lambda_i \lambda_j b_i^2 f_i(i, i), \quad i \neq j, \\ \left. \frac{\partial^2 M_{i+1}^{(N)}(\mathbf{z})}{\partial z_j \partial z_k} \right|_{\mathbf{z}=\mathbf{1}} = \lambda_i^2 (s_{i+1}^{(2)} + \beta^{(2)}) + f_i(i) \lambda_i^2 [2b_i (s_{i+1} + \beta) + b_i^{(2)}] \\ + \lambda_i^2 b_i^2 f_i(i, i), \quad i, j, k = \overline{1, N}.$$

The relations (14)–(17) give the system of linear algebraic equations to calculate the second-order moments $f_i(j, k)$, $i, j, k = \overline{1, N}$ which allow to find the mean waiting time W_i in queue Q_i by formula (see [8]).

$$W_i = \frac{f_i(i, i) - f_i}{2\lambda_i f_i} (1 + \rho_i), i = \overline{1, N}. \tag{18}$$

4 Numerical Results

To illustrate the results obtained, we shortly present the numerical results compared to the simulation.

Let the system have two queues. The service time has an exponential distribution and the average service time is $b = 0.01$, the server switching time has an exponential distribution and the average time to switch between queues is $s = 0.001$, the vacation time has an exponential distribution and the average vacation time of the server can be $\beta = 0$ or $\beta = 0.002$, the Poisson flow to each queue has a varying rate from $\lambda_2 = \lambda_1 = 5$ to $\lambda_2 = \lambda_1 = 40$. We compare the mean waiting time in the system obtained from formula (18) and the simulation (see Tables 1 and 2).

Table 1. $\beta = 0$

Arrival rate	Formula (18)	Simulation	Relative error, %
5	0.013	0.0128	1.5
10	0.0146	0.0144	1.4
15	0.0167	0.0164	1.8
20	0.0194	0.0192	1
25	0.0233	0.0230	1.3
30	0.0292	0.0289	1
35	0.0389	0.0384	1.3
40	0.0583	0.0576	1.2

Consider the case where the Poisson input to the first queue has a varying rate from $\lambda_1 = 5$ to $\lambda_1 = 40$ and the Poisson input to the second queue with a constant rate $\lambda_2 = 40$. The results obtained are presented in Table 3.

Now consider the system of $N = 3$ queues with exponential distribution of the service time with $b = 0.01$ (the parameter inverse), $\beta = 0.002$, $\lambda_1 = 30$, $\lambda_2 = 20$ and the variable parameter λ_3 (from 2 to 16). The results obtained for the mean waiting times W_i in queues $i = \overline{1, 3}$ calculated by (18) and the simulation results are presented in Table 4.

Table 2. $\beta = 0.002$

Arrival rate	Formula (18)	Simulation	Relative error, %
5	0.0137	0.0140	2.1
10	0.0154	0.0156	1.3
15	0.0177	0.0176	0.6
20	0.0206	0.0204	1
25	0.0247	0.0242	1.2
30	0.0309	0.0298	3.6
35	0.0411	0.0395	3.9
40	0.0612	0.0585	4.4

Table 3. $\beta = 0.002$

Arrival rate to the first queue	Formula (18)		Simulation results		Relative error, %	
	Q_1	Q_2	Q_1	Q_2	Q_1	Q_2
5	0.0216	0.0228	0.0221	0.0221	2.3	3.1
10	0.0239	0.0251	0.0239	0.0242	0	3.6
15	0.0268	0.0279	0.0262	0.0268	2.2	3.9
20	0.0303	0.0313	0.0293	0.0301	3.3	3.8
25	0.0348	0.0356	0.0333	0.0344	4.3	3.4
30	0.0408	0.0413	0.039	0.04	4.4	3.2
35	0.0490	0.0493	0.0467	0.0471	4.7	4.5
40	0.0612	0.0612	0.0585	0.0585	4.4	4.4

Table 4. The mean waiting time

λ_3	Formula (18)			Simulation			Relative error, %		
	Q_1	Q_2	Q_3	Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
2	0.0239	0.0263	0.0318	0.0236	0.0260	0.0305	1.3	1.1	4.1
4	0.0247	0.0273	0.0321	0.0242	0.0268	0.0314	2	1.8	2.2
6	0.0256	0.0284	0.0328	0.0254	0.0281	0.0323	0.8	1.1	1.5
8	0.0267	0.0296	0.0337	0.0263	0.0291	0.0333	1.5	1.7	1.2
10	0.0279	0.0311	0.0347	0.0274	0.0305	0.0342	1.8	1.9	1.4
12	0.0292	0.0327	0.0358	0.0288	0.0321	0.0355	1.4	1.8	0.8
14	0.0308	0.0345	0.0371	0.0303	0.0338	0.0366	1.6	2	1.4
16	0.0325	0.0366	0.0386	0.0321	0.0358	0.0379	1.2	2.2	1.9

5 Conclusion

In the paper we presented the method of probability generating functions to analyse the polling system with adaptive cyclic polling and gated service. The method allows to obtaining the system of linear algebraic equations for the first moments of the number of customers in the queues at the polling moments which can be used to calculate the mean waiting times. For the sake of brevity, we presented the results for the gated service only but the case of exhaustive service can be analysed in a similar way.

References

1. Takagi, H.: Analysis of Polling Systems. MIT Press, Cambridge (1986)
2. Levy, H., Sidi, M., Boxma, O.J.: Dominance relations in polling systems. *Queueing Syst.* **6**, 155–172 (1990)
3. Borst, S.C.: Polling Systems. Stichting Mathematisch Centrum, Amsterdam (1996)
4. Vishnevskii, V.M., Semenova, O.V.: Mathematical methods to study the polling systems. *Autom. Remote Control* **67**, 173–220 (2006)
5. Vishnevsky, V.M., Mishkoy, G.K., Semenova, O.V.: New models and methods to study polling systems. In: Proceedings of the International Conference on Distributed Computer and Communication Networks, DCCN 2009. Theory and Applications, Moscow, pp. 79–85 (2009)
6. Vishnevsky, V., Semenova, O.: Polling Systems: Theory and Applications for Broadband Wireless Networks, 317 p. LAMBERT Academic Publishing, London (2012)
7. Vishnevsky, V.M., Dudin, A.N., Klimenok, V.I., Semenova, O.V., Shpilev, S.: Approximate method to study $M/G/1$ -type polling system with adaptive polling mechanism. *Qual. Technol. Quant. Manag.* **2**, 211–228 (2012)
8. Yechiali, U.: Analysis and control of polling systems. In: Donatiello, L., Nelson, R. (eds.) Performance/SIGMETRICS 1993. LNCS, vol. 729, pp. 630–650. Springer, Heidelberg (1993). <https://doi.org/10.1007/BFb0013871>
9. Altman, E., Blanc, H., Khamisy, A., Yechiali, Y.: Gated-type polling systems with walking and switch-in times. *Commun. Stat. Stoch. Models* **10**, 741–763 (1994)
10. Eliazar, I., Yechiali, U.: Polling under randomly-timed gated regime. *Stoch. Models* **14**(1–2), 79–93 (1998)
11. Leung, K.K.: Cyclic-service systems with non-preemptive, time-limited service. *IEEE Trans. Commun.* **42**, 2521–2524 (1994)



Retrial Queue with Search of Interrupted Customers from the Finite Orbit

Dhanya Babu, Achyutha Krishnamoorthy, and Varghese C. Joshua^(✉)

Department of Mathematics, CMS College, Kottayam 686001, Kerala, India
{dhanyababu,krishnamoorthy,vcjoshua}@cmscollege.ac.in
<http://www.cmscollege.ac.in>

Abstract. In this paper we consider a single server retrial queue with two orbits of which the first orbit is occupied by primary customers who on arrival find the server busy or interrupted. The other orbit has finite capacity and consists of customers whose service get interrupted due to server breakdown. Interrupted customers are picked up with probability p by the server at the epoch at which he/she becomes free either by the successful completion of a service or by completion of repair. Also there arises a competition between primary customers, retrial customers from the first and the second orbit to access the server. Failed retrials result in the customers returning to the respective orbits. The primary customers arrive according to a Markovian arrival process (MAP), the interruption occur according to a Poisson process. Fixing of interruption takes a random duration having phase type distribution. The service time follows phase type distribution. Stability condition of the system is established. Steady-state system size distribution is obtained. Performance characteristics of the system are evaluated.

Keywords: Retrial queue · Server interruption · Repair
Orbital search

1 Introduction

Retrial queues play a vital role in the study of queueing models. This class of queues is characterized by the following feature: a customer on arrival, when all servers are busy leaves the service area but after some random time repeat his demand. This field have many applications in computer and communication networking, aircraft landing and take-off, and in several other areas. The first mathematical result about retrial queues were published in 1950s and applications in teletraffic theory were presented in the monograph of L. Kosten. Both single-server, multi-server retrial queueing models, their methods of analysis and results are described in [7,15]. The bibliographical information about retrial queues are given in [3–5]. Steady state solution of a single-server queue with linear repeated requests are described in [9]. Literature about retrial queues are referred in [14,30]. In the retrial queueing system customers arriving to a busy

system join a group of blocked customers called orbit and try to capture a free server after a random amount of time. Neuts and Ramalhoto in [23] introduces the idea of search for customers in classical queue. In retrial queues each service is preceded and followed by an idle period until a primary or secondary customer get absorbed into the service. However a concept of orbital search is introduced in [6] to minimize the idle time of the server. The mechanism is described as: On every service completion epoch the sever picks up a customer from the orbit with a probability p_j when there are j customers in the orbit and consequently the server remains idle with probability $1 - p_j$. In this model there are two objectives one is to introduce retrial queue with orbital search as an appropriate stochastic model for some practical repair models and the other is to provide a link between M/G/1 retrial queue and the standard M/G/1 queue. This is possible by choosing the recovery factor $p_j = r$ as 0 and 1. Chakravarthy et al. [10] considered a multi-server retrial queue with orbital search in which primary customers arrive according to MAP. A retrial queueing system with orbital search and Batch Markovian Arrival Process (BMAP) is analyzed in [13].

This is a retrial queueing model with server interruption. White introduced the queueing system with service interruption as a preemptive priority system in [12]. The term service interruption means that the service of a customer is interrupted either by a server vacation, breakdown of the server or by service of any priority customer. Aissani in [1] considered an unreliable M/G/1 retrial queue and redundancy problem. Aissani discussed a retrial queueing system with breakdown in [2]. Krishnamoorthy et al. in [19] proposed service interruption as disaster to the unit undergoing service. In [11] Choi discussed queueing system with feedback and Artalejo et al. considered retrial queueing system with two types of interruption in [8]. Li and Zhang in [22] considered an M/G/1 retrial G-queue with general retrial times, in which the server is still working in a low service rate even if the system is in breakdown and repair process starts immediately. Gaver in [17] studied a queueing problem with interruption in which on completion of interruption either the service can be repeated or resumed. But Krishnamoorthy et al. in [20] discussed queues with interruptions and repeat or resumption of service by setting a threshold clock, with Markov arrival process, phase type distributed interruption and phase type service time distribution. Many studies have been done about queues with interruption in random environment. Queues with service interruptions in an alternating random environment is discussed in [27]. Queueing systems with disruptive and non-disruptive interruption have been taken into consideration in [16]. Krishnamoorthy discussed queues with interruption in which service phases are divided into protected and unprotected groups where no more interruption affects the protected phases. But queues with customer induced interruption was taken for study by Krishnamoorthy et al. in [28]. Retrial queue with server breakdowns and repairs are analyzed using reliability theory in [29]. Neuts in [24] developed the theory of PH - distributions and related point processes. In stochastic modelling, PH-distributions lend themselves naturally to algorithmic implementation and have nice closure properties with a related matrix formalism that makes them attractive for practical use.

In this model we consider a single server retrial queueing system with two orbits in which the service is interrupted by server breakdown. A primary customer finding a busy server enter into an orbit of infinite capacity and attempt to retry for service. A customer whose service interrupted enter into another orbit of finite capacity from where they can make retrials. At every service completion epoch or after an exponential duration of repairing time the server pick an interrupted customer with a search probability p . Steady-state probabilities are computed using Matrix Geometric Methods in [25] and [26] by Neuts. The rate matrix is computed using Ramaswami's Logarithmic Reduction Algorithm by Latouche and Ramaswami in [21].

In Sect. 2 we described the model. Stability condition is derived in Sect. 3.1 and computation of steady-state vector have been done in Sect. 3.2. Some performance measures are evaluated in Sect. 4.

2 Model Description

We consider a single server retrial queue with two orbits say orbit I and orbit II. Orbit I is of infinite capacity and orbit II of finite capacity say N . We assume that the server undergoes interruption only when it is busy. Primary customers enter into orbit I either by finding a busy or an interrupted server. A customer whose service get interrupted can enter into orbit II whenever orbit size is less than or equal to N and when capacity is full, further interrupted customers are considered as lost forever. Whenever the server is interrupted, repairing process starts immediately. Retrials of customers are also possible from both the orbits. Interrupted customers are picked up with probability p by the server at the epoch at which he is free either by the successful completion of a service or repair. Primary customers arrive according to a Markovian arrival process (MAP) with representation (D_0, D_1) of order n . An arriving customer enter into service immediately when the server is free. The service time is assumed to follow phase distribution with representation $PH(\alpha, T)$ of order l . The vector T^0 is given by $T^0 = -Te$. Interruption occurs according to a Poisson process with parameter γ . Consequently, repairing process starts immediately with a phase type distributed amount of time with representation $PH(\beta, S)$ of order m . The vector S^0 is given by $S^0 = -Se$. At every service completion epoch or after repairing the server goes for search for interrupted customers with probability p and remains idle with probability $1 - p$. Search time is assumed to be negligible. Retrials of customers from orbit I and II are assumed to be exponential with rates μ_1 and μ_2 respectively. We assume that the server can get into interruption any number of times and service of an interrupted customer repeat identically. We intend to optimize p to minimize the idle time of the server, also to find an optimum capacity of finite orbit to accommodate the interrupted customers.

The MAP, a special class of tractable Markov renewal process, is a rich class of point processes that includes many well-known processes such as Poisson, PH-renewal processes, and Markov-Modulated Poisson Process (MMPP). One of the most significant features of the MAP is the underlying Markov structure

and fits ideally in the context of matrix-analytic solutions to stochastic models. Matrix analytic methods were first introduced and studied by Neuts. Poisson processes are the simplest and most tractable one used extensively in stochastic modelling. The idea of the MAP is to significantly generalize the Poisson processes and still keep the tractability for modelling purposes. In many practical applications, mainly in communications engineering, production and manufacturing engineering, the arrivals do not usually form a renewal process. So, MAP is a convenient tool to model both renewal and non-renewal arrivals. The customers arrive to the system with a stochastic process $\{\nu_t, t \geq 0\}$ with a state space $\{0, 1, 2, \dots, W\}$. The sojourn time of the chain in the state i is exponentially distributed with the positive finite parameter λ_i . When the sojourn time in the state i expires, with probability $p_0(i, j)$, the process ν_t jumps to the state j without generation of customers where $i, j = \{0, 1, 2, \dots, W\}; i \neq j$ and with probability $p_1(i, j)$ the process ν_t jumps to the state j with generation of customers where $i, j = \{0, 1, 2, \dots, W\}$.

The MAP process is completely characterized by the matrices D_0 and D_1 defined by

$$\begin{aligned} (D_0)_{i,i} &= -\lambda_i, i = 0, 1, 2, \dots, W \\ (D_0)_{i,j} &= \lambda_i p_0(i, j); i, j = 0, 1, 2, \dots, W, i \neq j \\ (D_1)_{i,j} &= \lambda_i p_1(i, j); i, j = 0, 1, 2, \dots, W \end{aligned}$$

The point process described by the MAP is a special class of Semi-Markov processes with transition probability matrix given by

$$\int_0^x e^{(D_0 t)} dt D_1$$

By assuming D_0 to be a non-singular matrix, the inter arrival times will be finite with probability one and the arrival process does not terminate. Hence, we see that D_0 is a stable matrix. The matrix $D(1) = D_0 + D_1$ represents the generator of the process $\{\nu_t, t \geq 0\}$. The average arrival rate λ is given by

$$\lambda = \theta D_1 \mathbf{e}$$

where θ is the invariant vector of the stationary distribution of the Markov chain $\{\nu_t, t \geq 0\}$. The vector θ is the unique solution to the system

$$\theta D(1) \mathbf{e} = 0, \theta \mathbf{e} = 1.$$

Here \mathbf{e} is a column vector of appropriate size consisting of 1's and $\mathbf{0}$ is a row vector of appropriate size consisting of zeros. The squared integral coefficient of variation of intervals between successive arrivals is given by $C_{var} = 2\lambda\theta(-D_0)^{-1}\mathbf{e} - 1$.

Notations

Let

- \mathbf{e} be a column vector all one's of appropriate order
- \mathbf{O} be a zero matrix of appropriate order
- I_r be an identity matrix of dimension r
- \otimes Kronecker product of two matrices
- If A is a matrix of order $m \times n$ and if B is a matrix of order $p \times q$, then $A \otimes B$ will denote a matrix of order $mp \times nq$ whose $(i, j)^{th}$ block matrix is given by $a_{ij}B$
- $N_1(t)$ be the number of customers in the orbit I at time t
- $C(t)$ be the status of the server

$$C(t) = \begin{cases} 0, & \text{if the server is idle} \\ 1, & \text{if the server is in service} \\ 2, & \text{if the server is under repair} \end{cases}$$

- $N_2(t)$ be the number of customers in the orbit II at time t
- $J_1(t)$ be the phase of the service process at time t
- $J_2(t)$ be the phase of the repair process at time t
- $J_3(t)$ be the phase of the arrival process at time t

The above model can be represented by the Markov process

$$X^* = \{X(t)/t \geq 0\} = \{(N_1(t), C(t), N_2(t), J_1(t), J_2(t), J_3(t)); t \geq 0\}$$

The state space is

$$\Omega = l^* \cup l(i) \text{ where } l^* = \{(i, 0, k, r) : i \geq 0, 1 \leq k \leq N, r = 1, 2, \dots, n\}$$

and

$$l(i) = \{(i, 1, k, p, r) \cup (i, 2, k, q, r); i \geq 0, 1 \leq k \leq N, p = 1, 2, \dots, l, q = 1, 2, \dots, m, r = 1, 2, \dots, n\}$$

This model is a level independent quasi birth and death process (LIQBD). Quasi birth death process can be conveniently and efficiently solved by the classical matrix analytic method.

3 Steady-State Analysis

Enumerating the states of a continuous time Markov chain in lexicographic order, the infinitesimal generator of the Markov chain is of the form:

$$Q = \begin{pmatrix} B & A_0 & & & & \\ & A_2 & A_1 & A_0 & & \\ & & A_2 & A_1 & A_0 & \\ & & & A_2 & A_1 & A_0 \\ & & & & \dots & \dots \\ & & & & & \dots \end{pmatrix}$$

where B, A_0, A_1, A_2 are square matrices of order K where $K = n[(N + 1)(l + 1) + Nm]$.

$$B = \begin{pmatrix} B^{(0,0)} & B^{(0,1)} & O \\ B^{(1,0)} & B^{(1,1)} & B^{(1,2)} \\ B^{(2,0)} & B^{(2,1)} & B^{(2,2)} \end{pmatrix}$$

$B^{(k,k^*)}$ - represents the matrix corresponding to the transitions in level 0 when server status changes from k to k^* , where $k, k^* \in \{0, 1, 2\}$.

$$\begin{aligned} B^{(0,0)} &= \begin{pmatrix} D_0 & O & O \\ O & D_0 - \mu_2 I_n & O \\ O & O & D_0 - \mu_2 I_n \end{pmatrix} \\ B^{(0,1)} &= \begin{pmatrix} \alpha \otimes D_1 & O & O \\ \mu_2 \alpha \otimes I_n & \alpha \otimes D_1 & O \\ O & \mu_2 \alpha \otimes I_n & \alpha \otimes D_1 \end{pmatrix} \\ B^{(1,0)} &= \begin{pmatrix} T^0 \otimes I_n & O & O \\ O & (1-p)T^0 \otimes I_n & O \\ O & O & (1-p)T^0 \otimes I_n \end{pmatrix} \\ B^{(1,1)} &= \begin{pmatrix} T \oplus (D_0 - \gamma I_n) & O & O \\ pT^0 \otimes \alpha & T \oplus (D_0 - \gamma I_n) & O \\ O & pT^0 \otimes \alpha & T \oplus (D_0 - \gamma I_n) \end{pmatrix} \\ B^{(1,2)} &= \begin{pmatrix} (\gamma\beta \otimes I_n) \otimes \mathbf{e} & O \\ O & (\gamma\beta \otimes I_n) \otimes \mathbf{e}_l \\ O & (\gamma\beta \otimes I_n) \otimes \mathbf{e}_l \end{pmatrix} \\ B^{(2,0)} &= \begin{pmatrix} O & (1-p)S^0 \otimes I_n & O \\ O & O & (1-p)S^0 \otimes I_n \end{pmatrix} \\ B^{(2,1)} &= \begin{pmatrix} pS^0 \alpha \otimes I_n & O & O \\ O & pS^0 \alpha \otimes I_n & O \end{pmatrix} \end{aligned}$$

$$B^{(2,2)} = \begin{pmatrix} S \oplus D_0 & O \\ O & S \oplus D_0 \end{pmatrix}$$

$$A_0 = \begin{pmatrix} O & O & O \\ O & I_{N+1} \otimes (I_n \otimes D_1) & O \\ O & O & I_N \otimes (I_n \otimes D_1) \end{pmatrix}$$

A_0 - represents the transition matrix corresponding to the arrival of primary customer to the first orbit.

$$A_1 = \begin{pmatrix} A_1^{(0,0)} & A_1^{(0,1)} & O \\ A_1^{(1,0)} & A_1^{(1,1)} & A_1^{(1,2)} \\ A_1^{(2,0)} & A_1^{(2,1)} & A_1^{(2,2)} \end{pmatrix}$$

$A_1^{(k,k^*)}$ - represents the matrix corresponding to the transitions in level i when server status changes from k to k^* where $k, k^* \in \{0, 1, 2\}$.

$$A_1^{(0,0)} = \begin{pmatrix} D_0 - \mu_1 I_n & O & O \\ O & D_0 - (\mu_1 + \mu_2) I_n & O \\ O & O & D_0 - (\mu_1 + \mu_2) I_n \end{pmatrix}$$

and

$$A_1^{(0,1)} = B^{(0,1)}, A_1^{(1,0)} = B^{(1,0)}$$

$$A_1^{(1,1)} = B^{(1,1)}, A_1^{(1,2)} = B^{(1,2)}$$

$$A_1^{(2,0)} = B^{(2,0)}, A_1^{(2,1)} = B^{(2,1)}, A_1^{(2,2)} = B^{(2,2)}$$

$$A_2 = \begin{pmatrix} O & I_{N+1} \otimes (\mu_1 \alpha \otimes I_n) & O \\ O & O & O \\ O & O & O \end{pmatrix}$$

A_2 - represents the transition matrix corresponding to the successful retrial of customer in the first orbit.

3.1 Stability Condition

Let π denote the steady- state probability vector of the generator matrix $A = A_0 + A_1 + A_2$ where

$$A = \begin{pmatrix} A^{(0,0)} & A^{(0,1)} & O \\ A^{(1,0)} & A^{(1,1)} & A^{(1,2)} \\ A^{(2,0)} & A^{(2,1)} & A^{(2,2)} \end{pmatrix}$$

where

$$A^{(0,0)} = A_1^{(0,0)}, A^{(0,1)} = I_{N+1} \otimes (\mu_1 \alpha \otimes I_n) + A_1^{(0,1)}$$

$$A^{(1,0)} = A_1^{(1,0)}, A^{(1,1)} = I_{N+1} \otimes (I_n \otimes D_1) + A_1^{(1,1)}$$

$$A^{(1,2)} = A_1^{(1,2)}, A^{(2,0)} = A_1^{(2,0)}$$

$$A^{(2,1)} = A_1^{(2,1)}, A^{(2,2)} = A_1^{(2,2)} + I_N \otimes (I_n \otimes D_1)$$

We see that A is an irreducible infinitesimal generator matrix and so there exists the stationary vector π of A such that

$$\pi A = 0$$

and

$$\pi \mathbf{e} = 1$$

where the vector π partitioned as

$$\pi = (\pi_0, \pi_1, \pi_2)$$

is computed by solving the equations

$$\begin{aligned} \pi_0 A^{(0,1)} + \pi_1 A^{(1,1)} + \pi_2 A^{(2,1)} &= 0 \\ \pi_0 A^{(0,2)} + \pi_1 A^{(1,2)} + \pi_2 A^{(2,2)} &= 0 \\ \pi_1 A^{(1,3)} + \pi_2 A^{(2,3)} &= 0 \end{aligned}$$

subject to

$$\pi_0 + \pi_1 = 1$$

Solving we get

$$\pi_0 \left[\sum_{i=0}^{N-1} \prod_{j=0}^{N-1-i} H_{N-1-j} + I \right] \mathbf{e} = 1 \tag{1}$$

where

$$\begin{aligned} H_1 &= -A^{(0,1)}[A^{(1,1)} + H_0 A^{(2,1)}]^{-1} \text{ and} \\ H_0 &= -A^{(1,2)}[A^{(2,2)}]^{-1} \end{aligned}$$

The system X^* is stable if and only if

$$\pi A_0 \mathbf{e} < \pi A_2 \mathbf{e}$$

i.e.

$$\pi_1 [I_{N+1} \otimes (I_n \otimes D_1)] + \pi_2 [I_N \otimes (I_n \otimes D_1)] < \pi_0 [I_{N+1} \otimes (I_n \otimes \mu_1 \alpha)]$$

3.2 Computation of the Steady-State Vector

Let \mathbf{x} be the steady-state probability vector of Q . Partition this vector as: $\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \dots)$, where $\mathbf{x}_i = (\mathbf{x}(i, 0), \mathbf{x}(i, 1), \mathbf{x}(i, 2))$ for $i \geq 0$.

Under the stability condition the steady-state probability vector is obtained as

$$\begin{aligned} \mathbf{x}_i &= \mathbf{x}_{i-1} R, i \geq 1 \\ \mathbf{x}_i &= \mathbf{x}_0 R^i, i \geq 1 \end{aligned}$$

where R is the minimal non negative solution to the matrix quadratic equation

$$R^2 A_2 + R A_1 + A_0 = 0 \tag{2}$$

R can be obtained by successive substitution procedure $R_0 = 0$ and

$$R_{k+1} = -V - R_k^2 W$$

where $V = A_2 A_1^{-1}$, $W = A_0 A_1^{-1}$ by *Logarithmic Reduction Algorithm* developed by Latouche and Ramaswamy. The vectors \mathbf{x}_0 and \mathbf{x}_1 are obtained by solving

$$\begin{aligned} \mathbf{x}_0 B + \mathbf{x}_1 A_2 &= 0 \\ \mathbf{x}_0 A_0 + \mathbf{x}_1 A_1 + \mathbf{x}_2 A_2 &= 0 \end{aligned}$$

subject to the normalizing condition

$$\mathbf{x}_0 (I - R)^{-1} \mathbf{e} = 1.$$

4 Performance Measures

1. Probability mass function of the number of customers in orbit I

$$\begin{aligned} \text{Pr}[i \text{ customers in orbit I}] \\ = \mathbf{x}(i) \mathbf{e} \end{aligned}$$

2. Probability mass function of the number of customers in orbit II

$$\begin{aligned} \text{Pr}[i \text{ customers in orbit II}] \\ = \mathbf{x}_{ij}(k) \mathbf{e} \end{aligned}$$

3. Expected Number of customers in orbit I

$$E[N_1] = \sum_{i=0}^{\infty} i \mathbf{x}(i) \mathbf{e}$$

4. Expected Number of customers in orbit II

$$E[N_2] = \sum_{k=0}^N k \sum_{i=0}^{\infty} \sum_{j=0}^2 \mathbf{x}_{ij}(k) \mathbf{e}$$

5. Expected Number of customers in the system when the server is idle

$$E[N_0] = \sum_{i=0}^{\infty} i \sum_{k=0}^N k \mathbf{x}_{i0}(k) \mathbf{e}$$

6. Expected Number of customers in the system when the server is in service

$$E[N_1] = \sum_{i=0}^{\infty} i \sum_{k=0}^N k \mathbf{x}_{i1}(k) \mathbf{e} + 1$$

7. Expected Number of customers in the system when the server is under repair

$$E[N_2] = \sum_{i=0}^{\infty} i \sum_{k=0}^N k \mathbf{x}_{i2}(k) \mathbf{e}$$

8. Expected Number of customers in the system

$$E[N] = E[N_0] + E[N_1] + E[N_2]$$

9. Probability that the server is idle

$$P_0 = \sum_{i=0}^{\infty} \mathbf{x}_i(0) \mathbf{e}$$

10. Probability that the server is in service

$$P_1 = \sum_{i=0}^{\infty} \mathbf{x}_i(1) \mathbf{e}$$

11. Probability that the server is under repair

$$P_2 = \sum_{i=0}^{\infty} \mathbf{x}_i(2) \mathbf{e}$$

12. The probability that an interrupted customer is blocked from entering into orbit II

$$P_b = \sum_{i=0}^{\infty} \sum_{j=0}^2 \mathbf{x}_{ij}(N) \mathbf{e}$$

4.1 Conclusion

We considered a single server retrial queue enhanced with the search mechanism for interrupted customers. Interrupted customers once entered into service repeat their service identically without getting any protection for further service. This model could be extended by providing separate service or protection for those interrupted customers who reentered into the service by any means i.e. by successful retrial or by search. Also we can evaluate the expected number of interruptions occurred to the server and expected number of searches, during a particular period of time. And so we can address a control problem to optimize the search probability and the capacity of the finite orbit.

Acknowledgments. The work of the third author is supported by the Maulana Azad National fellowship [F1 – 17.1/2015 – 16/MANF – 2015 – 17 – KER – 65493] of University Grants commission, India.

References

1. Aissani, A.: Unreliable queueing with repeated orders. *Micro. and Reli.* **33** (1993)
2. Aissani, A., Artalejo, J.R.: On the single server retrial queue subject to breakdowns. *Que. Syst.* **30** (1998)
3. Artalejo, J.R., Gomez-Corral, A.: *Retrial Queueing Systems: A Computational Approach*. Springer, Berlin (2008)
4. Artalejo, J.R.: Accessible bibliography of research on retrial queues. *Math. Comput. Model.* **30**, 1–6 (1999)
5. Artalejo, J.R.: A classified bibliography of research on retrial queues. *Progress 1990–1999 Top* **7**, 187–211 (1999)
6. Artalejo, J.R., Joshua, V.C, Krishnamoorthy, A.: An M/G/1 retrial queue with orbital search by server. In: *Adv. in Stoch. Model*, pp. 41–54. Notable Publications, New Jersey (2002)
7. Artalejo, J.R., Phung, T.D.: Single server retrial queues with two way communication. *Appl. Mathe. Model.* **37**, 1811–1822 (2013)
8. Artalejo, J.R.: New results in retrial queueing systems with breakdown of the servers. *Stati. Neerl.* **48**(I), 23–36 (1994)
9. Artalejo, J.R., Gomez-Corral, A.: Steady state solution of a single-server queue with linear repeated requests. *J. Appl. Prob.* **34**, 223–233 (1997)
10. Chakravathy, S.R., Krishnamoorthy, A., Joshua, V.C.: Analysis of a multi-server retrial queue with search of customers from the orbit. *Perf. Eval.* **63**(8), 776–798 (2006)
11. Choi, B.D., Kulkarni, B.D.: Feedback retrial queueing systems. In: *Queueing and Related Models*. Ox. Scie. Publi. (1992)
12. White, H.C., Christie, L.S.: Queueing with preemptive priorities or with breakdown. *Oper. Res.* **6**, 79–95 (1958)
13. Dudin, A.N., Krishnamoorthy, A., Joshua, V.C., Tsarenkov, G.: Analysis of BMAP/G/1 retrial system with search of customers from the orbit. *Eur. J. Oper. Res.* **157**, 169–179 (2004)
14. Falin, G.I.: A survey of Retrial queues. *Q. Syst.* **7**(2), 127–167 (1990)
15. Falin, G.I., Templeton, J.G.C.: *Retrial Queues*. Chapman and Hall, London (1997)
16. Fiems, D., Bruneel, H.: Queueing systems with different types of server interruptions. *Eur. J. Oper. Res.* **188**, 838–845 (2008)
17. Gaver, D.: A waiting line with interrupted service including priority. *J. Roy. Stat. Soc.* **B24**, 90 (1962)
18. Krishnamoorthy, A., Deepak, T.G., Joshua, V.C.: An M/G/1 retrial queue with non persistent customers and orbital search. *Stoch. Anal. Appl.* **23**, 975–997 (2005)
19. Krishnamoorthy, A., Ushakumari, P.V.: Reliability of a k-out-of -n system with repair and retrial of failed units. *Top* **7**, 293–304 (1999)
20. Krishnamoorthy, A., Pramod, P.K., Deepak, T.G.: On a queue with interruptions and repeat or resumption of service. *No. Lin. Anal.* (2009)
21. Latouche, G., Ramaswami, V.: *Introduction to Matrix analytic Methods in Stochastic Modelling*. Amer. Stat. Asso, Siam (1999)
22. Li, T., Zhang, L.: An M/G/1 Retrial G-queue with general retrial times and working breakdowns. *Math. Comput. Appl.* **22**(1), 15 (2017)
23. Neuts, M.F., Ramalhoto.: A service model in which the server is required to search for customers. *J. Appl. Prob.* **21**, 157–166 (1984)
24. Neuts, M.F.: Probability distributions of phase type. In: *Liber Amicorum Prof. Emeritus H. Florin*, Department of Mathematics, University of Louvain, pp. 173–206 (1975)

25. Neuts, M.F.: *Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach*. The Johns Hopkins University Press, Baltimore and London (1981)
26. Neuts, M.F.: Markov chains with applications in queueing theory which have a matrix-geometric invariant probability vector. *Adv. Appl. prob.* **10**, 185–212 (1978)
27. Sengupta, B.: A queue with service interruptions in an alternating random environment. *Oper. Res.* **38**, 308–318 (1989)
28. Jacob, V., Chakravarthy, S.R., Krishnamoorthy, A.: On a customer induced interruption in a service systems. *J. Stoch. Anal. Appl.* **30**, 1–13 (2012)
29. Wang, J., Cao, J., Li, Q.: Reliability analysis of the retrial queue with server breakdowns and repairs. *Que. Syst.* **38**, 363–380 (2001)
30. Yang, T., Templeton, J.G.: A survey on retrial queues. *Que. Sys.* **2**, 201–233 (1987)



Traffic Optimization and Multi-sided Pricing in Congested Networks

Haroun H. Salih, Dina A. Urusova, and Sergey A. Vasilyev^(✉)

Department of Applied Probability and Informatics, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow 117198, Russian Federation
harounhassan198@yahoo.fr, ns_urusovada@str.mos.ru,
vasilyev_sa@rudn.university

Abstract. The paper we consider traffic optimization problem for a model with multi-sided dynamic pricing in the telecommunications market with incomplete competition and taking into account congested networks. The model consists in the use of mathematical modeling methods, game theory and queueing theory. It is assumed that telecommunication companies agree on the rules of incoming and outgoing traffic charging in pairs, and this charging is built as a function of the tariffs that companies offer their subscribers for service. Companies are limited the agreement on mutual rules of reciprocal proportional charging for access traffic at first, which subsequently determine the tariffs for the network users. The reciprocity of the rules means that companies are subject to the same rules for the entire time interval during which the agreement is in force. Taking into account imperfect competition in the telecommunications market and using traffic and profit optimization method for each company the equilibrium tariffs and the volume of services are found with subject to congestion in multi-service networks. Numerical calculation is performed to illustrate the results.

Keywords: Queueing theory · Game theory · Optimization methods
Probability theory · Industrial market theory · Economic and mathematical modeling

1 Introduction

Methods of mathematical modeling in the economy of telecommunications are developed actively. Jean Tirole considers the impact of telecommunication technologies on competition in services and goods markets [10–14].

Doganoglu and Tauman [6] presented a model of two competing local telecommunications networks which are mandated to interconnect. After negotiating the access charges, the companies engage in price competition. Given the prices, each consumer selects a network and determines the consumption of phone calls. Using a discrete/continuous consumer choice model, it was shown

that a pure strategy equilibrium exists quite generally and satisfies desirable properties.

Dessein [4] considered network competition in nonlinear pricing, assuming linear pricing, and had shown that telecommunications networks may use a high access charge as an instrument of collusion. He showed that this conclusion is difficult to maintain when operators compete in nonlinear pricing: as long as subscription demand is inelastic, profits can remain independent of the access charge, even when customers are heterogeneous and networks engage in second-degree price discrimination. When demand for subscriptions is elastic, networks may increase profits by agreeing on an access charge below marginal cost (relative to cost-based access pricing) and welfare is typically increased by setting the access charge above marginal cost.

Wouter Dessein argued that the telecommunications industry is a fragmented of a market, characterized by a tremendous amount of customer heterogeneity [5]. He showed how such customer heterogeneity dramatically affects nonlinear pricing strategies. If there are unbalanced calling patterns between different customer types, networks make larger profits on the least attractive customers, the nature of the calling pattern substantially affects how networks discriminate implicitly between different customer types. Different customer types often perceive the substitutability of competing networks differently.

Hahn [7] considered two-way access pricing in a telecommunications market where consumers are heterogeneous in their demand for calls and firms are allowed to use non-linear tariffs. He investigated how the presence of access charges affects the tariffs offered by firms in symmetric equilibrium and showed that under certain conditions each firm's profit is independent of the level of (reciprocal) access charge and, therefore, collusion using access charges is not sustainable. This result suggests that efficient call-allocations can be achieved under a minimal regulatory intervention, i.e. recommending firms set access charges equal to call-termination cost.

Chuna Se-Hak considered optimal access charges for the provision of telecommunication network, mobile commerce, and cloud services [17]. Using theoretical analysis, Chuna Se-Hak investigated, when a regulator can set rational access pricing, considering the characteristics of access demand. Chuna Se-Hak demonstrated that optimal access prices depend on whether the final products or services are strategic independence or strategic substitutes. The results have implications for policy makers setting optimal access charges that maximize social welfare.

Lee, Jeong, Seo [15] considered optimal pricing and capacity partitioning for tiered access service in virtual networks. They showed that many Internet service providers offer some forms of tiered access service to meet diverse demands of users and to improve the efficiency of network resources. They found the optimal pricing and capacity partitioning by addressing the revenue maximization problem in the tiered service.

Mark Armstrong examined the use of nonlinear pricing as a method of price discrimination, both with monopoly and oligopoly supply [1].

Ushchev, Zenou [19] developed a product-differentiated model where the product space is a network defined as a set of varieties (nodes) linked by their degrees of substitutability (edges). They located consumers into this network, so that the location of each consumer (node) corresponds to her “ideal” variety. It was shown that there exists a unique Bertrand–Nash equilibrium where prices are determined by both the firms’ sign-alternating Bonacich centralities and the average willingness to pay across consumers. They also investigated how local product differentiation and the spatial discount factor affect the equilibrium prices. They showed that these effects non-trivially depend on the network structure. It was found that, in a star-shaped network, the central firm does not always enjoy higher monopoly power than the peripheral firms.

Cramton, Doyle [2] described an open access market for capacity. They assumed that open access means that in real-time, network capacity cannot be withheld—capacity is priced dynamically by the marginal demand during congestion. They offered the open access market as a means for managing growing spectrum demand and as an alternative to naked spectrum sharing.

In a network formation framework, where payoffs reflect an agent’s ability to access information from direct and indirect contacts. Mohlmeier, Rusinowska, Tanimura [9] integrated negative externalities due to connectivity associated with two types of effects: competition for the access to information, and rivalrous use of information.

Samouylov, Sevastianov, Kulyabov, Gaidamaka, Gudkova and other researchers built various multiservice models of networks, queuing systems and considered their dynamics [3, 8, 16].

This article a mathematical model of pricing for telecommunications services with overloads in networks is built. It is generalized the model that it was built earlier [18, 20].

It is assumed that telecommunications companies agree in pairs on the rules of charging for access traffic to the network other, and it is considered as a function of the tariffs that companies offer their consumers (subscribers) for services. Thus, these companies have contracts at the first step by agreements on reciprocal proportional access charge rules (RPACR), which subsequently allow them to determine the subscription rates. The ambiguity of the rules means that companies are subject to one and the same rules for the entire time interval during, which the agreement is valid.

RPACR may be seen as analogous to the regulatory policy of the state of the telecommunications industry. If telecommunication services, provided by different companies, are close substitutes, the use of RPACR by companies it leads to competitive prices in industry. However, if it is assumed that competing companies follow the policy of services differentiation, then it is required intervention of the state to preclude the use by companies of monopoly power.

It is also assumed that the utility function of subscribers consists of deterministic and stochastic parts. The deterministic part allows to find a linear function of subscribers demand for telecommunications services, which has a constant price elasticity. It allows to avoid unlimited growth of consumption of telecommunication services by subscribers at aspiration the corresponding tariffs to zero

and ensures the existence of a saturation point, i.e., for example, there is time limits that the subscriber uses for using telecommunication services. The Weibull distribution is used for the stochastic component of the utility function, which is convenient for further analysis. It is possible to find equilibrium tariffs and equilibrium demand for telecommunication services. This equilibrium is equilibrium in pure strategies and always exists, and the subscription rates are calculated explicitly. Numerical calculation is performed to illustrate the results.

2 Model of Telecommunications Industry

2.1 Multiservice Network Model

Let consider a network NW ($NW = \bigcup_{i=1}^n NW_i$) consisting of n equivalent multiservice network (numbered in a certain order multiservice network $SR = \bigcup_{s=1}^m SR_s$) belonging to different telecommunication companies T_i ($i = \overline{1, n}$), and it is assumed that in between all the networking companies are switching nodes.

Let $t \in \{1, 2, \dots, T_{\max}\}$ be time intervals (for example, the time period equals a week, a month or a year) equal to the length of time periods during which companies T_i independently decide on pricing for their services, and t_{\max} is the maximum planning horizon.

Let's assume that the network NW consists of a set of nodes $J^t = \bigcup_{i=1}^{s_j} J_i^t$ and a set of channels $L^t = \bigcup_{i=1}^{s_l} L_i^t$, and $NW = J^t \cup L^t$.

In the time period t each network NW_i of the company T_i ($i = \overline{1, n}$) is represented the set of nodes J_{ij}^t ($j = 1, \dots, s_i^J$) and channel set L_{ij}^t ($j = 1, \dots, s_i^L$), numbered in a certain way, where $J_i^t = \bigcup_{j=1}^{s_i^J} J_{ij}^t$, $L_i^t = \bigcup_{k=1}^{s_i^L} L_{ik}^t$ and $NW_i = J_i^t \cup L_i^t$, and the total number of nodes is $S_{NW}^J(t) = \sum_{i=1}^n s_i^J$, and the total number of channels is $S_{NW}^L(t) = \sum_{i=1}^n s_i^L$ for network NW .

Let H_{ij}^t be a capacity (bits/sec) of j -node ($j = \overline{1, J_{s_i^J}}$), and S_{ik}^t a throughput (bits/sec) k -link ($k = \overline{1, L_{s_i^L}}$) T_i of network NW_i company T_i in the time period t .

Two-point connections can be established to transmit information flows between the network nodes of network NW . Each connection is characterized by a route, i.e. a set of network links NW , through which connections are established.

Let $s = \{1, \dots, m\}$ be a set of services that offer companies for potential consumers (subscribers) during the period $t \in \{1, 2, \dots, T_{\max}\}$. Let b ($b \in (1, 2, \dots, B^t)$) be a set of consumers, who want to use the telecommunication services in the market.

2.2 Individual Consumer Demand and Network Traffic

Let's assume that the individual consumer demand function for the service $s = \{1, \dots, m\}$ has the form:

$$D_{bs}^t(p_s^t) = \frac{r_{bs}^t - p_s^t}{2s_{bs}^t} = a_{bs}^t - b_{bs}^t p_s^t, \quad a_{bs}^t = \frac{r_{bs}^t}{2s_{bs}^t}, \quad b_{bs}^t = \frac{1}{2s_{bs}^t}, \quad (1)$$

$D_{bs}^t(p_s^t)$ is a linear function of the price p_s^t , and $r_{bs}^t > 0$ and $s_{bs}^t > 0$ is positive coefficients, which are determined from the market research services SR in the period t .

A consumer b generates the traffic loading or the load using the service s in the period t . Let Y_{bs}^t be an individual traffic volume of a consumer b , and let $Y_{bs}^t = \bar{\lambda}_{bs}^t h_{bs}^t$ be the average value of Y_{bs}^t , where the parameter $\bar{\lambda}_{bs}^t$ is the average intensity of the flow of requests and the parameter h_{bs}^t is the average duration of service in the period t .

We assume that the average load is generated by the consumer b when using the service s in the period t , linearly depends on the corresponding demand function for this service s

$$Y_{bs}^t = \bar{\lambda}_{bs}^t h_{bs}^t = \theta_s D_{bs}^t(p_s^t) = \theta_s (a_{bs}^t - b_{bs}^t p_s^t), \quad (2)$$

where θ_s is the proportionality factor for the s service. It links the consumer demand for telecommunication services and the amount of traffic generated by this consumer in the network.

The total network traffic volume that it creates by a consumer in the period t during using the service s , is the sum of consumers network traffic volumes

$$Y_s^t = \sum_{b=1}^{B_t} Y_{bs}^t = \sum_{b=1}^{B_t} \theta_s (a_{bs}^t - b_{bs}^t p_s^t) = \bar{A}_s^t - \bar{B}_s^t p_s^t, \quad (3)$$

$$\bar{A}_s^t = \sum_{b=1}^{B_t} \theta_s a_{bs}^t, \quad \bar{B}_s^t = \sum_{b=1}^{B_t} \theta_s b_{bs}^t,$$

where \bar{a}_s^t, \bar{b}_s^t are parameters of the function Y_s^t .

The total consumers demand for the service s during the time t is the sum of all demand functions for the service s of all:

$$D_{bs}^t(p_s^t) = \sum_{b=1}^{B^t} D_{bs}^t(p_s^t) = \sum_{b=1}^{B^t} (a_{bs}^t - b_{bs}^t p_s^t), \quad (4)$$

$$D_{bs}^t(p_s^t) = (a_s^t - b_s^t p_s^t), \quad a_s^t = \sum_{b=1}^{B^t} a_{bs}^t, \quad b_s^t = \sum_{b=1}^{B^t} b_{bs}^t,$$

where the parameters $a_s^t \geq 0$ and $b_s^t \geq 0$ are determined from market research of services in the period t .

We can get a link between the network traffic volume $Y_s^t(p_s^t)$ and the demand function $D_{bs}^t(p_s^t)$ of the service s during the period t :

$$Y_s^t(p_s^t) = Q_{bs}^t(p_s^t) \theta_s D_{bs}^t(p_s^t) = \theta_s (a_s^t - b_s^t p_s^t) = A_s^t - B_s^t p_s^t, \quad (5)$$

where $Y_s^t(p_s^t)$ is linear price functions and $A_s^t = \theta_s a_s^t$, $B_s^t = \theta_s b_s^t$ are coefficients.

We can get the network traffic volume that is associated with the consumer b ($b = \overline{1, B^t}$)

$$\begin{aligned} Y_b^t &= \sum_{s=1}^m Y_{bs}^t = \sum_{s=1}^m \theta_s (a_{bs}^t - b_{bs}^t p_s^t) \leq \bar{A}_b^t - \bar{B}_b^t \bar{p}^t, \\ \bar{A}_b^t &= \sum_{s=1}^m \theta_s a_{bs}^t, \quad \bar{B}_b^t = \sum_{s=1}^m b_{bs}^t, \quad \bar{p}^t = \sum_{s=1}^m p_s^t, \quad \bar{B}_b^t \bar{p}^t \leq \sum_{s=1}^m \theta_s b_{bs}^t p_s^t. \end{aligned} \quad (6)$$

where $\bar{A}_b^t \geq 0$, $\bar{B}_b^t \geq 0$ is parameters load functions Y_b^t associated with the consumer b , and a parameter \bar{p}^t is a tariff for services SR (service package) during the time period t .

A consumer's b ($b = \overline{1, B^t}$) demand for SR -services in the considered the time period t has the form:

$$\begin{aligned} Q_b^t(p_b^t) &= \sum_{s=1}^m D_{bs}^t(p_s^t) = \sum_{s=1}^m (a_{bs}^t - b_{bs}^t p_s^t) \leq (a_b^t - b_b^t \bar{p}^t), \\ a_b^t &= \sum_{s=1}^m a_{bs}^t, \quad b_b^t = \sum_{s=1}^m b_{bs}^t, \quad b_b^t \bar{p}^t \leq \sum_{s=1}^m b_{bs}^t p_s^t. \end{aligned} \quad (7)$$

Aggregating the network traffic volume $Y_s^t(p_s^t)$ from (5) for all services $s = \{1, \dots, m\}$, we can get the total network traffic volume $Y(t)$ for the period t in the form:

$$\begin{aligned} Y(t) &= \sum_{s=1}^m Y_s^t(p_s^t) = \sum_{s=1}^m (a_s^t - b_s^t p_s^t) = \sum_{s=1}^m \theta_s (a_s^t - b_s^t p_s^t) = \bar{A}^t - \bar{B}^t \bar{p}^t, \\ \bar{A}^t &= \sum_{s=1}^m \theta_s a_s^t, \quad \bar{B}^t \bar{p}^t \geq \sum_{s=1}^m \theta_s b_s^t p_s^t, \quad \bar{B}^t = \sum_{s=1}^m \theta_s b_s^t, \end{aligned} \quad (8)$$

where $\bar{A}^t \geq 0$ and $\bar{B}^t \geq 0$ are aggregated parameters of function $Y(t)$, and where function of aggregated demand for services SR (service package) has the form:

$$\begin{aligned} D(t) &= \sum_{s=1}^m (a_s^t - b_s^t p_s^t) = \bar{a}^t - \bar{b}^t \bar{p}^t, \\ \bar{a}^t &= \sum_{s=1}^m a_s^t, \quad \bar{b}^t \bar{p}^t \geq \sum_{s=1}^m b_s^t p_s^t, \quad \bar{b}^t = \sum_{s=1}^m b_s^t, \end{aligned} \quad (9)$$

where the parameters $\bar{a}^t \geq 0$ and $\bar{b}^t \geq 0$ are aggregated parameters of the demand function $D(t)$.

2.3 Reciprocal Proportional Access Charge Rules and Multi-sided Pricing

We can assume that for each company T_i ($i = \overline{1, n}$) has a function of consumer demand for services SR (service package) during the time period t .

Let D_{sii} ($i \in \{1, \dots, n\}$) be a demand function of services $SR = \bigcup_{s=1}^m SR_s$ provided by the company T_i using its NW_i network resource only, and let

D_{sij}^t ($i, j \in \{1, \dots, n\}, i \neq j$) be a demand function of services provided together with a network NW_i of a company T_i and a network NW_j of a company T_j ($i, j \in \{1, \dots, n\}, i \neq j$). Thus, there is a question of access of one company to resources of a network of other company.

We assume that the companies T_i and T_j to ($i, j \in \{1, \dots, n\}, i \neq j$) agree on the charges \hat{a}_{ij}^t and \hat{a}_{ji}^t , where \hat{a}_{ij}^t is a charge, which company T_i pays the company T_j of ($i, j \in \{1, \dots, n\}, i \neq j$) for the use of its network resources in connection with the service of $s \in \{1, \dots, m\}$ (traffic from the network NW_i to the network NW_j or outgoing traffic for the company T_i and incoming traffic the company T_j), and \hat{a}_{ji}^t is a corresponding charge at which the company T_j pays the company T_i ($i, j \in \{1, \dots, n\}, i \neq j$) for the using of network resources in connection with the provision of a similar service $s \in \{1, \dots, m\}$ (traffic from the network NW_j to the network NW_i or outgoing traffic for the company T_j and incoming traffic the company T_i) during the time period t .

Suppose that any two companies T_i and T_j ($i, j \in \{1, \dots, n\}, i \neq j$) charges \hat{a}_{ij}^t and \hat{a}_{ji}^t depend on tariffs \bar{p}_i^t and \bar{p}_j^t , and $\hat{a}_{ij}^t = a_i^t(\bar{p}_i^t, \bar{p}_j^t)$ for any ($i, j \in \{1, \dots, n\}, i \neq j$) and $s \in \{1, \dots, m\}$ at any time $t \in \{1, 2, \dots, T_{\max}\}$.

We assume that there is the proportional dependence between \hat{a}_{ij}^t and \bar{p}_i^t , then $\hat{a}_{ij}^t = a_i^t \bar{p}_i^t$, where the proportionality factor is $0 \leq a_i^t \leq 1$ for $i \in \{1, \dots, n\}$ and $s \in \{1, \dots, m\}$.

3 Multiservice Demand Function

Suppose that each consumer can use telecommunication multiservice network of companies T_i ($i \in \{1, \dots, n\}$) at any time period t . Let's assume that each consumer has individual tastes and preferences in relation to these services SR . We assume that the consumer b ($b \in \{1, \dots, B^t\}$), which is ready to choose one service from the set $s \in \{1, \dots, m\}$ of the company T_i ($i \in \{1, \dots, n\}$), has the following utility function:

$$\begin{aligned} u_{ibs}^t &= U_{ibs}^t e^{\eta_s \epsilon_{ibs}^t} = U_{bs}^t(Q_{bs}^t(p_{is}^t), p_{is}) e^{\eta_s \epsilon_{ibs}^t}, \\ U_{ibs}^t &= [r_{bs}^t - s_{bs}^t Q_{bs}^t(p_s^t)] Q_{bs}^t(p_s^t) - p_s^t Q_{bs}^t(p_s^t), \end{aligned} \tag{10}$$

where the random parameter ϵ_{ibs}^t characterizes individual tastes and preferences of the consumer. Let's consider that ϵ_{ibs}^t has a Weibull distribution. The value of η_s gives the characteristic measures of the dispersion of tastes and preferences of the consumers, that is η_s allows us to estimate the substitutability telecommunication services $s \in \{1, \dots, m\}$ that provide companies T_i and T_j ($i, j \in \{1, \dots, n\}, i \neq j$). The services $s \in \{1, \dots, m\}$ of companies become total substitutes with $\eta_s \rightarrow 0$, and it is total complementary with $\eta_s \rightarrow \infty$.

Let each consumer b ($b \in \{1, \dots, B^t\}$) makes a choice the company T_i and rejects the company T_j ($i, j \in \{1, \dots, t\}, i \neq j$) at the period t then there is inequality

$$U_{ibs}^t e^{\eta_s \epsilon_{ibs}^t} \geq U_{jbs}^t e^{\eta_s \epsilon_{jbs}^t}.$$

Thus, the probability P_{ibs}^t that the consumer b gives preference to the company T_i and reject the company T_j ($i, j \in \{1, \dots, n\}$, $i \neq j$) equals to

$$P_{ibs}^t = \text{P}\{U_{ibs}^t e^{\eta_s \epsilon_{ibs}} > U_{jbs}^t e^{\eta_s \epsilon_{jbs}}\}. \quad (11)$$

Since the values ϵ_{ibs} are independent and have a Weibull distribution we have that

$$P_{ibs}^t = \frac{1}{1 + \left(\frac{U_{ibs}^t}{U_{jbs}^t}\right)^{\frac{1}{\eta_s}}} = \frac{(r_{bs}^t p_{is}^t)_s^\tau}{(r_{bs}^t p_{is}^t)_s^\tau + (r_{bs}^t - p_{js}^t)_s^\tau}, \quad (12)$$

where $\tau_s = 2/\eta_s$. Similarly for the company T_j we have that same

$$P_{jbs}^t = \frac{1}{1 + \left(\frac{U_{jbs}^t}{U_{ibs}^t}\right)^{\frac{1}{\eta_s}}} = \frac{(r_{bs}^t p_{js}^t)_s^\tau}{(r_{bs}^t p_{js}^t)_s^\tau + (r_{bs}^t - p_{is}^t)_s^\tau}. \quad (13)$$

Thus, each consumer chooses one service s in the company T_i with probability p_{ibs} and in the company T_j with probability p_{jbs} .

We can generalize this approach for the case when the consumer chooses one company T_i from the set of companies $\{T_1, \dots, T_n\}$ to obtain the service s , and we can get the probability in case the consumer gives preference to the company T_i :

$$P_{ibs}^t = \frac{(r_{bs}^t - p_{is}^t)_s^\tau}{\sum_{j=1}^n (r_{bs}^t - p_{js}^t)_s^\tau}. \quad (14)$$

The probability that the consumer chooses one company T_i from a set of companies $\{T_1, \dots, T_n\}$ to receive service package SR have the form:

$$P_{ib}^t = \frac{\sum_{s=1}^m (r_{bs}^t - p_{is}^t)_s^\tau}{\sum_{s=1}^m \sum_{j=1}^n (r_{bs}^t - p_{js}^t)_s^\tau}. \quad (15)$$

The expected value of consumers $b_i(t)$ who chooses a company T_i is determined by the probability P_{ib}^t , which can be considered as the market share m_i^t of a company T_i has form

$$m_i^t = P_{ib}^t = \frac{\sum_{s=1}^m (r_{bs}^t - p_{is}^t)_s^\tau}{\sum_{s=1}^m \sum_{j=1}^n (r_{bs}^t - p_{js}^t)_s^\tau}, \quad \sum_{i=1}^n m_i^t = 1. \quad (16)$$

The demand of consumers for services $s \in \{1, \dots, m\}$ of the company T_i ($i \in \{1, \dots, n\}$) has the form:

$$D_{ibs}^t(p_{is}^t) = \frac{B^t P_{ib}^t}{2s_{bs}^t} (r_{bs}^t - p_{is}^t) = \frac{B^t m_i^t}{2s_{bs}^t} (r_{bs}^t - p_{is}^t). \quad (17)$$

Demand function of the consumers D_{sii}^t , who have plan to use the service SR of a company T_i , which may be implemented within network NW_i , and demand function of the consumer D_{ij}^t , who has plan to use the service SR implemented with resources of the networks NW_i and NW_j , have the form:

$$D_{sii}^t = \frac{B^t m_i^{t2}}{2s_{bs}^t} (r_{bs}^t - p_{is}^t), D_{ij}^t = \frac{B^t m_i^t m_j^t}{2s_{bs}^t} (r_{bs}^t - p_{is}^t), \tag{18}$$

where the aggregated s -service demand D_{is}^t has form:

$$D_{is}^t = D_{sii}^t + \sum_{j=1}^n D_{sij}^t = \frac{B^t m_i^{t2}}{2s_{bs}^t} (r_{bs}^t - p_{is}^t) + \sum_{j=1; i \neq j}^n \frac{B^t m_i^t m_j^t}{2s_{bs}^t} (r_{bs}^t - p_{is}^t), \tag{19}$$

and the total network traffic volume demand D_i^t for company T_i has form:

$$D_i^t = \sum_{s=1}^m \left[D_{sii}^t + \sum_{j=1}^n D_{sij}^t \right] \\ = \sum_{s=1}^m \left[\frac{B^t m_i^{t2}}{2s_{bs}^t} (r_{bs}^t - p_{is}^t) + \sum_{j=1; i \neq j}^n \frac{B^t m_i^t m_j^t}{2s_{bs}^t} (r_{bs}^t - p_{is}^t) \right],$$

where

$$D_{ii}^t = \sum_{s=1}^m D_{sii}^t, D_{ij}^t = \sum_{s=1}^m D_{sij}^t,$$

and the total network traffic volume for a company T_i has form:

$$Y_i^t = \theta D_i^t = \sum_{s=1}^m \theta_s D_{is}^t \\ = \sum_{s=1}^m \theta_s \left[\frac{B^t m_i^{t2}}{2s_{bs}^t} (r_{bs}^t - p_{is}^t) + \sum_{j=1; i \neq j}^n \frac{B^t m_i^t m_j^t}{2s_{bs}^t} (r_{bs}^t - p_{is}^t) \right], \tag{20}$$

where θ is an ‘‘average’’ linking parameter for function Y_i^t and D_i^t .

4 Revenue, ARPU, Profit

Revenue function TR_i^t companies T_i ($i \in \{1, \dots, n\}$) at the period t ($t = 1, 2, \dots, T_{\max}$) has the form:

$$TR_i^t = \sum_{i,j=1; i \neq j}^n \left[\bar{p}_i^t D_{ii}^t (\bar{p}_i^t) + (\bar{p}_i^t - \delta_{ij}^t \bar{p}_j^t) D_{ij}^t (\bar{p}_i) + \delta_{ij}^t \bar{p}_i^t D_{ji}^t (\bar{p}_j^t) \right], \tag{21}$$

where $\delta_{ij}^t \in [0, 1]$ is a parameter to be defined during negotiations between companies T_i and T_j . We assume that the cost of an access service to the competitor’s network is a value proportional to the cost of servicing by this company of its consumers.

Average revenue per user (ARPU) $ARPU_i^t$ companies T_i ($i \in \{1, \dots, n\}$) at the period t ($t = 1, 2, \dots, T_{\max}$) has the form:

$$ARPU_i^t = \frac{TR_i^t}{B^t m_i^t} = (B^t m_i^t)^{-1} \sum_{i,j=1; i \neq j}^n [\bar{p}_i^t D_{ii}^t(\bar{p}_i^t) + (\bar{p}_i^t - \delta_{ij}^t \bar{p}_j^t) D_{ij}^t(\bar{p}_i^t) + \delta_{ij}^t \bar{p}_i^t D_{ji}^t(\bar{p}_j^t)]. \quad (22)$$

Profit function Π_i^t of companies T_i ($i \in \{1, \dots, n\}$) at the period t ($t = 1, 2, \dots, T_{\max}$) has the form:

$$\begin{aligned} \Pi_i^t &= TR_i^t - TC^t(w_{J_{ik}}^t, H_{ik}^t, w_{L_{ik}}^t, c_{ik}^t, F^t), \\ TC^t &= \left(\sum_{k=1}^{s_i^J} w_{J_{ik}}^t H_{ik}^t + \sum_{k=1}^{s_i^L} w_{L_{ik}}^t c_{ik}^t \right) + F^t, \end{aligned} \quad (23)$$

where TC^t is a total costs function and F^t is a fix cost.

5 Congested Traffic Networks with RPACR

5.1 Traffic Optimization Problem in Congested Networks with RPACR

We can formulate an optimization problem for each company T_i ($i \in \{1, \dots, n\}$) at any time $t \in \{1, 2, \dots, T_{\max}\}$:

$$\begin{cases} \partial \Pi_i^t / \partial p_i^t &= 0; \\ \partial^2 \Pi_i^t / \partial p_i^{t2} &< 0 \\ Y_i^t &\leq \bar{Y}_i^t, \end{cases} \quad (24)$$

where \bar{Y}_i^t is the maximum peak of the total network traffic volume for a company T_i .

The following theorem holds true.

Provided that the parameters $\theta_s > 0$, $\bar{a}^t > 0$, $\bar{b}^t > 0$, $\delta_{ij}^t \in [0, 1]$, $w_{J_{ij}}^t \geq 0$, $w_{L_{ij}}^t \geq 0$, $F^t \geq 0$, $\bar{Y}_i^t > 0$ there is a unique solution of the problem (24) in the form of the equilibrium value of the tariff for the use of services SR of company $i \in \{1, \dots, n\}$ during the period t :

$$\bar{p}_{it}^* = \left(m_i^t + \sum_{j=1; i \neq j}^n \delta_{ij}^t m_j^t \right) \frac{\bar{a}^t}{2\bar{b}^t}.$$

Proof. Let's write out the profit function of i company in the form of:

$$\begin{aligned} \Pi_i^t &= \sum_{i,j;i \neq j}^n [\bar{p}_i^t m_i^{t2} (\bar{a}^t - \bar{a}^t \bar{p}_i^t) + m_i^t m_j^t (\bar{p}_i^t - \delta_{ij}^t \bar{p}_j^t) (\bar{a}^t - \bar{b}^t \bar{p}_i^t) \\ &\quad + \delta_{ij}^t m_j^t m_i^t \bar{p}_i^t (\bar{a}^t - \bar{b}^t \bar{p}_j^t)] - \left(\sum_{k=1}^{s_i^J} w_{J_{ik}}^t H_{ik}^t + \sum_{k=1}^{s_i^L} w_{L_{ik}}^t c_{ik}^t \right) - F^t, \end{aligned}$$

We can calculate the derivatives of \bar{p}_i^t and equals them to zero, we obtain a system of algebraic equations of the form:

$$m_i^t (\bar{a}^t - 2\bar{b}^t \bar{p}_i^t) + \sum_{j=1; j \neq i}^n [m_j^t (\bar{a}^t - 2\bar{b}^t \bar{p}_i^t + \delta_{ij}^t \bar{b}^t \bar{p}_j^t) + \delta_{ij}^t m_j^t (\bar{a}^t - \bar{b}^t \bar{p}_j^t)] = 0,$$

and the equilibrium value of the tariff has form:

$$\bar{p}_{it}^* = \left(m_i^t + \sum_{j=1; j \neq i}^n \delta_{ij}^t m_j^t \right) \frac{\bar{a}^t}{2\bar{b}^t}.$$

We can obtain for $\partial^2 \Pi_i^t / \partial \bar{p}_i^{t2}$,

$$\frac{\partial^2 \Pi_i^t}{\partial \bar{p}_i^{t2}} = \sum_{i, j; i \neq j} [-m_i^{t2} 2\bar{b}^t - m_i^t m_j^t 2\bar{b}^t - \delta_{ij}^t m_j^t m_i^t \bar{b}^t \bar{p}_j^t] < 0.$$

The theorem is proved.

We can formulate an optimization problem for each company T_i ($i \in \{1, \dots, n\}$) at any time $t \in \{1, 2, \dots, T_{\max}\}$ for the tariff value \bar{p}_{it}^* :

$$\begin{cases} \partial \Pi_i^t(\bar{p}_i^*, \delta_{ij}^t) / \partial \delta_{ij}^t = 0; \\ \partial^2 \Pi_i^t(\bar{p}_i^*, \delta_{ij}^t) / \partial \delta_{ij}^{t2} < 0; \\ Y_i^t \leq \bar{Y}_i^t, \end{cases}$$

which allows maximizing the profit of each company of T_i using the parameter δ_{ij}^t with condition $Y_i^t \leq \bar{Y}_i^t$.

After substituting the corresponding equilibrium tariffs \bar{p}_{it}^* in the profit function, we obtain the following equation

$$\begin{aligned} \Pi_i^t = \sum_{i, j; i \neq j}^n \frac{\bar{a}^{t2} m_i^t [m_i^t + m_j^t]}{2\bar{b}^t} \Phi(m_i^t, m_j^t, \delta_{ij}^t) (1 - 0.5\Phi(m_i^t, m_j^t, \delta_{ij}^t)) \\ - \left(\sum_{k=1}^{s_i^J} w_{J_{ik}}^t H_{ik}^t + \sum_{k=1}^{s_i^L} w_{L_{ik}}^t C_{ik}^t \right) - F^t, \end{aligned}$$

where

$$\Phi(m_i^t, m_j^t, \delta_{ij}^t) = m_i^t + \sum_{j=1; j \neq i}^n \delta_{ij}^t m_j^t,$$

and differentiating by δ_{ij}^t and equating to zero, we have a system of algebraic equations, solving which, we obtain an equilibrium value of $\delta_i^* = 0.5$.

The equilibrium tariff \bar{p}_{it}^* for the services of company T_i , taking into account the optimal value $\delta_i^* = 0.5$ during the period t , has the form:

$$\bar{p}_{it}^* = (m_i^t + 1) \frac{\bar{a}^t}{4\bar{b}^t},$$

The equilibrium demand function for the company T_i ($i \in \{1, \dots, n\}$) services SR at any t can be represented as follows:

$$D_{it}^* (\bar{p}_{it}^*) = m_i^t D_t (\bar{p}_{it}^*) = 0.25 m_i^t \bar{a}^t (3 - m_i^t),$$

and the total network traffic volume for a company T_i with the equilibrium tariff has form:

$$Y_i^t = \theta D_{it}^* = 0.25 \theta m_i^t \bar{a}^t (3 - m_i^t) \leq \bar{Y}_i^t.$$

The total equilibrium market demand function D_t^* and the total equilibrium traffic volume Y_t^* for services SR at any t has the form:

$$D_t^* = \bar{a}^t \left(3 - \sum_{i=1}^n m_i^{t2} \right), \quad Y_t^* = \theta \bar{a}^t \left(3 - \sum_{i=1}^n m_i^{t2} \right)$$

and we can show that with a uniform distribution of customers between all companies T_i ($i \in \{1, \dots, n\}$) the total equilibrium market demand function the total equilibrium traffic volume for services SR reaches maximum.

If the network bandwidth of companies is less than the traffic volume that subscribers generate, then companies can manage the overload by creating such tariffs that reduces the overload on the network.

5.2 Numerical Analysis of Traffic Optimization Problem in Congested Networks

Let's consider this model in the case of the oligopoly, when two companies are present in the market of telecommunication services, using numerical analysis. Let the duration of the calculations be $t_{max} = 36$ then the equilibrium tariff \bar{p}_{it}^* for the services of company T_i ($i = 1, 2$), taking into account the optimal value $\delta_t^* = 0.5$ during this $t = 0, 1, \dots, 36$, has the form:

$$\bar{p}_{it}^* = (m_i^t + 1) \frac{\bar{a}^t}{4b^t}, \quad i = 1, 2.$$

Let's suppose that the market share between two companies changes in such way as it is presented (see Fig. 1).

In this case the network traffic volume dynamics for a company T_i ($i = 1, 2$) with subject to the equilibrium tariff it is presented at (Fig. 2). We can see that each company try to increase the capacity of own network. There is fluctuation of network capacity in the long term, which may be due to the periodic transition of users from one company to the other and back.

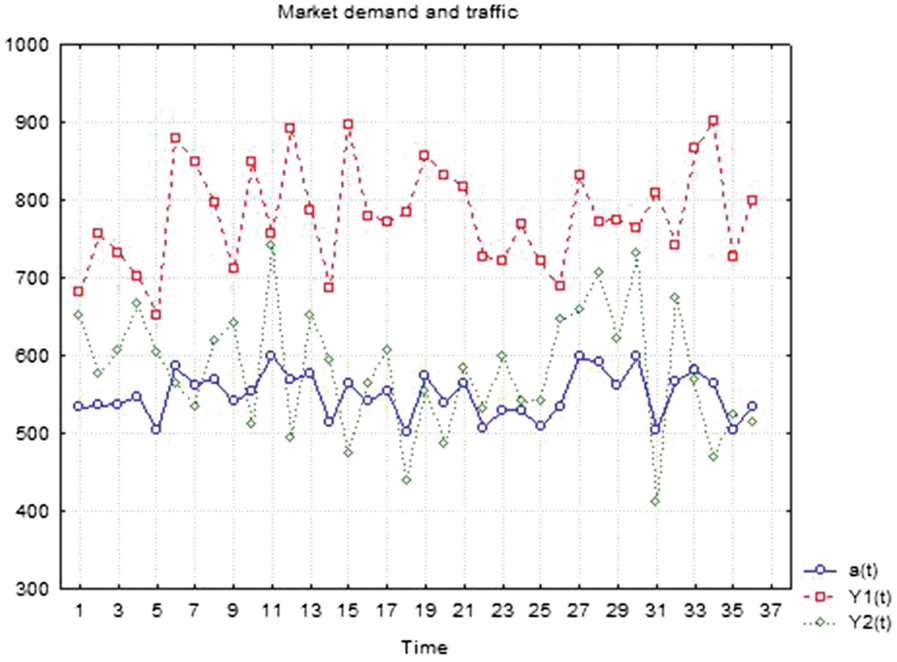


Fig. 1. Market share dynamics.

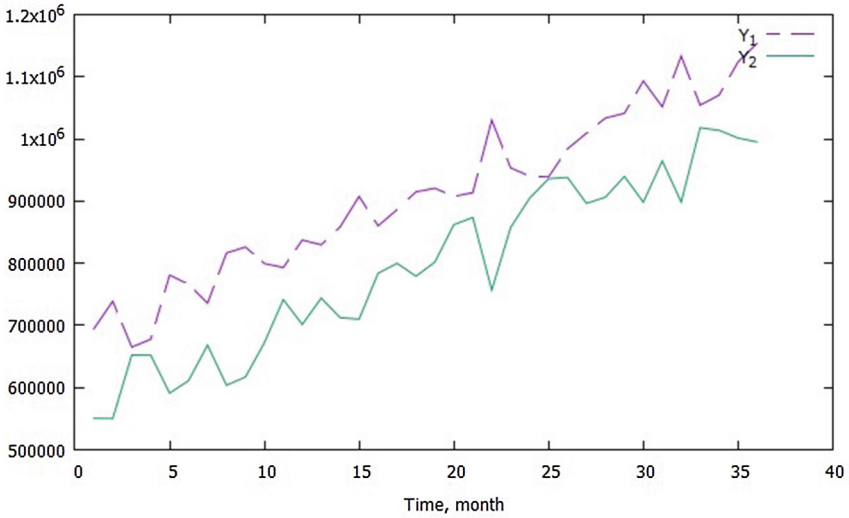


Fig. 2. Network traffic volume dynamics.

6 Conclusions

In this paper a mathematical model of the telecommunications market is constructed taking into account congested in networks. It is carried out the analysis of equilibrium tariffs for telecommunications services for this type of market with multi-sided pricing.

The applied value of the model is that the use of PACR telecommunication companies does not require detailed information market telecommunications, as the number of parameters of the model is optimized. This model proved to be effective in the analysis the dynamics of the telecommunications market, as it allows companies to respond flexibly to external changes, which allows timely to change the strategy. The proposed model can serve as a tool for analyzing the existence of collusion between companies in the telecommunications industry market with congested networks. Numerical calculation is performed to illustrate the results.

Acknowledgments. The publication has been prepared with the support of the “RUDN University Program 5-100”.

References

1. Armstrong, M.: Nonlinear pricing. *Ann. Rev. Econ.* **8**, 583–614 (2016)
2. Cramton, P., Doyle, L.: Open access wireless markets. *Telecommun. Policy* **41**(5–6), 379–390 (2017)
3. Gaidamaka, Y., Sopin, E., Talanova, M.: Approach to the analysis of probability measures of cloud computing systems with dynamic scaling. In: Vishnevsky, V., Kozyrev, D. (eds.) *DCCN 2015. CCIS*, vol. 601, pp. 121–131. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30843-2_13
4. Dessein, W.: Network competition in nonlinear pricing. *Rand J. Econ.* **34**, 593–611 (2003)
5. Dessein, W.: Network competition with heterogeneous customers and calling patterns. *Inform. Econ. Policy* **16**, 323–345 (2004)
6. Doganoglu, T., Tauman, Y.: Network competition and access charge rules. *Manch. Sch.* **70**, 16–35 (2002)
7. Hahn, J.-H.: Network competition and interconnection with heterogeneous subscribers. *Int. J. Ind. Organ.* **22**, 611–631 (2004)
8. Korolkova, A.V., Eferina, E.G., Laneev, E.B., Gudkova, I.A., Sevastianov, L.A., Kulyabov, D.S.: Stochastization of one-step processes in the occupations number representation. In: *Proceedings - 30th European Conference on Modelling and Simulation, ECMS 2016*, pp. 698–704 (2016)
9. Mohlmeier, P., Rusinowska, A., Tanimura, E.: Competition for the access to and use of information in networks. *Math. Soc. Sci.* **92**(C), 48–63 (2018)
10. Laffont, J.-J., Tirole, J.: Access pricing and competition. *Eur. Econ. Rev.* **38**, 1673 (1994)
11. Laffont, J.-J., Rey, P., Tirole, J.: Network competition I: overview and nondiscriminatory pricing. *Rand J. Econ.* **29**, 1–37 (1999)
12. Laffont, J.-J., Rey, P., Tirole, J.: Network competition II: price discrimination. *Rand J. Econ.* **29**, 38–56 (1998)

13. Laffont, J.-J., Tirole, J.: Internet interconnection and the off-net-cost pricing principle. *Rand J. Econ.* **34**, 73–95 (2003)
14. Laffont, J.-J., Tirole, J.: Receiver-pays principle. *Rand J. Econ.* **35**, 85–110 (2004)
15. Lee, S.-H., Jeong, H.-Y., Seo, S.-W.: Optimal pricing and capacity partitioning for tiered access service in virtual networks. *Comput. Netw.* **57**(18), 3941–3956 (2013)
16. Samouylov, K., Naumov, V., Sopin, E., Gudkova, I., Shorgin, S.: Sojourn time analysis for processor sharing loss system with unreliable server. In: Wittevrongel, S., Phung-Duc, T. (eds.) *ASMTA 2016. LNCS*, vol. 9845, pp. 284–297. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43904-4_20
17. Se-Hak, C.: Network capacity and access pricing for cloud services. *Procedia Soc. Behav. Sci.* **109**, 1348–1352 (2014)
18. Sevastianov, L.A., Vasilyev, S.A.: Large-scale queuing systems and services pricing. In: 9th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, (ICUMT 2017), pp. 7–12 (2017)
19. Ushchev, P., Zenou, Y.: Price competition in product variety networks. *Games Econ. Behav.* **110**, 226–247 (2018)
20. Vasilyev, S.A., Sevastianov, L.A., Urusova, D.A.: Economics and mathematical modeling of oligopoly telecommunication market. *Math. Inform. Sci. Phys.* **2**, 59–69 (2011). *Bulletin of Peoples Friendship University of Russia*



Retrial Queueing System of MMPP/M/2 Type with Impatient Calls in the Orbit

Olga Vygovskaya, Elena Danilyuk^(✉), and Svetlana Moiseeva

National Research Tomsk State University, Lenina Avenue, 36, 634050 Tomsk, Russia
osipovich.olga@bk.ru, daniluc.elena@sibmail.com, smoiseeva@mail.ru

Abstract. In the paper, the retrial queueing system of MMPP/M/2 type with input MMPP-flow of events and impatient calls is considered. The delay time of calls in the orbit, the calls service time and the impatience time of calls in the orbit have exponential distribution. Asymptotic analysis method is proposed for the solving problem of finding distribution of the number of calls in the orbit under a system heavy load and long time patience of calls in the orbit condition. The theorem about the Gauss form of the asymptotic probability distribution of the number of calls in the orbit is formulated and proved. Numerical illustrations, results are also given.

Keywords: Two-server retrial queueing system · Orbit
Asymptotic analysis · Impatient calls

1 Introduction

Queueing systems with repeated calls, or Retrial Queueing Systems, are mathematical models widely used for many real objects, systems and processes analysis and optimization, especially telecommunication systems, networks, mobile networks, call-centres, manufacturing, economics. In these queueing systems unserved calls are not lost when there are not available service devices (servers are busy or broken). So, the customers that don't get a service repeat to occupy server after a random time.

There are many papers devoted to RQ-systems study [1–14, 18, 19, 22–25]. The main results and comprehensive description of retrial queues are contained in the books [5, 6].

Models with calls leaved RQ-system after failed attempt to get a service was considered by many scientists [12–17], etc. In these studies, an arriving call joints the orbit with some probability p and leaves the system with the probability $1-p$ when there are not available service devices at the time. Some authors name such customers as non-persistent or p -non-persistent customers.

We consider a different model which was not been investigated early. So, in present research impatient customer is a customer in the orbit that can repeat an attempt to reach the server again or can leave the orbit after a random time without server recalling.

Classical retrial models consist of one server but real telecommunication systems are usually multiserver retrial queue [9, 18–21]. In the proposed paper RQ-system consisting of two service devices is considered. Applying the Poisson input flow gives a big determination error of the characteristics of the service quality in real systems. Validity of the models with Markov Modulated Poisson Process as input flow for multi-server queueing systems description is shown in papers [22–24]. And we are considering the MMPP input flow with two servers and impatient calls under system heavy load condition (for example, the problem for MMPP input flow with one server under system heavy load condition is solved in [25]).

Asymptotic analysis method is widely applied for RQ-systems research. The method makes it possible to produce analytical result for different types of queueing systems and networks under given asymptotic condition. More information about the asymptotic analysis method is provided in [5–7, 9, 17, 26], etc.

The general information about mathematical model of the retrial queueing system discussed in the paper and the problem statement are presented in the Sect. 2. In the Sect. 3 the detailed derivation of the model and the system of Kolmogorov equations for the stationary state probabilities are cited. The Sect. 4 consists of the decision of the problem under study by the asymptotic analysis method. As a result of the section the Theorem about stationary probability distribution of the calls number in the orbit for Retrial queueing system of MMPP/M/2 type with impatient calls in the orbit under a system heavy load and long time patience of calls in the orbit condition is formulated and proved. Some numerical results, graphs, that proved the theoretical results, are performed in the Sect. 5. Section 6 concludes the paper.

2 Mathematical Model

A retrial queueing system consisting of an infinite orbit and two servers is considered. The input flow is defined by the Markov Modulated Poisson Process and it is defined by matrix $\mathbf{Q} = ||q_{ij}||$, $i, j = 1, 2, \dots, S$, and matrix $\mathbf{\Lambda} = \text{diag} \{ \lambda_1 \lambda_2 \dots \lambda_S \}$. The rate of MMPP-flow is defined as $\lambda = \mathbf{r}\mathbf{\Lambda}\mathbf{e}$, where S -sized vectors \mathbf{r} and \mathbf{e} are given below. The service times on every of the two servers are exponentially distributed with parameter μ . A customer which arrives into the system, when at least one of the two servers is free, instantly occupies this server. If all of the devices are busy, the call goes to the orbit, where it stays during a random time distributed exponentially with parameter σ . After the delay the customer makes an attempt to reach any server again. If it is free, the call occupies it, otherwise the call immediately joins the orbit. From the orbit calls (impatient calls) can leave the system after a random time distributed exponentially with parameter α .

The structure of the model is presented in Fig. 1.

The problem is to get stationary probability distribution of the number of calls in the orbit for the system under review.

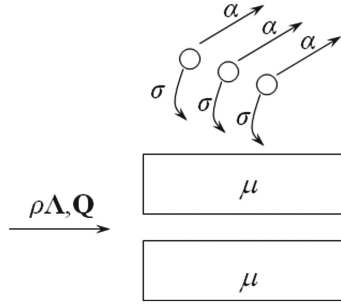


Fig. 1. Retrial queue MMPP/M/2 with impatient calls in the orbit

3 Process of the System States: Stationary Distribution

Let us consider Markovian process $\{n(t), s(t), i(t)\}$ determined states of the Retrial queue MMPP/M/2 with impatient customer in the orbit where the random process $i(t)$ is the number of calls in the orbit at the moment t , $i(t) = 0, 1, 2, 3, \dots$, $s(t)$ is the value of the Markov chain state that specifies MMPP-flow at the moment t , $s(t) = 1, 2, \dots, S$, the random process $n(t)$ defines device state at the moment t and takes one of the following values

$$n(t) = \begin{cases} 0, & \text{if all servers are free at the moment } t; \\ 1, & \text{if only one of two servers is busy at the moment } t; \\ 2, & \text{if both servers are busy at the moment } t. \end{cases}$$

Denote the probability that, at the moment t , the server (service device) is in the state n , $n = 0, 1, 2$, the Markov chain managing the input calls flow is in the state s , $s = 1, 2, \dots, S$, and there are i calls in the orbit, $i = 0, 1, 2, \dots$, as $P\{n(t) = n, s(t) = s, i(t) = i\} = P_n(s, i, t)$. We write the following system of equations for $s = 1, 2, \dots, S$

$$\left\{ \begin{aligned} \frac{\partial P_0(s, i, t)}{\partial t} &= -(\lambda_s + i\sigma + i\alpha - q_{ss}) P_0(s, i, t) + (i + 1)\alpha P_0(s, i + 1, t) \\ &+ \mu P_1(s, i, t) + \sum_{\substack{\nu=1, \\ \nu \neq s}}^S P_0(\nu, i, t) q_{\nu s}, \\ \frac{\partial P_1(s, i, t)}{\partial t} &= -(\lambda_s + i\sigma + i\alpha + \mu - q_{ss}) P_1(s, i, t) + (i + 1)\sigma P_0(s, i + 1, t) \\ &+ \lambda_s P_0(s, i, t) + (i + 1)\alpha P_1(s, i + 1, t) + 2\mu P_2(s, i, t) + \sum_{\substack{\nu=1, \\ \nu \neq s}}^S P_1(\nu, i, t) q_{\nu s}, \\ \frac{\partial P_2(s, i, t)}{\partial t} &= -(\lambda_s + i\alpha + 2\mu - q_{ss}) P_2(s, i, t) + (i + 1)\sigma P_1(s, i + 1, t) \\ &+ \lambda_s P_1(s, i, t) + (i + 1)\alpha P_2(s, i + 1, t) + \lambda_s P_2(s, i - 1, t) + \sum_{\substack{\nu=1, \\ \nu \neq s}}^S P_2(\nu, i, t) q_{\nu s}. \end{aligned} \right. \tag{1}$$

Let the row-vector \mathbf{r} is the stationary probability distribution of the underlying process $s(t)$ and it is defined as the unique solution of the system

$$\begin{cases} \mathbf{r}\mathbf{Q} = \mathbf{0}, \\ \mathbf{r}\mathbf{e} = 1, \end{cases}$$

where \mathbf{e} is unit column-vector, $\mathbf{0}$ is zero row-vector.

Denote the row-vectors $\mathbf{\Pi}_n(i) = \{ \Pi_n(1, i) \ \Pi_n(2, i) \ \dots \ \Pi_n(S, i) \}$, $n = 0, 1, 2$, $i = 0, 1, 2, \dots$, where $\Pi_n(s, i) = \lim_{t \rightarrow \infty} P_n(s, i, t)$. Then the system of Kolmogorov equations for the stationary state probabilities $\mathbf{\Pi}_n(i)$ of the process $\{n(t), s(t), i(t)\}$ is written as follows

$$\begin{cases} \mathbf{\Pi}_0(i) (-\mathbf{\Lambda} - i(\sigma + \alpha)\mathbf{I} + \mathbf{Q}) + \mu\mathbf{\Pi}_1(i) + (i + 1)\alpha\mathbf{\Pi}_0(i + 1) = \mathbf{0}, \\ \mathbf{\Pi}_1(i) (-\mathbf{\Lambda} - i(\sigma + \alpha)\mathbf{I} + \mathbf{Q} - \mu\mathbf{I}) + \mathbf{\Pi}_0(i)\mathbf{\Lambda} + (i + 1)\sigma\mathbf{\Pi}_0(i + 1) \\ + (i + 1)\alpha\mathbf{\Pi}_1(i + 1) + 2\mu\mathbf{\Pi}_2(i) = \mathbf{0}, \\ \mathbf{\Pi}_2(i) (-\mathbf{\Lambda} - i\alpha\mathbf{I} + \mathbf{Q} - 2\mu\mathbf{I}) + \mathbf{\Pi}_1(i)\mathbf{\Lambda} + (i + 1)\sigma\mathbf{\Pi}_1(i + 1) \\ + (i + 1)\alpha\mathbf{\Pi}_2(i + 1) + \mathbf{\Pi}_2(i - 1)\mathbf{\Lambda} = \mathbf{0}, \end{cases} \tag{2}$$

where \mathbf{I} is the identity matrix.

We get in (2) the indefinite dimensional system of matrix difference equations with variable coefficients. In common case it is not possible to produce the exact solution of this system. To find solution of (2), we will use the method of asymptotic analysis under a system heavy load and long time patience of calls in the orbit condition.

4 Asymptotic Analysis Method

We introduce the partial characteristic functions

$$H_n(s, u) = \sum_{i=0}^{\infty} e^{ju i} \Pi_n(s, i), \quad H_n(s, 0) = \sum_{i=0}^{\infty} \Pi_n(s, i) = R_n, \tag{3}$$

where $j = \sqrt{-1}$, $n = 0, 1, 2$, $s = 1, 2, \dots, S$, and R_n are stationary state probabilities of the process $n(t)$.

Using (3), row-vectors $\mathbf{H}_n(u) = \{H_n(1, u) \ H_n(2, u) \ \dots \ H_n(S, u)\}$ and $\mathbf{H}'_n(u) = \left\{ \frac{\partial H_n(1, u)}{\partial u} \ \frac{\partial H_n(2, u)}{\partial u} \ \dots \ \frac{\partial H_n(S, u)}{\partial u} \right\}$, where $\frac{\partial H_n(s, u)}{\partial u} =$

$j \sum_{i=0}^{\infty} i e^{ju i} \Pi_n(s, i)$, $s = 1, 2, \dots, S$, $n = 0, 1, 2$, we can write the system (2) as

$$\begin{cases} j \left(\sigma + \alpha (1 - e^{-ju}) \right) \mathbf{H}'_0(u) + \mu\mathbf{H}_1(u) + \mathbf{H}_0(u) (\mathbf{Q} - \mathbf{\Lambda}) = \mathbf{0}, \\ j \left(\sigma + \alpha (1 - e^{-ju}) \right) \mathbf{H}'_1(u) + \mathbf{H}_1(u) (\mathbf{Q} - \mathbf{\Lambda}) - j\sigma e^{-ju} \mathbf{H}'_0(u) - \mu\mathbf{H}_1(u) \\ + \mathbf{H}_0(u)\mathbf{\Lambda} + 2\mu\mathbf{H}_2(u) = \mathbf{0}, \\ j\alpha (1 - e^{-ju}) \mathbf{H}'_2(u) + \mathbf{H}_2(u) (-2\mu\mathbf{I} + \mathbf{Q} - \mathbf{\Lambda} (1 - e^{ju})) \\ + \mathbf{H}_1(u)\mathbf{\Lambda} - j\sigma e^{-ju} \mathbf{H}'_1(u) = \mathbf{0}. \end{cases} \tag{4}$$

In adding the third equation by the first and the second equations of (4) we get the system below

$$\left\{ \begin{aligned} & j(\sigma + \alpha(1 - e^{-ju})) \mathbf{H}'_0(u) + \mu \mathbf{H}_1(u) + \mathbf{H}_0(u) (\mathbf{Q} - \mathbf{\Lambda}) = \mathbf{0}, \\ & j(\sigma + \alpha(1 - e^{-ju})) \mathbf{H}'_1(u) + \mathbf{H}_1(u) (\mathbf{Q} - \mathbf{\Lambda}) - j\sigma e^{-ju} \mathbf{H}'_0(u) - \mu \mathbf{H}_1(u) \\ & + \mathbf{H}_0(u) \mathbf{\Lambda} + 2\mu \mathbf{H}_2(u) = \mathbf{0}, \\ & j\alpha(1 - e^{-ju}) \sum_{k=0}^2 \mathbf{H}'_k(u) + j\sigma \sum_{k=0}^1 \mathbf{H}'_k(u) - j\sigma e^{-ju} \sum_{k=0}^1 \mathbf{H}'_k(u) \\ & + \sum_{k=0}^2 \mathbf{H}_k(u) \mathbf{Q} - (1 - e^{ju}) \mathbf{H}_2(u) \mathbf{\Lambda} = \mathbf{0}. \end{aligned} \right. \tag{5}$$

The system in (5) is the base system for analysis of Retrial queueing system of MMPP/M/2 type with impatient calls in the orbit under a system heavy load ($\lambda \gg 2\mu$) and long time patience of calls in the orbit ($\alpha \rightarrow 0$) condition.

Theorem 1. *Stationary probability distribution of the calls number in the orbit for Retrial queueing system of MMPP/M/2 type with impatient calls in the orbit under a system heavy load and long time patience of calls in the orbit condition can be approximated by the Gaussian distribution with mean and variance equal to $\frac{\lambda - 2\mu}{\alpha}$ and $\frac{\lambda}{\alpha}$ respectively, where $\lambda = \mathbf{r}\mathbf{\Lambda}\mathbf{e}$, and $\mathbf{\Lambda}$ is the matrix of the input calls flow parameters, \mathbf{r} is the row-vector of the stationary probability distribution of the process $s(t)$, \mathbf{e} is the unit row-vector, μ, σ, α are the exponential distribution parameters, accordingly, of the calls service time, the calls delay time in the orbit, the calls leaving the system from the orbit.*

Proof. The Theorem 1 proving will carried out in two stages.

Stage 1. Let to denote $\alpha = \varepsilon, u = \varepsilon w, \mathbf{H}_0(u) = \varepsilon^2 \mathbf{F}_0(w, \varepsilon), \mathbf{H}_1(u) = \varepsilon \mathbf{F}_1(w, \varepsilon), \mathbf{H}_2(u) = \mathbf{F}_2(w, \varepsilon)$, where $\mathbf{F}_n(w, \varepsilon) = \{F_n(1, w, \varepsilon) \ F_n(2, w, \varepsilon) \ \dots \ F_n(S, w, \varepsilon)\}, n = 0, 1, 2$, and $\varepsilon \rightarrow 0$ is infinitesimal.

Since $\mathbf{H}'_0(u) = \varepsilon \frac{\partial \mathbf{F}_0(w, \varepsilon)}{\partial w}, \mathbf{H}'_1(u) = \frac{\partial \mathbf{F}_1(w, \varepsilon)}{\partial w}, \mathbf{H}'_2(u) = \frac{1}{\varepsilon} \frac{\partial \mathbf{F}_2(w, \varepsilon)}{\partial w}$, where $\frac{\partial \mathbf{F}_n(w, \varepsilon)}{\partial w} = \left\{ \frac{\partial F_n(1, w, \varepsilon)}{\partial w} \ \frac{\partial F_n(2, w, \varepsilon)}{\partial w} \ \dots \ \frac{\partial F_n(S, w, \varepsilon)}{\partial w} \right\}, n = 0, 1, 2$, the equations system (5) can be written as

$$\left\{ \begin{aligned} & j(\sigma + \varepsilon(1 - e^{-j\varepsilon w})) \frac{\partial \mathbf{F}_0(w, \varepsilon)}{\partial w} + \mu \mathbf{F}_1(w, \varepsilon) + \varepsilon \mathbf{F}_0(w, \varepsilon) (\mathbf{Q} - \mathbf{\Lambda}) = \mathbf{0}, \\ & j(\sigma + \varepsilon(1 - e^{-j\varepsilon w})) \frac{\partial \mathbf{F}_1(w, \varepsilon)}{\partial w} + \varepsilon \mathbf{F}_1(w, \varepsilon) (\mathbf{Q} - \mathbf{\Lambda}) - j\sigma e^{-j\varepsilon w} \varepsilon \frac{\partial \mathbf{F}_0(w, \varepsilon)}{\partial w} \\ & - \mu \varepsilon \mathbf{F}_1(w, \varepsilon) + \varepsilon^2 \mathbf{F}_0(w, \varepsilon) \mathbf{\Lambda} + 2\mu \mathbf{F}_2(w, \varepsilon) = \mathbf{0}, \\ & j\varepsilon(1 - e^{-j\varepsilon w}) \sum_{k=0}^2 \varepsilon^{1-k} \frac{\partial \mathbf{F}_k(w, \varepsilon)}{\partial w} + j\sigma(1 - e^{-j\varepsilon w}) \sum_{k=0}^1 \varepsilon^{1-k} \frac{\partial \mathbf{F}_k(w, \varepsilon)}{\partial w} \\ & + \sum_{k=0}^2 \varepsilon^{2-k} \mathbf{F}_k(w, \varepsilon) \mathbf{Q} - (1 - e^{j\varepsilon w}) \mathbf{F}_2(w, \varepsilon) \mathbf{\Lambda} = \mathbf{0}. \end{aligned} \right. \tag{6}$$

The transformation of the first and the second equations of (6) under $\varepsilon \rightarrow 0$ with $\mathbf{F}_n(w) = \lim_{\varepsilon \rightarrow 0} \mathbf{F}_n(w, \varepsilon)$, $n = 0, 1, 2$, leads to equations system as follows

$$\begin{cases} j\sigma \frac{d\mathbf{F}_0(w)}{dw} = -\mu\mathbf{F}_1(w), \\ j\sigma \frac{d\mathbf{F}_1(w)}{dw} = -2\mu\mathbf{F}_2(w), \\ j^2\varepsilon w \sum_{k=0}^2 \varepsilon^{2-k} \frac{d\mathbf{F}_k(w)}{dw} + j^2\sigma\varepsilon w \sum_{k=0}^1 \varepsilon^{1-k} \frac{d\mathbf{F}_k(w)}{dw} + \sum_{k=0}^2 \varepsilon^{2-k} \mathbf{F}_k(w)\mathbf{Q} \\ + j\varepsilon w \mathbf{F}_2(w)\mathbf{\Lambda} = \mathbf{0}. \end{cases} \tag{7}$$

In multiplying the last equation in the (7) by the unit column-vector \mathbf{e} and using $\mathbf{Q}\mathbf{e} = \mathbf{0}$

$$j\varepsilon w \sum_{k=0}^2 \varepsilon^{2-k} \frac{d\mathbf{F}_k(w)}{dw} \mathbf{e} + j\sigma\varepsilon w \sum_{k=0}^1 \varepsilon^{1-k} \frac{d\mathbf{F}_k(w)}{dw} \mathbf{e} + \varepsilon w \mathbf{F}_2(w)\mathbf{\Lambda}\mathbf{e} = \mathbf{0},$$

or under $\varepsilon \rightarrow 0$

$$j \frac{d\mathbf{F}_2(w)}{dw} \mathbf{e} + j\sigma \frac{d\mathbf{F}_1(w)}{dw} \mathbf{e} + \mathbf{F}_2(w)\mathbf{\Lambda}\mathbf{e} = \mathbf{0}. \tag{8}$$

We suggest to find the Eq. (8) solution $\mathbf{F}_2(w)$ in the form

$$\mathbf{F}_2(w) = \mathbf{r}\Phi(w). \tag{9}$$

Substituting (9) and (7) in (8) with expression $\mathbf{r}\mathbf{e} = 1$ we have

$$\mathbf{F}_2(w) = R_2 \exp\{(\lambda - 2\mu)jw\} \mathbf{r}, \tag{10}$$

where $\lambda = \mathbf{r}\mathbf{\Lambda}\mathbf{e}$, and R_2 is defined above.

Pre-limit characteristic function $h(u)$ is approximately equal to

$$h(u) = \{\mathbf{H}_0(u) + \mathbf{H}_1(u) + \mathbf{H}_2(u)\} \mathbf{e} = \mathbf{F}_2\left(\frac{u}{\varepsilon}\right) \mathbf{e} + o(\varepsilon) \approx \mathbf{F}_2\left(\frac{u}{\varepsilon}\right) \mathbf{e}.$$

So, the first-order asymptotic characteristic function $h^{(1)}(u)$ of the probability distribution of the number of calls in the orbit under the system heavy load and long time patience of calls in the orbit condition can be presented as

$$h^{(1)}(u) = \mathbf{F}_2\left(\frac{u}{\varepsilon}\right) \mathbf{e} = R_2 \exp\left\{(\lambda - 2\mu)\frac{j u}{\varepsilon}\right\} \mathbf{r}\mathbf{e} = R_2 \exp\left\{(\lambda - 2\mu)\frac{j u}{\alpha}\right\}. \tag{11}$$

Stage 2. Denoting in the base system of Eqs. (5) with (11)

$$\mathbf{H}_n(u) = R_2 \exp\left\{(\lambda - 2\mu)\frac{j u}{\alpha}\right\} \mathbf{H}_n^{(2)}(u), \quad n = 0, 1, 2, \tag{12}$$

and making some transformations with this system we get (13)

$$\left\{ \begin{aligned} & j(\sigma + \alpha(1 - e^{-ju})) \left(\mathbf{H}_0^{(2)'}(u) + j \frac{\lambda - 2\mu}{\alpha} \mathbf{H}_0^{(2)}(u) \right) \\ & + \mu \mathbf{H}_1^{(2)}(u) + \mathbf{H}_0^{(2)}(u) (\mathbf{Q} - \mathbf{\Lambda}) = \mathbf{0}, \\ & j(\sigma + \alpha(1 - e^{-ju})) \left(\mathbf{H}_1^{(2)'}(u) + j \frac{\lambda - 2\mu}{\alpha} \mathbf{H}_1^{(2)}(u) \right) + \mathbf{H}_1^{(2)}(u) (\mathbf{Q} - \mathbf{\Lambda}) \\ & - j\sigma e^{-ju} \left(\mathbf{H}_0^{(2)'}(u) + j \frac{\lambda - 2\mu}{\alpha} \mathbf{H}_0^{(2)}(u) \right) - \mu \mathbf{H}_1^{(2)}(u) \\ & + \mathbf{H}_0^{(2)}(u) \mathbf{\Lambda} + 2\mu \mathbf{H}_2^{(2)}(u) = \mathbf{0}, \\ & j\alpha(1 - e^{-ju}) \sum_{k=0}^2 \left[\mathbf{H}_k^{(2)'}(u) + j \frac{\lambda - 2\mu}{\alpha} \mathbf{H}_k^{(2)}(u) \right] - (1 - e^{ju}) \mathbf{H}_2^{(2)}(u) \mathbf{\Lambda} \\ & + j\sigma(1 - e^{-ju}) \sum_{k=0}^1 \left[\mathbf{H}_k^{(2)'}(u) + j \frac{\lambda - 2\mu}{\alpha} \mathbf{H}_k^{(2)}(u) \right] + \sum_{k=0}^2 \mathbf{H}_k^{(2)}(u) \mathbf{Q} = \mathbf{0}. \end{aligned} \right. \tag{13}$$

Let $\alpha = \varepsilon^2$, $u = \varepsilon w$, $\mathbf{H}_0^{(2)}(u) = \varepsilon^4 \mathbf{F}_0^{(2)}(w, \varepsilon)$, $\mathbf{H}_1^{(2)}(u) = \varepsilon^2 \mathbf{F}_1^{(2)}(w, \varepsilon)$, $\mathbf{H}_2^{(2)}(u) = \mathbf{F}_2^{(2)}(w, \varepsilon)$, where $\varepsilon \rightarrow 0$ is infinitesimal.

Taking into account $\mathbf{H}_0^{(2)'}(u) = \varepsilon^3 \mathbf{F}_0^{(2)'}(w, \varepsilon)$, $\mathbf{H}_1^{(2)'}(u) = \varepsilon \mathbf{F}_1^{(2)'}(w, \varepsilon)$, $\mathbf{H}_2^{(2)'}(u) = \frac{1}{\varepsilon} \mathbf{F}_2^{(2)'}(w, \varepsilon)$, we get the system (13) in the form below

$$\left\{ \begin{aligned} & j(\sigma + \varepsilon^2(1 - e^{-j\varepsilon w})) \left[\varepsilon^3 \mathbf{F}_0^{(2)'}(w, \varepsilon) + j(\lambda - 2\mu) \varepsilon^2 \mathbf{F}_0^{(2)}(w, \varepsilon) \right] \\ & + \mu \varepsilon^2 \mathbf{F}_1^{(2)}(w, \varepsilon) + \varepsilon^4 \mathbf{F}_0^{(2)}(w, \varepsilon) (\mathbf{Q} - \mathbf{\Lambda}) = \mathbf{0}, \\ & j(\sigma + \varepsilon^2(1 - e^{-j\varepsilon w})) \left[\varepsilon \mathbf{F}_1^{(2)'}(w, \varepsilon) + j(\lambda - 2\mu) \mathbf{F}_1^{(2)}(w, \varepsilon) \right] \\ & - j\sigma e^{-j\varepsilon w} \left[\varepsilon^3 \mathbf{F}_0^{(2)'}(w, \varepsilon) + j(\lambda - 2\mu) \varepsilon^2 \mathbf{F}_0^{(2)}(w, \varepsilon) \right] \\ & + \varepsilon^2 \mathbf{F}_1^{(2)}(w, \varepsilon) (\mathbf{Q} - \mathbf{\Lambda}) - \mu \varepsilon^2 \mathbf{F}_1^{(2)}(w, \varepsilon) + \varepsilon^4 \mathbf{F}_0^{(2)}(w, \varepsilon) \mathbf{\Lambda} \\ & + 2\mu \mathbf{F}_2^{(2)}(w, \varepsilon) = \mathbf{0}, \\ & j\varepsilon^2(1 - e^{-j\varepsilon w}) \sum_{k=0}^2 \left(\varepsilon^{3-2k} \mathbf{F}_k^{(2)'}(w, \varepsilon) + j(\lambda - 2\mu) \varepsilon^{2-2k} \mathbf{F}_k^{(2)}(w, \varepsilon) \right) \\ & + j\sigma(1 - e^{-j\varepsilon w}) \sum_{k=0}^1 \left(\varepsilon^{3-2k} \mathbf{F}_k^{(2)'}(w, \varepsilon) + j(\lambda - 2\mu) \varepsilon^{2-2k} \mathbf{F}_k^{(2)}(w, \varepsilon) \right) \\ & + \sum_{k=0}^2 \varepsilon^{4-2k} \mathbf{F}_k^{(2)}(w, \varepsilon) \mathbf{Q} - (1 - e^{j\varepsilon w}) \mathbf{F}_2^{(2)}(w, \varepsilon) \mathbf{\Lambda} = \mathbf{0}. \end{aligned} \right. \tag{14}$$

In multiplying the equations in the (14) by the unit column-vector \mathbf{e} we can write

$$\left\{ \begin{aligned} & j(\sigma + \varepsilon^2(1 - e^{-j\varepsilon w})) \left[\varepsilon^3 \mathbf{F}_0^{(2)'}(w, \varepsilon) \mathbf{e} + j(\lambda - 2\mu) \varepsilon^2 \mathbf{F}_0^{(2)}(w, \varepsilon) \right] \mathbf{e} \\ & + \mu \varepsilon^2 \mathbf{F}_1^{(2)}(w, \varepsilon) \mathbf{e} + \varepsilon^4 \mathbf{F}_0^{(2)}(w, \varepsilon) (\mathbf{Q} - \mathbf{\Lambda}) \mathbf{e} = \mathbf{0}, \\ & j(\sigma + \varepsilon^2(1 - e^{-j\varepsilon w})) \left[\varepsilon \mathbf{F}_1^{(2)'}(w, \varepsilon) \mathbf{e} + j(\lambda - 2\mu) \mathbf{F}_1^{(2)}(w, \varepsilon) \right] \mathbf{e} \\ & - j\sigma e^{-j\varepsilon w} \left[\varepsilon^3 \mathbf{F}_0^{(2)'}(w, \varepsilon) \mathbf{e} + j(\lambda - 2\mu) \varepsilon^2 \mathbf{F}_0^{(2)}(w, \varepsilon) \right] \mathbf{e} \\ & + \varepsilon^2 \mathbf{F}_1^{(2)}(w, \varepsilon) (\mathbf{Q} - \mathbf{\Lambda}) \mathbf{e} - \mu \varepsilon^2 \mathbf{F}_1^{(2)}(w, \varepsilon) \mathbf{e} + \varepsilon^4 \mathbf{F}_0^{(2)}(w, \varepsilon) \mathbf{\Lambda} \mathbf{e} \\ & + 2\mu \mathbf{F}_2^{(2)}(w, \varepsilon) \mathbf{e} = \mathbf{0}, \\ & j\varepsilon^2(1 - e^{-j\varepsilon w}) \sum_{k=0}^2 \left(\varepsilon^{3-2k} \mathbf{F}_k^{(2)'}(w, \varepsilon) + j(\lambda - 2\mu) \varepsilon^{2-2k} \mathbf{F}_k^{(2)}(w, \varepsilon) \right) \mathbf{e} \\ & + j\sigma(1 - e^{-j\varepsilon w}) \sum_{k=0}^1 \left(\varepsilon^{3-2k} \mathbf{F}_k^{(2)'}(w, \varepsilon) + j(\lambda - 2\mu) \varepsilon^{2-2k} \mathbf{F}_k^{(2)}(w, \varepsilon) \right) \mathbf{e} \\ & + \sum_{k=0}^2 \varepsilon^{4-2k} \mathbf{F}_k^{(2)}(w, \varepsilon) \mathbf{Q} \mathbf{e} - (1 - e^{j\varepsilon w}) \mathbf{F}_2^{(2)}(w, \varepsilon) \mathbf{\Lambda} \mathbf{e} = \mathbf{0}. \end{aligned} \right. \tag{15}$$

Let divide each equation of the system (15) by the ε to the minimum power and then we can obtain (16) by a limiting process $\varepsilon \rightarrow 0$ in (15) with $e^{\pm j\varepsilon w} = 1 \pm j\varepsilon w + o(\varepsilon^2)$

$$\left\{ \begin{aligned} & \sigma(\lambda - 2\mu) \mathbf{F}_0^{(2)}(w) - \mu \mathbf{F}_1^{(2)}(w) = \mathbf{0}, \\ & \sigma(\lambda - 2\mu) \mathbf{F}_1^{(2)}(w) - 2\mu \mathbf{F}_2^{(2)}(w) = \mathbf{0}, \\ & \sigma(\lambda - 2\mu) \mathbf{F}_1^{(2)}(w) \mathbf{e} - (\lambda - 2\mu) \mathbf{F}_2^{(2)}(w) \mathbf{e} + \mathbf{F}_2^{(2)}(w) \mathbf{\Lambda} \mathbf{e} = \mathbf{0}, \end{aligned} \right. \tag{16}$$

where $\mathbf{F}_n^{(2)}(w) = \lim_{\varepsilon \rightarrow 0} \mathbf{F}_n^{(2)}(w, \varepsilon)$, $n = 0, 1, 2$.

The solving of equations system (15) has the following form

$$\mathbf{F}_n^{(2)}(w, \varepsilon) = \mathbf{F}_n^{(2)}(w) + j\varepsilon w \mathbf{f}_n(w) + o(\varepsilon^2), \quad n = 0, 1, 2. \tag{17}$$

Using (16) and (17), $\mathbf{Q} \mathbf{e} = \mathbf{0}$ and $\mathbf{F}_n^{(2)'}(w, \varepsilon) = \mathbf{F}_n^{(2)'}(w) + j\varepsilon \mathbf{f}_n(w) + j\varepsilon w \mathbf{f}'_n(w)$, $n = 0, 1, 2$, we can write (15) as

$$\left\{ \begin{aligned} & \sigma \mathbf{F}_0^{(2)'}(w) + j(\lambda - 2\mu) w \sigma \mathbf{f}_0(w) + \mu w \mathbf{f}_1(w) = \mathbf{0}, \\ & \sigma \mathbf{F}_1^{(2)'}(w) + j(\lambda - 2\mu) w \sigma \mathbf{f}_1(w) + 2\mu w \mathbf{f}_2(w) = \mathbf{0}, \\ & \left(\mathbf{F}_2^{(2)'}(w) - \lambda w \mathbf{f}_2(w) \right) \mathbf{e} + w \mathbf{F}_2^{(2)}(w) \mathbf{\Lambda} \mathbf{e} + w \mathbf{f}_2(w) \mathbf{\Lambda} \mathbf{e} = \mathbf{0}. \end{aligned} \right. \tag{18}$$

In adding the second equation by the third equation of the (18) we obtain the equation for $\mathbf{F}_2^{(2)}(w)$ as follows

$$\mathbf{F}_2^{(2)'}(w) \mathbf{e} + w \mathbf{F}_2^{(2)}(w) \mathbf{\Lambda} \mathbf{e} + w \mathbf{f}_2(w) (\mathbf{\Lambda} - \lambda \mathbf{I}) \mathbf{e} = \mathbf{0},$$

and it is easy to get the solution of the equation as $\mathbf{F}_2^{(2)}(w) = R_2 \exp \{-\lambda w^2/2\} \mathbf{r}$ in suggestion that $\mathbf{F}_2^{(2)}(w) = \mathbf{r} \Phi(w)$, $\mathbf{f}_2(w) = \mathbf{r} \phi(w)$, where $\lambda = \mathbf{r} \mathbf{\Lambda} \mathbf{e}$, and R_2 is defined above.

Pre-limit characteristic function $h(u)$ is approximately equal to

$$\begin{aligned}
 h(u) &= \{\mathbf{H}_0(u) + \mathbf{H}_1(u) + \mathbf{H}_2(u)\} \mathbf{e} = R_2 \exp \left\{ (\lambda - 2\mu) \frac{ju}{\alpha} \right\} \mathbf{F}_2^{(2)} \left(\frac{u}{\varepsilon} \right) \mathbf{e} + o(\varepsilon) \\
 &\approx R_2 \exp \left\{ (\lambda - 2\mu) \frac{ju}{\alpha} \right\} \mathbf{F}_2^{(2)} \left(\frac{u}{\varepsilon} \right) \mathbf{e}.
 \end{aligned}$$

So, the second-order asymptotic characteristic function $h^{(2)}(u)$ of the probability distribution of the number of calls in the orbit under the system heavy load and long time patience of calls in the orbit condition can be presented as

$$\begin{aligned}
 h^{(2)}(u) &= R_2 \exp \left\{ (\lambda - 2\mu) \frac{ju}{\alpha} \right\} \mathbf{F}_2^{(2)} \left(\frac{u}{\varepsilon} \right) \mathbf{e} \\
 &= R_2^2 \exp \left\{ (\lambda - 2\mu) \frac{ju}{\alpha} + \frac{\lambda (ju)^2}{\alpha \cdot 2} \right\}.
 \end{aligned} \tag{19}$$

The Theorem 1 is proved.

5 Numerical Results

In this section, some numerical examples are presented. It demonstrate the applicability area of the asymptotic results depending on parameters of the Retrial queueing system of MMPP/M/2 type with impatient customer in the orbit.

So, we compare asymptotic and exact distributions for different values of parameters λ and α using the Kolmogorov distance between respective cumulative distribution functions

$$\Delta = \max_{0 \leq i < \infty} \left| \sum_{\nu=0}^i D_\nu - \sum_{\nu=0}^i P_\nu \right|$$

where D_ν and P_ν are an exact and an asymptotic probability distributions respectively.

Let the system parameters be

$$\mathbf{\Lambda} = \begin{pmatrix} 5 & 0 & 5 \\ 0 & 7.5 & 0 \\ 0 & 0 & 2.5 \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} -4 & 3 & 1 \\ 2 & -6 & 4 \\ 5 & 3 & -8 \end{pmatrix}, \quad \mu = 1, \quad \sigma = 1, \tag{20}$$

then parameter $\lambda = \mathbf{r}\mathbf{\Lambda}\mathbf{e} = 5.297$ and values of the Kolmogorov distance for that example is presented in Table 1.

In Fig. 2 there are examples of comparison of the asymptotic and the exact distribution densities.

Let the system parameters be

$$\mathbf{\Lambda} = \begin{pmatrix} 10 & 0 & 5 \\ 0 & 15 & 0 \\ 0 & 0 & 5 \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} -4 & 3 & 1 \\ 2 & -6 & 4 \\ 5 & 3 & -8 \end{pmatrix}, \quad \mu = 1, \quad \sigma = 1, \tag{21}$$

Table 1. Kolmogorov distances between asymptotic and exact distributions under given parameters (20)

	$\alpha = 2$	$\alpha = 1$	$\alpha = 0.5$	$\alpha = 0.1$
$\lambda = 5.297$	0.082	0.025	0.021	0.013

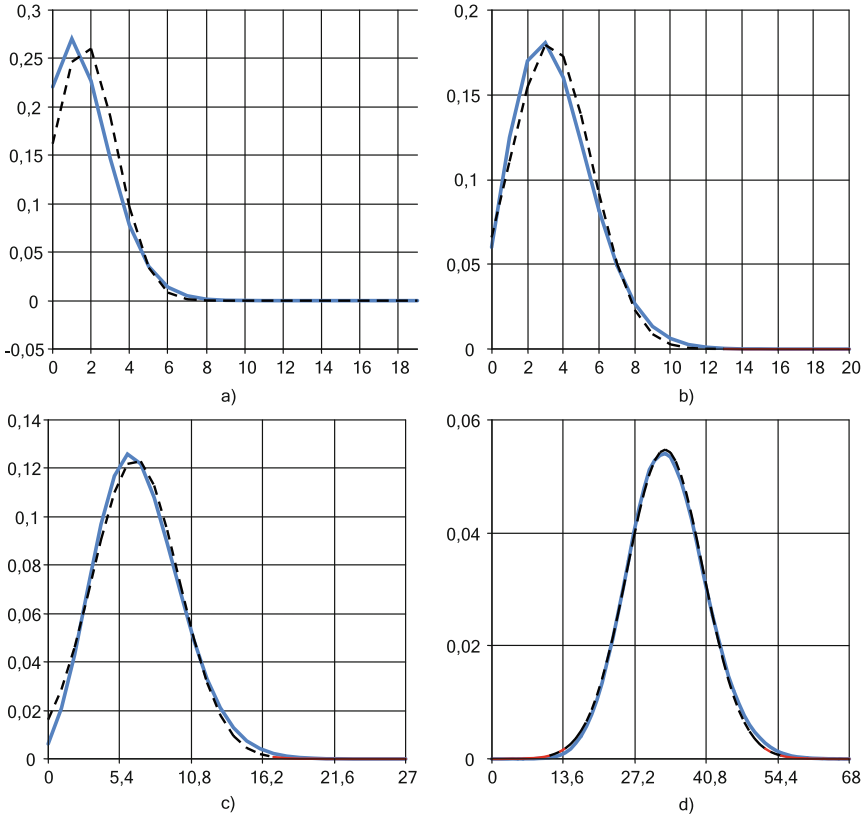


Fig. 2. Comparisons of the asymptotic (dashed line) and the exact (solid line) probability densities when (a) $\alpha = 2$, (b) $\alpha = 1$, (c) $\alpha = 0.5$, (d) $\alpha = 0.1$

Table 2. Kolmogorov distances between asymptotic and exact distributions under given parameters (21)

	$\alpha = 2$	$\alpha = 1$	$\alpha = 0.5$	$\alpha = 0.1$
$\lambda = 10.575$	0.040	0.019	0.016	0.017

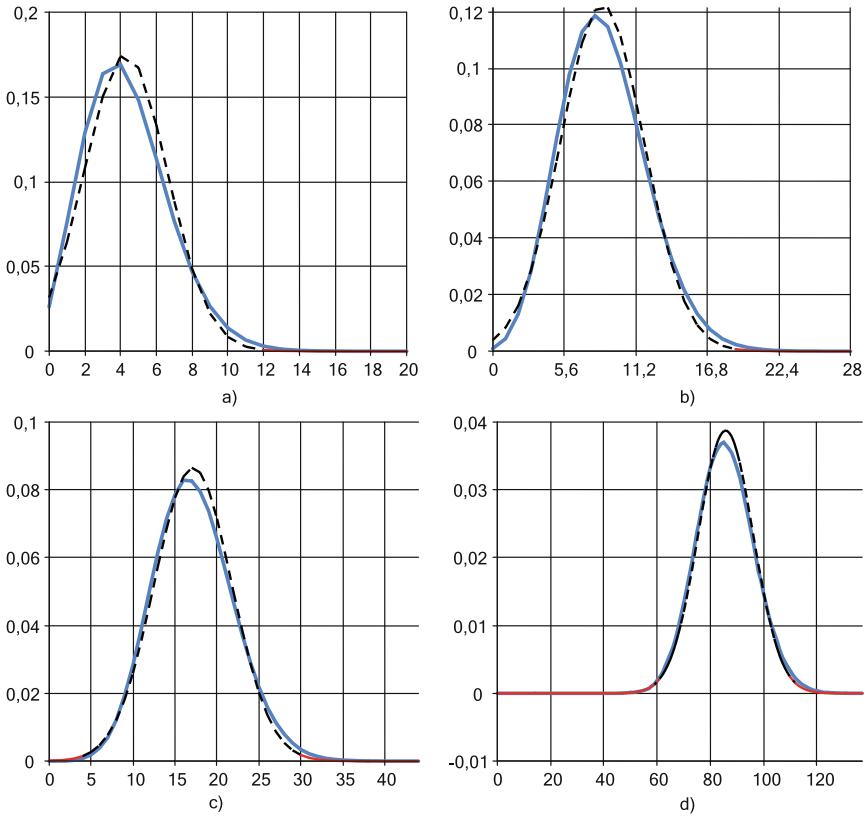


Fig. 3. Comparisons of the asymptotic (dashed line) and the exact (solid line) probability densities when (a) $\alpha = 2$, (b) $\alpha = 1$, (c) $\alpha = 0.5$, (d) $\alpha = 0.1$

then parameter $\lambda = \mathbf{r}\mathbf{\Lambda}\mathbf{e} = 10.575$ and values of the Kolmogorov distance for that example is presented in Table 2.

If we suppose the Kolmogorov distance equal to 0.05 and less as acceptable accuracy of a result, we can find parameters values in which the approximation (14) can be applied. Figures 2 and 3 show that increasing of the parameter λ when parameter α is fixed leads to reduction of the Kolmogorov distances between asymptotic and exact distributions, and decreasing of the parameter α when parameter λ is fixed leads to reduction of the Kolmogorov distances between asymptotic and exact distributions.

6 Conclusion

In the present paper, two servers retrial queueing system of MMPP/M/2 type with impatient customer in the orbit is considered. It is proved that the probability distribution of the calls number in the orbit can be approximated by the

Gaussian distribution under the system heavy load and long time patience of calls in the orbit condition.

Numerical results that allow to draw a conclusion about an applicability area of the asymptotic result is the purpose of the future studies.

References

1. Wilkinson, R.I.: Theories for toll traffic engineering in the USA. *Bell Syst. Tech. J.* **35**(2), 421–507 (1956)
2. Cohen, J.W.: Basic problems of telephone traffic and the influence of repeated calls. *Philips Telecommun. Rev.* **18**(2), 49–100 (1957)
3. Gosztony, G.: Repeated call attempts and their effect on traffic engineering. *Budavox Telecommun. Rev.* **2**, 16–26 (1976)
4. Elldin, A., Lind, G.: *Elementary Telephone Traffic Theory*. Ericsson Public Telecommunications, Stockholm (1971)
5. Artalejo, J.R., Gomez-Corral, A.: *Retrial Queueing Systems. A Computational Approach*. Springer, Heidelberg (2008). <https://doi.org/10.1007/978-3-540-78725-9>
6. Falin, G.I., Templeton, J.G.C.: *Retrial Queues*. Chapman & Hall, London (1997)
7. Artalejo, J.R., Falin, G.I.: Standard and retrial queueing systems: a comparative analysis. *Rev. Mat. Complut.* **15**, 101–129 (2002)
8. Roszik, J., Sztrik, J., Kim, C.: Retrial queues in the performance modelling of cellular mobile networks using MOSEL. *Int. J. Simul.* **6**, 38–47 (2005)
9. Aguir, S., Karaesmen, F., Askin, O.Z., Chauvet, F.: The impact of retrials on call center performance. *OR Spektrum* **26**, 353–376 (2004)
10. Nazarov, A., Sztrik, J., Kvach, A.: Comparative analysis of methods of residual and elapsed service time in the study of the closed retrial queueing system $M/GI/1//N$ with collision of the customers and unreliable server. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) *ITMM 2017. CCIS*, vol. 800, pp. 97–110. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_8
11. Dudin, A., Deepak, T.G., Joshua, V.C., Krishnamoorthy, A., Vishnevsky, V.: On a $BMAP/G/1$ retrial system with two types of search of customers from the orbit. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) *ITMM 2017. CCIS*, vol. 800, pp. 1–12. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_1
12. Dudin, A.N., Klimenok, V.I.: Queueing system $BMAP/G/1$ with repeated calls. *Math. Comput. Model.* **30**(3–4), 115–128 (1999)
13. Yang, T., Posner, M., Templeton, J.: The $M/G/1$ retrial queue with non-persistent customers. *Queueing Syst.* **7**(2), 209–218 (1990)
14. Krishnamoorthy, A., Deepak, T.G., Joshua, V.C.: An $M/G/1$ retrial queue with non-persistent customers and orbital search. *Stoch. Anal. Appl.* **23**, 975–997 (2005)
15. Kim, J.: Retrial queueing system with collision and impatience. *Commun. Korean Math. Soc.* **4**, 647–653 (2010)
16. Martin, M., Artalejo, J.: Analysis of an $M/G/1$ queue with two types of impatient units. *Adv. Appl. Probab.* **27**, 647–653 (1995)
17. Kumar, M., Arumuganathan, R.: Performance analysis of single server retrial queue with general retrial time, impatient subscribers, two phases of service and Bernoulli schedule. *Tamkang J. Sci. Eng.* **13**(2), 135–143 (2010)
18. Fedorova, E., Voytikov, K.: Retrial queue $M/G/1$ with impatient calls under heavy load condition. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) *ITMM 2017. CCIS*, vol. 800, pp. 347–357. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_28

19. Bérczes, T., Sztrik, J., Tóth, Á., Nazarov, A.: Performance modeling of finite-source retrial queueing systems with collisions and non-reliable server using MOSEL. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) DCCN 2017. CCIS, vol. 700, pp. 248–258. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66836-9_21
20. Artalejo, J.R., Pozo, M.: Numerical calculation of the stationary distribution of the main multiserver retrial queue. *Ann. Oper. Res.* **116**, 41–56 (2002)
21. Neuts, M.F., Rao, B.M.: Numerical investigation of a multiserver retrial model. *Queueing Syst.* **7**(2), 169–189 (1990)
22. Klimenok, V.I., Orlovsky, D.S., Dudin, A.N.: BMAP/PH/N system with impatient repeated calls. *Asia Pac. J. Oper. Res.* **24**(3), 293–312 (2007)
23. Kim, C.S., Klimenok, V., Dudin, A.: Retrial queueing system with correlated input, finite buffer, and impatient customers. In: Dudin, A., De Turck, K. (eds.) ASMTA 2013. LNCS, vol. 7984, pp. 262–276. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39408-9_19
24. Dudin, A., Klimenok, V.: Retrial queue of BMAP/PH/N type with customers balking, impatience and non-persistence. In: 2013 Conference on Future Internet Communications (CFIC), pp. 1–6. IEEE (2013)
25. Fedorova, E.: The second order asymptotic analysis under heavy load condition for retrial queueing system MMPP/M/1. In: Dudin, A., Nazarov, A., Yakupov, R. (eds.) ITMM 2015. CCIS, vol. 564, pp. 344–357. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25861-4_29
26. Borovkov, A.A.: *Asymptotic Methods in Queueing Theory*. Wiley, New York (1984)

Author Index

- Babu, Dhanya 360
Bérczes, Tamás 236
Bui, Duy T. 348
- Chakravarthy, Srinivas 143
Covington, R. 195
- Dammer, Diana 71
Danilyuk, Elena 387
Dragieva, Velika 263
Dudin, Alexander 117
- Enakoutsa, K. 195, 305
- Fokina, Nadezhda P. 184
- Joshua, Varghese C. 39, 360
- Kerobyan, K. 195, 305
Kerobyan, R. 195, 305
Kirpichnikov, Alexander 225
Kitaeva, Anna 248
Klimenok, Valentina 117
Krishna Kumar, B. 333
Krishnamoorthy, Achyutha 39, 55, 360
Kuki, Attila 236
Kvach, Anna 1, 172
- Lebedev, Eugene 27
Lisovskaya, Ekaterina 129
Livinska, Hanna 27
Livshits, Klimentii 248
- Mathew, Ambily P. 39
Medvedev, Gennady 16
Melikov, Agassi 55
Mikheev, Pavel 274
Moiseev, Alexander 321
Moiseeva, Svetlana 129, 321, 387
Morozov, Evsey 143
- Nazarov, Anatoly 1, 71, 172
Nezhelskaya, Luydmila 93, 157
- Osipova, Marina 106
- Pagano, Michele 129
Paul, Svetlana 213
Phung-Duc, Tuan 213
Pichugina, Anastasiya 274
- Remnev, Stanislav 143
Rukmani, R. 333
Rumyantsev, Alexander 143
Ryzhikov, Yury I. 83
- Salih, Haroun H. 372
Sankar, R. 333
Semenova, Olga V. 348
Shahmaliyev, Mammad 55
Shklennik, Maria 321
Shumchenia, Uladzimir 117
Sidorova, Ekaterina 157
Suschenko, Sergey 274
Sztrik, János 1, 172, 236
- Tananko, Igor E. 184
Titovtsev, Anton 225
Tóth, Ádám 236
Tsitsiashvili, Gurami 106
Tulubaev, D. A. 290
Tumashkina, Diana 93
- Ulyanova, Ekaterina 248
Urusova, Dina A. 372
- Vasilyev, Sergey A. 372
Vygovskaya, Olga 387
- Yakimov, Igor 225
- Zadorozhnyi, V. N. 290
Zakharenkova, T. R. 290