Alexander Dudin · Anatoly Nazarov · Alexander Moiseev (Eds.)

Communications in Computer and Information Science 2163

Information Technologies and Mathematical Modelling

Queueing Theory and Applications

22nd International Conference, ITMM 2023
and 14th International Workshop, WRQ 2023
Tomsk, Russia, December 4–9, 2023
Revised Selected Papers

Springer

Communications in Computer and Information Science 2163

Rationale

The CCIS series is devoted to the publication of proceedings of computer science conferences. Its aim is to efficiently disseminate original research results in informatics in printed and electronic form. While the focus is on publication of peer-reviewed full papers presenting mature work, inclusion of reviewed short papers reporting on work in progress is welcome, too. Besides globally relevant meetings with internationally representative program committees guaranteeing a strict peer-reviewing and paper selection process, conferences run by societies or of high regional or national relevance are also considered for publication.

Topics

The topical scope of CCIS spans the entire spectrum of informatics ranging from foundational topics in the theory of computing to information and communications science and technology and a broad variety of interdisciplinary application fields.

Information for Volume Editors and Authors

Publication in CCIS is free of charge. No royalties are paid, however, we offer registered conference participants temporary free access to the online version of the conference proceedings on SpringerLink (http://link.springer.com) by means of an http referrer from the conference website and/or a number of complimentary printed copies, as specified in the official acceptance email of the event.

CCIS proceedings can be published in time for distribution at conferences or as post-proceedings, and delivered in the form of printed books and/or electronically as USBs and/or e-content licenses for accessing proceedings at SpringerLink. Furthermore, CCIS proceedings are included in the CCIS electronic book series hosted in the SpringerLink digital library at http://link.springer.com/bookseries/7899 Conferences publishing in CCIS are allowed to use Online Conference Service (OCS) for managing the whole proceedings lifecycle (from submission and reviewing to preparing for publication) free of charge.

Publication process

The language of publication is exclusively English. Authors publishing in CCIS have to sign the Springer CCIS copyright transfer form, however, they are free to use their material published in CCIS for substantially changed, more elaborate subsequent publications elsewhere. For the preparation of the camera-ready papers/files, authors have to strictly adhere to the Springer CCIS Authors' Instructions and are strongly encouraged to use the CCIS LaTeX style files or templates.

Abstracting/Indexing

CCIS is abstracted/indexed in DBLP, Google Scholar, EI-Compendex, Mathematical Reviews, SCImago, Scopus. CCIS volumes are also submitted for the inclusion in ISI Proceedings.

How to start

To start the evaluation of your proposal for inclusion in the CCIS series, please send an e-mail to ccis@springer.com.

Alexander Dudin · Anatoly Nazarov ·
Alexander Moiseev
Editors

Information Technologies and Mathematical Modelling

Queueing Theory and Applications

22nd International Conference, ITMM 2023
and 14th International Workshop, WRQ 2023
Tomsk, Russia, December 4–9, 2023
Revised Selected Papers

Springer

Editors
Alexander Dudin
Belarusian State University
Minsk, Belarus

Anatoly Nazarov
Tomsk State University
Tomsk, Russia

Alexander Moiseev
Tomsk State University
Tomsk, Russia

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-031-65384-1 ISBN 978-3-031-65385-8 (eBook)
https://doi.org/10.1007/978-3-031-65385-8

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, expressed 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

If disposing of this product, please recycle the paper.

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 *Communications in Computer and Information Science* since 2014. The conference series was named after 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 conference has 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 ones at the conference. These sections gather many scientists from different countries. Many foreign participants come to this Siberian conference every year because of our warm welcome and serious scientific discussions. In 2023, the conference was held in Tomsk together with the 14th International Workshop on Retrial Queues and Related Topics (WRQ). This workshop is aimed at a specific area of queueing theory. The conference was organized by the National Research Tomsk State University of Russia, Peoples' Friendship University of Russia (RUDN University), Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, and Karshi State University of Uzbekistan.

This volume presents selected papers from the 22th ITMM conference. The conference received 96 submissions, from which 23 were selected to be published in the current collection. Papers have passed single-blind peer review and each of them had at least three reviewers.

The papers are devoted to new results in queueing theory and its applications, and also related areas of probabilistic analysis. 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.

December 2023

Alexander Dudin
Anatoly Nazarov
Alexander Moiseev

Organization

International Program Committee Chairs

Alexander Dudin	Belarusian State University, Belarus
Anatoly Nazarov	Tomsk State University, Russia
Alexander Moiseev	Tomsk State University, Russia
Svetlana Moiseeva	Tomsk State University, Russia

International Program Committee

Abdurahim Abdushukurov	Romanovsky Institute of Mathematics of the Academy of Sciences of the Republic of Uzbekistan, Uzbekistan
Ivan Atencia	University of Málaga, Spain
Shavkat Ayupov	Uzbekistan Academy of Sciences, Uzbekistan
Abdulla Azamov	Romanovsky Institute of Mathematics of the Academy of Sciences of the Republic of Uzbekistan, Uzbekistan
Srinivas Chakravarthy	Kettering University, USA
Rui Dinis	Universidade Nova de Lisboa, Portugal
Dmitry Efrosinin	Johannes Kepler University Linz, Austria
Mais Farhadov	Institute of Control Sciences, Russian Academy of Sciences, Russia
Shakir Formanov	Romanovsky Institute of Mathematics of the Academy of Sciences of the Republic of Uzbekistan, Uzbekistan
Yulia Gaydamaka	Peoples' Friendship University of Russia (RUDN University), Russia
Erol Gelenbe	Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Poland
Tareq Hamadneh	Al-Zaytoonah University, Jordan
Azam Imomov	Karshi State University, Uzbekistan
Bara Kim	Korea University, South Korea
Che Soong Kim	Sangji University, South Korea
Udo Krieger	Universität Bamberg, Germany
B. Krishna Kumar	Anna University, India
Achyutha Krishnamoorthy	Cochin University of Science and Technology, India

Quan-Lin Li	Yan Shan University, China
Yury Malinkovsky	Francisk Skorina Gomel State University, Belarus
Natalia Markovich	Institute of Control Sciences, Russian Academy of Sciences, Russia
Agassi Melikov	National Aviation Academy of Azerbaijan, Azerbaijan
Paulo Montezuma-Carvalho	Universidade Nova de Lisboa, Portugal
Evsey Morozov	Institute of Applied Mathematical Research, Karelian Research Centre of Russian Academy of Sciences, Russia
Rein Nobel	Vrije Universiteit Amsterdam, The Netherlands
Michele Pagano	University of Pisa, Italy
Svetlana Paul	Tomsk State University, Russia
Tuan Phung-Duc	University of Tsukuba, Japan
Gul'nara Raimova	Romanovsky Institute of Mathematics of the Academy of Sciences of the Republic of Uzbekistan, Uzbekistan
Jacques Resing	Eindhoven University of Technology, The Netherlands
Svetlana Rozhkova	Tomsk Polytechnical University, Russia
Vladimir Rykov	Gubkin Russian State University of Oil and Gas, Russia
Konstantin Samouylov	Peoples' Friendship University of Russia (RUDN University), Russia
Daria Semenova	Siberian Federal University, Russia
Stanislav Shidlovskiy	Tomsk State University, Russia
Sergey Suschenko	Tomsk State University, Russia
János Sztrik	University of Debrecen, Hungary
Henk Tijms	Vrije Universiteit Amsterdam, The Netherlands
Gurami Tsitsiashvili	Institute of Applied Mathematics, Far Eastern Branch of Russian Academy of Sciences, Russia
Vladimir Vishnevsky	Institute of Control Sciences, Russian Academy of Sciences, Russia
Anton Voitishek	Institute of Computational Mathematics and Mathematical Geophysics, Siberian Branch, Russian Academy of Sciences, Russia
Alexander Zamyatin	Tomsk State University, Russia
Amdjed Zraiqat	Al-Zaytoonah University, Jordan

Local Organizing Committee

Ekaterina Fedorova (Chair)
Ivan Lapatin (Co-chair)
Valentina Broner
Elena Danilyuk
Yana Izmailova
Irina Kochetkova
Andrey Larionov
Ekaterina Lisovskaya
Olga Lizyura
Anna Morozova
Ekaterina Pankratova
Svetlana Paul
Daria Semenova
Eduard Sopin
Radmir Salimzyanov
Dmitry Shashev
Maria Shklennik
Alexey Shkurkin
Konstantin Voytikov
Lyubov Zadiranova

Contents

Information Technologies and Mathematical Modeling

Workshop on Retrial Queues

Information Technologies and Mathematical Modeling

Analysis of a k-out-of-n Reliability System with Single Server, Internal and External Service, N-Policy, and Multiple Server Vacations

Binumon Joseph[1] and K. P. Jose[2(✉)]

[1] Government Engineering College Idukki, Painavu, Idukki 685603, Kerala, India
[2] PG and Research Department of Mathematics, St. Peter's College, Kolenchery 682311, Kerala, India
kpjspc@gmail.com

Abstract. This study involves the evaluation of the reliability of a k-out-of-n repairable system with a single server responsible for repairing failed components. The server provides service to external customers in addition to servicing failed components in the system. To optimise revenue from external services without compromising system reliability, we introduce the N-policy. The repair of internal failed components starts only with the accumulation of N failed components. Once the service of internal components is started, all failed internal components are repaired one by one. As soon as the system is free of failed internal and external components, the server takes multiple vacations. Whenever the number of internal failed components reaches N, if the server is on vacation, it is interrupted, and the server immediately serves the internal component. Meanwhile, the server is in an external service, and then the service is preempted to serve the internal components. The failure times of the components of the system follow an exponential distribution, and external customers arrive according to the Poisson process. The service times of both internal and external customers follow an exponential distribution. The vacation time follows an exponential distribution. By using the Matrix Analytic Method, we discuss system stability and steady-state distribution. The N-policy level is optimised numerically using a suitable cost function.

Keywords: k-out-of-n system · Multiple vacation · N-Policy · Matrix-Analytic Method

1 Introduction

In order to maintain system functionality in the event of a failure, redundancy is provided by multiple identical components that are connected in a way that

The authors acknowledge the financial support provided by FIST Program, Department of Science and Technology, Government of India, to the PG & Research Department of Mathematics through SR/FST/College-2018- XA 276(C).

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 3–18, 2024.
https://doi.org/10.1007/978-3-031-65385-8_1

allows them to share the increased load resulting from fewer operating components. Redundancy is also a very economical way to raise the system's reliability level. k-out-of-n reliability systems, which require at least k of n components to be operational for the system to function, are a common type of redundancy. They have been widely studied in the context of system optimisation and reliability computation. Chakravarthy et al. [2] analysed a system with N machines having an exponential failure rate serviced by an unreliable server. The service time of a failed machine and the repair time of the server follow a phase-type distribution. Chakravarthy et al. [3] studied a k-out-of-n system with an unreliable server that takes multiple vacations under (N,T)policy.

In a world of intense competition, businesses place a great priority on serving both internal and external clients. The increased revenue obtained from outside services is one of the key goals of this. Attending more diversified services, may also be assumed to increase the server's level of experience. Krishnamoorthy et al. [8] studied a k-out-of-n system extending service to external customers with MAP arrival. The service of both failed components of the system and external customers follows phase-type distributions. Dudin et al. [4] analysed a k-out-of-n system with utilization of idle time by providing service to external customers. If the server is busy, the external customers with BMAP arrival directed to an orbit. Krishnamoorthy et al. [9] analysed the reliability of a k-out-of-n system serving external customers, and obtained various performance measures using the Matrix geometric method. Switching of server between internal and external customers is controlled by N-policy. By providing vacations to a heterogeneous multi-server system, Jose and Beena [7] effectively use the idle time in a production inventory system. Beena and Jose [1] analyse a production inventory system with multiple servers under multiple vacations. The arrival of customers is constituted by the Markovian Arrival Process(MAP). Wu et al. [13] discussed a single-vacation, k-out-of-n:G repairable system whose vacation and repair times are distributed according to general distributions. By employing the supplementary variable technique, a range of reliability metrics, including availability, failure rate, and mean time to first failure of the system, are obtained in steady-state.

Jain and Jain [6] analysed a machine repair problem with multiserver and asynchronous vacation for servers. The servers are subject to breakdown. In the analysis of a standby system with a single repairman, Yang et al. [14] introduced a working vacation, in which the repairman offers service at a lower rate during vacation. Wang et al. [12] analysed a repairable system with non-identical components under phase-type distributed multiple vacations of a single server. Liu et al. [10] introduced a mixed redundancy strategy in the reliability analysis of a multistate system under phase-type distributed multiple vacations of a single server. Eryilmaz [5] investigated a k-out-of-n system with multiple types of components having nonidentical failure distributions, and obtained the number of failed components present in the system at a time. This paper analyses a k-out-of-n system in which a single server serves external customers in addition to internal failed components. The server takes a vacation when the system is free from failed components. An N-policy is introduced to maintain system reliability by balancing the main and external services.

The rest of the paper is organised as follows: The mathematical model is defined and analysed in Sect. 2. Section 3 deduces the stability conditions and steady-state probability vector of the systems. In Sect. 4, we derive some essential system performance measures. Section 5 discussed the numerical study of the model. The impact of N-policy and external customer service on system reliability is investigated. A cost function is discussed for determining the value of N.

2 Mathematical Modelling and Analysis of the Problem

We consider a k-out-of-n system in which all components work well initially. The components of the system are subject to failure. When i components are operational, their lifetimes are independent and exponentially distributed random variables with parameter λ_1/i. Hence, λ_1 failures occur on average per unit time when i components are working, and the failure rate of internal components follows an exponential distribution with parameter λ_1. The server offers service to the failed components from outside during idle time. This will improve the server's performance by gaining more experience from various external service situations. Also, this will improve the revenue without compromising the reliability of the system. The arrival of failed components from outside the system also follows an exponential distribution with parameter λ_2. Although the system offers service to external components to generate additional income, we have an N policy for the service of the failed components to ensure system reliability. That means whenever the number of failed components of the system reaches N, the service to the external components is preempted and the server repairs all the N system components one by one. To ensure the proper working of the server, a vacation is taken after servicing N internal failed components. The vacation time follows an exponential distribution with the parameter ν.

The server takes a vacation after service when the system is free from failed outside and internal customers. After completing one vacation, the server searches for failed components, and if there are no failed units in the external components, and the number of failed system components is less than N, then the server takes another vacation. Also, if the server is on vacation, whenever the number of failed components of the system reaches N, the vacation is interrupted, and the server immediately services all the N internal failed components one by one. When the server is busy with internal components, the external components do not join the system for service. Otherwise, the external components join a queue of infinite length. The service times of internal and external components are exponentially distributed with parameters μ_1 and μ_2 respectively.

Let $N_1(t)$ be the number of external failed components in the system, $N_2(t)$ be the number of internal failed components and $S(t)$ the status of the server at time t.

$$S(t) = \begin{cases} 0, & \text{if the server is in vacation,} \\ 1, & \text{if the server services the internal components,} \\ 2, & \text{if the server services the external components.} \end{cases}$$

Then $\{X(t), t \geq 0\}$ where $X(t) = (N_1(t), S(t), N_2(t))$ is a continuous time Markov chain with the state space $\{(i,0,j)/i \geq 0, 0 \leq j \leq N-1\} \cup \{(i,1,j)/i \geq 0, 1 \leq j \leq n-k+1\} \cup \{(i,2,j)/i \geq 1, 0 \leq j \leq N-1\}$

In sequel, we use the following notations:

1. I_n - n^{th} order identity matrix.
2. E_k - k^{th} order square matrix defined as

$$E_k(i,j) = \begin{cases} 1, & \text{if } j = i+1; 1 \leq i \leq k-1, \\ -1, & \text{if } j = i; 1 \leq i \leq k, \\ 0, & \text{otherwise.} \end{cases}$$

3. E'_k - transpose of E_k.
4. $r_k(i)$ - 1×k order row matrix with i^{th} element is 1 and all elements are zeros.
5. $c_k(i)$ - transpose of $r_k(i)$.
6. $\mathbf{e}$ - a column matrix of appropriate order.
7. $\otimes$ - Kronecker product of matrices.

The block tridiagonal infinitesimal generator matrix of $\{X(t), t \geq 0\}$ is

$$Q = \begin{pmatrix} B_1 & B_0 & & & \\ B_2 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \text{ where}$$

$$B_1 = \begin{pmatrix} B_{11} & B_{12} \\ B_{13} & B_{14} \end{pmatrix},$$

$$B_{11} = \lambda_1 E_N - \lambda_2 I_N,$$

$$B_{12} = \lambda_1 \Big(r_{n-k+1}(N) \otimes C_N(N) \Big),$$

$$B_{13} = \mu_1 \Big(r_N(1) \otimes C_{n-k+1}(1) \Big),$$

$$B_{14} = \lambda_1 E_{n-k+1} + \mu_1 E'_{n-k+1} + \lambda_1 \Big(r_{n-k+1}(n-k+1) \otimes C_{n-k+1}(n-k+1) \Big),$$

$$B_0 = \begin{pmatrix} \lambda_2 I_N & 0_{N\times(n-k+1)} & 0_{N\times N} \\ 0_{(n-k+1)\times N} & 0_{(n-k+1)\times(n-k+1)} & 0_{(n-k+1)\times N} \end{pmatrix},$$

$$B_2 = \begin{pmatrix} 0_{N\times N} & 0_{N\times(n-k+1)} \\ 0_{(n-k+1)\times N} & 0_{(n-k+1)\times(n-k+1)} \\ \mu_2 I_N & 0_{N\times(n-k+1)} \end{pmatrix},$$

$$A_1 = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{14} & A_{15} & A_{16} \\ A_{17} & A_{18} & A_{19} \end{pmatrix},$$

$$A_{11} = \lambda_1 E_N - \big(\lambda_2 + \nu\big) I_N,$$

$$A_{12} = \lambda_1 \Big(r_{n-k+1}(N) \otimes C_N(N)\Big),$$

$$A_{13} = \nu I_N,$$

$$A_{14} = \mu_1 \Big(r_N(1) \otimes C_{n-k+1}(1)\Big),$$

$$A_{15} = \lambda_1 E_{n-k+1} + \mu_1 E'_{n-k+1} + \lambda_1 \Big(r_{n-k+1}(n-k+1) \otimes C_{n-k+1}(n-k+1)\Big),$$

$$A_{16} = 0_{(n-k+1)\times N},$$

$$A_{17} = 0_{N\times N},$$

$$A_{18} = \lambda_1 \Big(r_{n-k+1}(N) \otimes C_N(N)\Big),$$

$$A_{19} = \lambda_1 E_N - \big(\lambda_2 + \mu_2\big) I_N,$$

$$A_0 = \begin{pmatrix} \lambda_2 I_N & 0_{N\times(n-k+1)} & 0_{N\times N} \\ 0_{(n-k+1)\times N} & 0_{(n-k+1)\times(n-k+1)} & 0_{(n-k+1)\times N} \\ 0_{N\times N} & 0_{N\times(n-k+1)} & \lambda_2 I_N \end{pmatrix},$$

$$A_2 = \begin{pmatrix} 0_{N\times N} & 0_{N\times(n-k+1)} & 0_{N\times N} \\ 0_{(n-k+1)\times N} & 0_{(n-k+1)\times(n-k+1)} & 0_{(n-k+1)\times N} \\ 0_{N\times N} & 0_{N\times(n-k+1)} & \mu_2 I_N \end{pmatrix}.$$

3 Stability Condition

Theorem 1. *The steady state probablity vector* $\boldsymbol{\Pi} = (\Pi_0, \Pi_1, \Pi_2) = \Big((\pi_{(0,0)}, \pi_{(0,1)}, \pi_{(0,2)} \ldots, \pi_{(0,N-1)}\big), \big(\pi_{(1,1)},\ \pi_{(1,2)},\ \pi_{(1,3)} \ldots, \pi_{(1,N)}, \ldots, \pi_{(1,n-k+1)}\big), \big(\pi_{(2,0)}, \pi_{(2,1)},\ \pi_{(2,2)} \ldots, \pi_{(2,N-1)}\big)\Big)$ *corresponding to the generator matrix* $A = A_2 + A_1 + A_0$ *is given by*

$$\pi_{(0,0)} = \frac{\left(\frac{\mu_1-\lambda_1}{\lambda_1+\nu}\right)}{\left[N\left(\frac{\lambda_1}{\mu_1}\right) - \left(\left(\frac{\lambda_1}{\mu_1}\right)^{n-k+1-N} - \left(\frac{\lambda_1}{\mu_1}\right)^{n-k+1}\right)\left(\frac{\lambda_1}{\mu_1-\lambda_1}\right)\right]},$$

$$\pi_{(0,i)} = \left(\frac{\lambda_1}{\lambda_1+\nu}\right)^i \pi_{(0,0)}, \qquad i = 1,2,3,\ldots,N-1,$$

$$\pi_{(1,i)} = \begin{cases} \left[1-\left(\frac{\lambda_1}{\mu_1}\right)^i\right]\left(\frac{\lambda_1+\nu}{\mu_1-\lambda_1}\right)\pi_{(0,0)}, & i = 1,2,3,\ldots,N, \\ \left(\frac{\lambda_1}{\mu_1}\right)^{i-N}\left[1-\left(\frac{\lambda_1}{\mu_1}\right)^N\right]\left(\frac{\lambda_1+\nu}{\mu_1-\lambda_1}\right)\pi_{(0,0)}, & \\ \qquad i = N+1,\ N+2,\ldots,n-k+1, & \end{cases} \tag{1}$$

$$\pi_{(2,i)} = \left[1-\left(\frac{\lambda_1}{\mu_1}\right)^{i+1}\right]\left(\frac{\lambda_1+\nu}{\lambda_1}\right)\pi_{(0,0)}, \qquad i = 0,1,2,\ldots,N.$$

Proof. The generator matrix $A = A_2 + A_1 + A_0 =$

$$\begin{pmatrix} \lambda_1 E_N - \nu I_N & \lambda_1\Big(r_{n-k+1}(N)\otimes C_N(N)\Big) & \nu I_N \\ \mu_1\Big(r_N(1)\otimes C_{n-k+1}(1)\Big) & \begin{matrix}\lambda_1 E_{n-k+1} + \mu_1 E'_{n-k+1} + \\ \lambda_1\Big(r_{n-k+1}(n-k+1)\otimes \\ C_{n-k+1}(n-k+1)\Big)\end{matrix} & 0_{(n-k+1)\times N} \\ 0_{N\times N} & \lambda_1\Big(r_{n-k+1}(N)\otimes C_N(N)\Big) & \lambda_1 E_N \end{pmatrix}.$$

Let the steady state vector of A be $\Pi = (\Pi_0, \Pi_1, \Pi_2) = \Big((\pi_{(0,0)}, \pi_{(0,1)}, \pi_{(0,2)} \ldots, \pi_{(0,N-1)}\Big), \Big(\pi_{(1,1)}, \pi_{(1,2)}, \pi_{(1,3)} \ldots, \pi_{(1,N)}, \ldots, \pi_{(1,n-k+1)}\Big), \Big(\pi_{(2,0)}, \pi_{(2,1)}, \pi_{(2,2)} \ldots, \pi_{(2,N-1)}\Big)\Big)$. Then

$$\Pi A = 0 \text{ and } \Pi\ \mathbf{e} = 1,$$

gives

$$\begin{aligned}
\pi_{(0,0)}(-\lambda_1-\nu)+\pi_{(1,1)}\mu_1&=0,\\
\pi_{(0,i-1)}\lambda_1+\pi_{(0,i)}(-\lambda_1-\nu)&=0,\\
\text{for } i=1,2,3\ldots,N-1,&\\
\pi_{(1,1)}(-\lambda_1-\mu_1)+\pi_{(1,2)}\mu_1&=0,\\
\pi_{(1,i-1)}\lambda_1+\pi_{(1,i)}(-\lambda_1-\mu_1)+\pi_{(1,i+1)}\mu_1&=0,\\
\text{for } i=2,3\ldots,N-1,&\\
\pi_{(0,N-1)}\lambda_1+\pi_{(1,N-1)}\lambda_1\pi_{(1,N)}(-\lambda_1-\mu_1)+\pi_{(1,N+1)}\mu_1+\pi_{(2,N-1)}\lambda_1&=0,\\
\pi_{(1,i-1)}\lambda_1+\pi_{(1,i)}(-\lambda_1-\mu_1)+\pi_{(1,i+1)}\mu_1&=0,\\
\text{for } i=N+1,N+2,\ldots,n-k,&\\
\pi_{(1,n-k)}\lambda_1+\pi_{(1,n-k+1)}(-\mu_1)&=0,\\
\pi_{(0,0)}\nu+\pi_{(2,0)}(-\lambda_1)&=0,\\
\pi_{(0,i)}\nu+\pi_{(2,i-1)}(-\lambda_1)+\pi_{(2,i)}(-\lambda_1)&=0,\\
\text{for } i=1,2,3,\ldots,N-1.&
\end{aligned} \tag{2}$$

From Eq. (2)

$$\pi_{(0,i)}=\left(\frac{\lambda_1}{\lambda_1+\nu}\right)^i\pi_{(0,0)}, \qquad i=1,2,3,\ldots,N-1,$$

$$\pi_{(1,i)}=\begin{cases}\left[1-\left(\dfrac{\lambda_1}{\mu_1}\right)^i\right]\left(\dfrac{\lambda_1+\nu}{\mu_1-\lambda_1}\right)\pi_{(0,0)}, & i=1,2,3,\ldots,N,\\[2ex] \left(\dfrac{\lambda_1}{\mu_1}\right)^{i-N}\left[1-\left(\dfrac{\lambda_1}{\mu_1}\right)^N\right]\left(\dfrac{\lambda_1+\nu}{\mu_1-\lambda_1}\right)\pi_{(0,0)}, & \\ \qquad i=N+1,N+2,\ldots,n-k+1, & \end{cases} \tag{3}$$

$$\pi_{(2,i)}=\left[1-\left(\frac{\lambda_1}{\mu_1}\right)^{i+1}\right]\left(\frac{\lambda_1+\nu}{\lambda_1}\right)\pi_{(0,0)}, \qquad i=0,1,2,\ldots,N.$$

Using normalization $\mathit{\Pi}\mathbf{e}=1$, gives

$$\pi_{(0,0)}=\frac{\left(\dfrac{\mu_1-\lambda_1}{\lambda_1+\nu}\right)}{N\left(\dfrac{\lambda_1}{\mu_1}\right)-\left[\left(\dfrac{\lambda_1}{\mu_1}\right)^{n-k+1-N}-\left(\dfrac{\lambda_1}{\mu_1}\right)^{n-k+1}\right]\left(\dfrac{\lambda_1}{\mu_1-\lambda_1}\right)}. \tag{4}$$

Theorem 2. *The system is stable if and only if*

$$\frac{\mu_2}{\nu}\left[1-\left(\frac{\lambda_1}{\lambda_1+\mu_1}\right)^N\right]<N\left(\frac{\mu_2-\lambda_1}{\lambda_1}\right). \tag{5}$$

Proof. The above Markov chain is stable if and only if $\Pi A_0 e < \Pi A_2 e$, which gives

$$\lambda_2 \left[\sum_{j=0}^{N-1} \pi_{(0,j)} + \sum_{j=0}^{N-1} \pi_{(2,j)} \right] < \mu_2 \sum_{j=0}^{N-1} \pi_{(2,j)}.$$

Using equations (3) and (4), the Markov chain under consideration is stable if and only if

$$\frac{\mu_2}{\nu} \left[1 - \left(\frac{\lambda_1}{\lambda_1 + \mu_1} \right)^N \right] < N \left(\frac{\mu_2 - \lambda_1}{\lambda_1} \right).$$

4 Steady State Probability Vector

The Markov process $\{X(t),\ t \geq 0\}$ is a level-independent QBD process. The stationery distribution when it exists, has a matrix geometric solution. Let $\mathbf{x} = (x_0, x_1, x_2, \dots)$ be the probability steady state vector of Q, the generator matrix of the process. Then $\mathbf{x}$ satisfies the equations $\mathbf{x}Q = 0$ and the normalizing condition $\mathbf{xe} = 1$. Here $\mathbf{e}$ represents the column matrix of 1's with infinite order. Then

$$x_{i+1} = x_i \ R \ \forall \ i \geq 1,$$

where R is the minimal nonnegative solution of the matrix equation $A_0 + A_1 R + A_2 R^2 = 0$. The boundary probability vectors $(x_0,\ x_1)$ are obtained from the equations

$$x_0 B_0 + x_1 B_2 = 0,$$
$$x_0 B_1 + x_1 (R A_2 + A_1) = 0.$$

Using normalization, $x_0 \mathbf{e} + x_1 (I - R)^{-1} \mathbf{e} = 1$, one can solve the equations for x_0 and x_1.

5 System Performance Measures

Theorem 3. *Expected server busy period of the server with internal failed components, that start with an arbitrary number of external customers is* $E_B = E(T) \left[\sum_{i=0}^{\infty} x(i, 0, N-1) + \sum_{i=1}^{\infty} x(i, 2, N-1) \right]$.

Proof. The server busy period with the internal failed components begins with the accumulation of N failed components and ends when all components are served. Let $T(i), i \geq 0$ be the server busy period with internal components, which begins with i external customers are in the system. But the busy period with internal components is not dependes on the number of customers from outside. Hence, $T(i) = T, \forall i \geq 0$. Consider a Markov chain $X_B(t)$, which denotes the number of failed internal components of the system with $\{0, 1, 2, \dots, N, N+1, \dots, n-k+1\}$ as the state space. The infinitesimal generator of the Markov chain is

$\bar{Q} = \begin{pmatrix} 0 & 0 \\ -\bar{B}\mathbf{e} & \bar{B} \end{pmatrix}$, where

$\bar{B} = \lambda_1 E_{n-k+1} + \mu_1 E'_{n-k+1} + \lambda_1 \Big(r_{n-k+1}(n-k+1) \otimes C_{n-k+1}(n-k+1) \Big)$.

Then 0 is the absorbing state and T is time until absorption time. Thus T follows a phase type distribution $PH(\alpha, \bar{B})$ with the initial probability vector $\alpha = (0, 0, 0, \ldots, 1, \ldots, 0)$ having 1 as the N^{th} element and all other elements are zeros. Then the expectation of the server busy period is $E(T) = -\alpha \bar{B}^{-1}\mathbf{e}$.

The expected value E_B, of the server busy period in internal service, when the service begins with any arbitrary number of failed external components is

$$E_B = E(T) \left[\sum_{i=0}^{\infty} x(i, 0, N-1) + \sum_{i=1}^{\infty} x(i, 2, N-1) \right].$$

The important system performance measures are given below.

1. Portion of time the system was down $P_F = \sum_{i=0}^{\infty} x(i, 1, n-k+1)$.
2. Reliability of the system $P_R = 1 - P_F$.
3. The average number of outside units in the queue $N_Q = \sum_{i=1}^{\infty} i \sum_{j=1}^{n-k+1} x(i, 1, j) + \sum_{i=0}^{\infty} i \sum_{j=0}^{N-1} x(i, 0, j) + \sum_{i=2}^{\infty} (i-1) \sum_{j=0}^{N-1} x(i, 2, j)$.
4. The average number of failed main components $N_{IF} = \sum_{j=0}^{N-1} j \sum_{i=0}^{\infty} x(i, 0, j) + \sum_{j=0}^{n-k+1} j \sum_{i=0}^{\infty} x(i, 1, j) + \sum_{j=0}^{N-1} j \sum_{i=1}^{\infty} x(i, 2, j)$.
5. Fraction of time the server in an external service $P_{EB} = \sum_{i=1}^{\infty} \sum_{j=0}^{N-1} x(i, 2, j)$.
6. Probability that the server was found on vacation $P_v = \sum_{i=0}^{\infty} \sum_{j=0}^{N-1} x(i, 0, j)$.
7. Expected rate of external customer loss $E_{EL} = \lambda_2 \sum_{i=0}^{\infty} \sum_{j=1}^{n-k+1} x(i, 1, j)$.
8. Average number of external customers waiting while the server is on vacation $EE_v = \sum_{i=0}^{\infty} i \sum_{j=0}^{N-1} x(i, 0, j)$.
9. Average number of internal components waiting while the server is on vacation $EI_v = \sum_{j=0}^{N-1} j \sum_{i=0}^{\infty} x(i, 0, j)$.

6 Numerical Analysis

The numerical experiments that were carried out to look at how different parameter adjustments affected the performance measures are described in this part.

6.1 Various Performance Measures for Different Values of N

In Table 1, various performance measures are listed for different values of the policy level N. The second column corresponding to P_F displays the values of the probability that the system will fail. As expected, the value of P_F increases with an increase in N. The column corresponding to N_Q expresses the variation in the expected number of outside units waiting for service in the queue. For lower values of N, the server spends more time serving the main components, and hence the number of outside units waiting for service increases. The fourth column represents the change in P_{EB}, the fraction of time the server is busy with external customers, corresponding to different values of N. As N increases, P_{EB} also slightly increases. The expected loss rate of external customers, E_{EL}, decreases with an increase in N. Also, the average number of internal failed components, N_{IF}, increases with an increase in N. The fraction of time the

Table 1. Perfomance measures and the N-policy level. $\lambda_1 = 5, \lambda_2 = 3, \mu_1 = 6, \mu_2 = 5, n = 50, k = 20, \nu = 6$.

N	P_F	N_Q	P_{EB}	E_{EL}	N_{IF}	P_v	P_R	EE_v	EI_v
2	0.00064569	10.2670	0.10032	2.4984	5.3983	0.066882	0.99935	0.596040	0.024477
3	0.00071245	4.7809	0.10036	2.4982	5.8896	0.066904	0.99929	0.231160	0.052057
4	0.00078830	3.5184	0.10039	2.4980	6.3798	0.066929	0.99921	0.148040	0.081701
5	0.00087462	2.9727	0.10044	2.4978	6.8688	0.066958	0.99913	0.112380	0.112760
6	0.00097304	2.6755	0.10049	2.4976	7.3565	0.066991	0.99903	0.093032	0.144810
7	0.00108540	2.4922	0.10054	2.4973	7.8426	0.067028	0.99891	0.081088	0.177540
8	0.00121400	2.3698	0.10061	2.4970	8.3270	0.067071	0.99879	0.073074	0.210760
9	0.00136140	2.2833	0.10068	2.4966	8.8094	0.067120	0.99864	0.067370	0.244340
10	0.00153050	2.2195	0.10077	2.4962	9.2896	0.067177	0.99847	0.063126	0.278200
11	0.00172490	2.1710	0.10086	2.4957	9.7671	0.067242	0.99828	0.059859	0.312280
12	0.00194870	2.1329	0.10097	2.4951	10.2420	0.067316	0.99805	0.057276	0.346580
13	0.00220680	2.1025	0.10110	2.4945	10.7130	0.067402	0.99779	0.055190	0.381070
14	0.00250500	2.0777	0.10125	2.4937	11.1810	0.067502	0.99750	0.053477	0.415760
15	0.00284990	2.0572	0.10142	2.4929	11.6440	0.067617	0.99715	0.052053	0.450690
16	0.00324960	2.0399	0.10162	2.4919	12.1020	0.067750	0.99675	0.050858	0.485880
17	0.00371360	2.0253	0.10186	2.4907	12.5540	0.067905	0.99629	0.049850	0.521370
18	0.00425310	2.0128	0.10213	2.4894	13.0000	0.068084	0.99575	0.048998	0.557230
19	0.00488150	2.0019	0.10244	2.4878	13.4390	0.068294	0.99512	0.048279	0.593520
20	0.00561490	1.9924	0.10281	2.4860	13.8680	0.068538	0.99439	0.047676	0.630340
21	0.00647250	1.9839	0.10324	2.4838	14.2880	0.068824	0.99353	0.047179	0.667810
22	0.00747770	1.9764	0.10374	2.4813	14.6960	0.069159	0.99252	0.046780	0.706060
23	0.00865850	1.9697	0.10433	2.4784	15.0910	0.069553	0.99134	0.046474	0.745260
24	0.01004900	1.9636	0.10502	2.4749	15.4700	0.070016	0.98995	0.046260	0.785650
25	0.01169200	1.9579	0.10585	2.4708	15.8310	0.070564	0.98831	0.046140	0.827470

system takes on vacation, P_v, increases with an increase in N, while the reliability of the system, P_R, shows a slight decrease with an increase in N.

6.2 N-Policy and the Number of External Failed Units in the Queue

As per the N policy, as the value of N increases, the server gets more time to attend to the service of external customers. This will reduce the number of external components waiting for service in the queue. First three columns of the Table 2 lists N_Q, the expected number of external customers in the queue for different values of the vacation parameter ν. The table also shows the relationship between the vacation parameter and the expected number of external

Table 2. Expected number of external customers in the queue and the N-policy level. $\lambda_2 = 3, \mu_1 = 6, \mu_2 = 5, n = 50, k = 20$.

		N_Q			N_Q	
N	ν=5	ν=6	ν=7	$\lambda_1 = 4$	$\lambda_1 = 5$	$\lambda_1 = 6$
2	29.318	10.267	6.6436	5.8715	10.267	28.488
3	6.7555	4.7809	3.8601	3.6159	4.7809	6.6043
4	4.4057	3.5184	3.0312	2.8942	3.5184	4.3228
5	3.5287	2.9727	2.6453	2.5509	2.9727	3.4701
6	3.0796	2.6755	2.4279	2.3557	2.6755	3.0326
7	2.8117	2.4922	2.2911	2.2323	2.4922	2.7710
8	2.6365	2.3698	2.1984	2.1486	2.3698	2.5995
9	2.5146	2.2833	2.1323	2.0887	2.2833	2.4799
10	2.4258	2.2195	2.0830	2.0441	2.2195	2.3925
11	2.3589	2.1710	2.0453	2.0098	2.1710	2.3264
12	2.3070	2.1329	2.0155	1.9828	2.1329	2.2749
13	2.2657	2.1025	1.9915	1.9611	2.1025	2.2339
14	2.2324	2.0777	1.9719	1.9433	2.0777	2.2007
15	2.2049	2.0572	1.9556	1.9285	2.0572	2.1732
16	2.1820	2.0399	1.9418	1.9160	2.0399	2.1503
17	2.1627	2.0253	1.9301	1.9054	2.0253	2.1308
18	2.1462	2.0128	1.9200	1.8963	2.0128	2.1141
19	2.1320	2.0019	1.9112	1.8885	2.0019	2.0997
20	2.1196	1.9924	1.9034	1.8816	1.9924	2.0872
21	2.1087	1.9839	1.8966	1.8755	1.9839	2.0762
22	2.0990	1.9764	1.8904	1.8702	1.9764	2.0664
23	2.0904	1.9697	1.8849	1.8654	1.9697	2.0577
24	2.0826	1.9636	1.8798	1.8610	1.9636	2.0499
25	2.0756	1.9579	1.8750	1.8570	1.9579	2.0429

customers in the queue. As the vacation parameter increases, the vacation duration decreases. The decrease in the vacation duration ensures more time for the service of external customers. The table illustrates that the increase in the vacation parameter results in a decrease in the expected number of external customers in the queue. The last three columns of the Table 2 lists N_Q for different values of the vacation parameter λ_1. This shows that as λ_1 increases, N_Q also increases.

6.3 N-Policy and the Reliability of the System

When N increased, we anticipated that the system's reliability would decrease. This is because we believe that a decrease in system reliability could result from the server spending more time servicing external customers as N increases. Table 3 provides support for this. As N increases, P_{EB}, the portion of time spent

Table 3. System realibility and the N-policy level. $\lambda_1 = 5, \lambda_2 = 3, \mu_1 = 6, \mu_2 = 5, \nu = 6, k = 20$.

		PR			P_{EB}	
N	n=45	n=50	n=55	n=45	n=50	n=55
2	0.99839	0.99935	0.99974	0.10081	0.10032	0.10013
3	0.99822	0.99929	0.99971	0.10089	0.10036	0.10014
4	0.99803	0.99921	0.99968	0.10099	0.10039	0.10016
5	0.99781	0.99913	0.99965	0.10110	0.10044	0.10018
6	0.99756	0.99903	0.99961	0.10122	0.10049	0.10019
7	0.99728	0.99891	0.99957	0.10136	0.10054	0.10022
8	0.99695	0.99879	0.99951	0.10152	0.10061	0.10024
9	0.99658	0.99864	0.99946	0.10171	0.10068	0.10027
10	0.99615	0.99847	0.99939	0.10193	0.10077	0.10031
11	0.99565	0.99828	0.99931	0.10217	0.10086	0.10034
12	0.99508	0.99805	0.99922	0.10246	0.10097	0.10039
13	0.99442	0.99779	0.99912	0.10279	0.10110	0.10044
14	0.99365	0.99750	0.99900	0.10318	0.10125	0.10050
15	0.99275	0.99715	0.99886	0.10362	0.10142	0.10057
16	0.99171	0.99675	0.99871	0.10414	0.10162	0.10065
17	0.99050	0.99629	0.99852	0.10475	0.10186	0.10074
18	0.98907	0.99575	0.99831	0.10546	0.10213	0.10084
19	0.98740	0.99512	0.99807	0.10630	0.10244	0.10097
20	0.98542	0.99439	0.99778	0.10729	0.10281	0.10111
21	0.98308	0.99353	0.99745	0.10846	0.10324	0.10128
22	0.98030	0.99252	0.99706	0.10985	0.10374	0.10147
23	0.97697	0.99134	0.99661	0.11151	0.10433	0.10170
24	0.97297	0.98995	0.99608	0.11351	0.10502	0.10196
25	0.96813	0.98831	0.99546	0.11593	0.10585	0.10227

with external customers, increases. As a result, the fraction of time spent serving the main customers gets reduced, resulting in a decline in the system's reliability. When the total number of system components n was high, the magnitude of the reliability diminished and was found to be less. Table 3 illustrates, in summary, that an increase in the N-policy level has little impact on system reliability. The k-out-of-n system's reliability increases as its total number of components does. The extent of the reliability decrease also decreases with increasing n. The reason for this is because n-k+1 increases as n increases, with k being constant. So, when N failed components accumulate, the server begins to handle them and spends more time on them, which maintains system reliability even as N increases.

6.4 Cost Function

Table 2 shows us that by raising N, we may give externally failed components more time to be attended. Table 3 shows that when N increases, the percentage of time the server spends serving external clients also rises. Still, when N increases, system reliability is slightly decreased. Finding out whether the N-policy level

Table 4. Cost variation corresponding to failure rate λ_1. $C_1 = 3000$, $C_2 = 2000$, $C_3 = 1000$, $C_4 = 2000$, $C_5 = 500$, $C_6 = 500$.

N	$\lambda_1 = 4$	$\lambda_1=5$	$\lambda_1=6$
2	23060	34269	76958
3	18798	23543	33313
4	17605	21263	28872
5	17168	20417	27287
6	17028	20067	26531
7	17031	19944	26125
8	17114	19942	25898
9	17244	20010	25772
10	17405	20124	25710
11	17586	20266	25687
12	17782	20428	25692
13	17988	20604	25716
14	18202	20790	25752
15	18422	20981	25798
16	18647	21178	25850
17	18875	21377	25906
18	19106	21577	25964
19	19339	21777	26024
20	19573	21976	26085
21	19808	22172	26144
22	20044	22366	26202
23	20279	22555	26258
24	20514	22738	26312
25	20746	22914	26362

has an optional value is therefore worthwhile. To do this, we create a suitable cost function.

Let C_1 denote the cost per unit time incurred if the system fails. Let C_2 represent the cost of holding each external customer in the queue for one unit of time; C_3 represent the cost of starting a failed system component service; C_4 represent the cost of losing one external customer; C_5 represent the cost of holding each failed internal component for one unit of time, and C_6 represent the cost per unit of time if the server is on vacation.

The expected cost/unit time $= C_1 P_F + C_2 N_Q + \dfrac{C_3}{E_B} + C_4 E_{EL} + C_5 N_{IF} + C_6 P_v$.

Table 4 examines how the cost function varies with N. We examine the cost function for various component failure rates. The table shows that, up to a certain point, the cost decreases as N increases and the policy level increases, but after that point, the cost increases as N increases. As a result, the cost curve has a concave shape and obtains an optimal value for N.

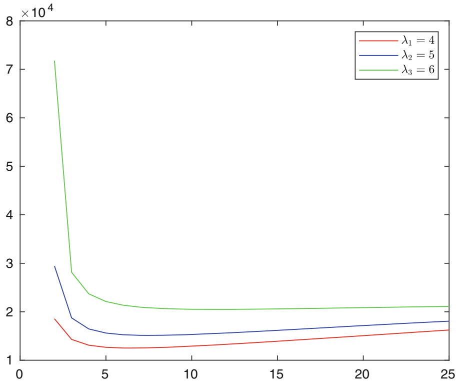


Fig. 1. Cost variation corresponding to failure rate λ_1. $C_1 = 3000$, $C_2 = 2000$, $C_3 = 1000$, $C_4 = 2000$, $C_5 = 500$, $C_6 = 500$.

7 Conclusion

An efficient way to make use of server idle time and increase system revenue would be to render services to external customers and undergo server vacation. But when a system requires a minimal number of components to function, the vacation duration and external services need to be handled cautiously to avoid negatively impacting the system's reliability. We have chosen to manage the external service using the N-Policy in this study. Specifically, we assume that the server only begins to respond to failed components of the system when N of them have failed. Then it provides service to external components, if any, during this idle time. If the system is free from failed external customers and the number of internal failed components is below N, the server takes a vacation. Also, after completing the service for N components of the system, the server takes a vacation. A continuous-time Markov chain has been used to model this situation. We also assume that the external service is preempted when N internal failed components are accumulated and the external arrivals are prevented from accessing the system when they find the server occupied with failed main system components. From the numerical analysis, we see that by implementing the N-policy and vacationing the server, we may maintain system reliability while optimising system cost obtained from a cost function by providing services to external clients. We intend to investigate in the future how an unreliable server affects the N-policy and the system's reliability. The study in this paper can also be extended by considering PH distributions for the service time.

References

1. Beena,P., Jose, K. P.: A MAP/PH (1), PH (2)/2 production inventory model with inventory dependent production rate and multiple servers. In: AIP Conference Proceedings, vol. 2261, no. 1. AIP Publishing (2020)
2. Chakravarthy, S.R., Agarwal, A.: Analysis of a machine repair problem with an unreliable server and phase type repairs and services. Naval Res. Logist. (NRL) **50**(5), 462–80 (2003)
3. Chakravarthy, S.R., Krishnamoorthy, A., Ushakumari, P.V.: A k-out-of-n reliability system with an unreliable server and phase type repairs and services: the (N, T) policy. J. Appl. Math. Stochast. Anal. **14**(4), 361–380 (2001)
4. Dudin, A.N., Krishnamoorthy, A., Narayanan, V.C.: Idle time utilization through service to customers in a retrial queue maintaining high system reliability. J. Math. Sci. **191**, 506–17 (2013)
5. Eryilmaz, S.: The number of failed components in a k-out-of-n system consisting of multiple types of components. Reliab. Eng. Syst. Saf. **175**, 246–250 (2018)
6. Jain, A., Jain, M.: Multi server machine repair problem with unreliable server and two types of spares under asynchronous vacation policy. Int. J. Math. Oper. Res. **10**(3), 286–315 (2017)
7. Jose, K.P., Beena, P.: On a retrial production inventory system with vacation and multiple servers. Int. J. Appl. Comput. Math. **6**(4), 108 (2020)
8. Krishnamoorthy, A., Narayanan, V.C., Deepak, T.G.: Reliability of a k-out-of-n system with repair by a service station attending a queue with postponed work. Int. J. Reliab. Qual. Saf. Eng. **14**(04), 379–98 (2007)
9. Krishnamoorthy, A., Sathian, M.K. : Reliability of a k-out-of-n system with repair by a single server extending service to external customers with pre-emption. Reliab. Theory Appl. **11**(2(41)), 61–93 (2016)
10. Liu, B., Wen, Y., Qiu, Q., Shi, H., Chen, J.: Reliability analysis for multi-state systems under k-mixed redundancy strategy considering switching failure. Reliab. Eng. Syst. Saf. **228**, 108814 (2022)
11. Neuts, M.F.: Matrix Geometric solutions in stochastic processes-An Algorithmic Approach, The John Hopkins University Press (1981)
12. Wang, G., Hu, L., Zhang, T., Wang, Y.: Reliability modeling for a repairable (k1, k2)-out-of-n: G system with phase-type vacation time. Appl. Math. Model. **91**, 311–21 (2021)

13. Wu, W., Tang, Y., Yu, M., Jiang, Y.: Reliability analysis of a k-out-of-n: G repairable system with single vacation. Appl. Math. Modelling **38**(24), 6075–6097 (2014)
14. Yang, D.Y., Tsao, C.L.: Reliability and availability analysis of standby systems with working vacations and retrial of failed components. Reliab. Eng. Syst. Saf. **182**, 46–55 (2019)

The Limit Distribution of the Queue Length in a Priority System with Autoregressive Arrivals Under the Heavy Traffic Condition

Alexey Bergovin[1](✉) and Vladimir Ushakov[1,2]

[1] Lomonosov Moscow State University, Moscow, Russia
alexey.bergovin@gmail.com

[2] The Institute of Informatics Problems of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia

Abstract. In this paper a one-line queueing system with two priority classes, relative priority, Poissonian input flow with random intensity and infinite number of places in queue for waiting is considered. The current intensity value is taken at the beginning of the time reckoned for the arrival of the next requirement. Successive values of the flow intensity form a Markov chain of a special kind. This input flow structure allows to take in consideration not only mathematical expectation and variance, but also correlation between interval of two next arrivals. The main result is the limit distribution of the queue length for the least priority class, it is obtained in an explicit form. Also, analytical expressions for the density function, mathematical expectation and variance are given. Numerical examples, which show difference among limit distributions (for different parameters) for studied cases are provided.

Keywords: Poissonian flow · random intensity · relative priority · queue length · heavy traffic

1 Introduction

The main goal of this paper is to study the behaviour of the queue length of the lowest priority class in the queueing system with the autoregressive input flow (will be fully described in Sect. 2). The relevance of this research is determined by the fact that the vast majority of real service systems in many applied areas operate under conditions of either the heavy load or close to the heavy load. For instance, the widespread distribution of communication networks has led to a sharp increase in the volume of network traffic, and, therefore, an increase in the load on these networks, most of which have become highly loaded. It also lead us to the relevance of considering exactly this structure of the incoming flow. In [2,3] the authors have shown that there is a stochastic dependence between an inter-arrival time of two adjacent requests.

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 19–32, 2024.
https://doi.org/10.1007/978-3-031-65385-8_2

The key purpose of this study is to find the limit distribution of the queue length of the lowest priority class. In addition, the probability density function, mathematical expectation and variance are given for the limit distribution. It should be mentioned that there is a huge amount of papers and monographs dedicated to studying queueing systems under the heavy traffic condition [4–16].

The paper has the following structure. In Sect. 2 the studied system is fully described. In Sect. 3, the results from [1], which will be used during the research, are given. Section 4 consists of two parts: in the first one some auxiliary expansions in a series are obtained, in the second one the main theorem is proved. In Sect. 5, some numerical examples are given.

2 System Definition

In this work sequence of queuing systems (the series scheme) is researched; m-th queueing system has the following structure. The structure of arrivals is time z_{m1} before the arrival of the first requirement and interval z_{mn} between $(n-1)$ - th and n-th requirement have an exponential distribution with random parameter $a_m^{(n)},\ n = 1, 2, \dots$. Value $a_m^{(n)}$ is selected just before the beginning of interval z_{mn} such that, $\mathbb{P}(a_m^{(1)} = a_{mj}) = c_{mj},\ a_{mi} \neq a_{mj},\ i \neq j,\ c_{mj} > 0,\ j = \overline{1, N},\ \sum\limits_{j=1}^{N} c_{mj} = 1$ and $a_m^{(n)} = \xi_m \cdot a_m^{(n-1)} + (1-\xi_m) \cdot b_m^{(n)}$, where $b_m^{(n)}, n = 1, 2, \dots,$ $a_m^{(n)}, n = 1, 2, \dots$, and ξ_m are independent random variables. The distribution of the random variables $a_m^{(n)}$ and $b_m^{(n)}$ coincides with the distribution of $a_m^{(1)}$, $n = 1, 2, \dots$, and ξ_m has Bernoulli distribution with parameter p_m.

It is easy to show that

$$\mathbb{P}(z_{mn} < t) = \sum_{j=1}^{N} c_{mj}(1 - e^{-a_{mj}t})$$

$$\mathbb{P}(z_{mn} < t_1, z_{m,n+k} < t_2) = (1 - p_m^k) \sum_{j=1}^{N} c_{mj}(1 - e^{-a_{mj}t_1}) \sum_{k=1}^{N} c_{mk}(1 - e^{-a_{mj}t_2}) + \\ + p_m^k \sum_{k=1}^{N} c_{mk}(1 - e^{-a_{mj}t_1})(1 - e^{-a_{mj}t_2})$$

$$\mathbb{E}z_{mn} = \sum_{j=1}^{N} \frac{c_{mj}}{a_{mj}}, \quad \mathbb{D}z_{mn} = \sum_{j=1}^{N} \frac{c_{mj}}{a_{mj}^2}, \quad corr(z_{mn}, z_{m,n+k}) = \frac{p_m^k}{2} \left(1 - \frac{(\mathbb{E}z_{mn})^2}{\mathbb{D}z_{mn}}\right) \tag{1}$$

Further m will be used in indexes only where it is necessary to highlight dependence on m.

There are two special cases: $p = 0$ and $p = 1$, in the first one the input flow is hyper-exponential, in the second one: we obtain a system such that the initial intensity is randomly selected from the set $\{a_1, \dots, a_N\}$ with probabilities

$c_1, \dots, c_N$ respectively, and afterward acts as a system with a Poissonian input flow with the chosen intensity (this case is not considered in this work).

From (1), it follows that if $\mu > 0$ and $\sigma > \mu$, there exists second-order flow ($N = 2$) such that it belongs to the class of flows considered in this work, and its expectation and variance of intervals between arrivals are equal to μ and σ^2, respectively. The correlation coefficient is equal to $\frac{p}{2}\left(1 - \left(\frac{\mu}{\sigma}\right)^2\right)$. While mathematical model of real system is being constructed, it is possible to adjust the first two moments of the real arrival process and their dependence.

All arriving requirements are divided into 2 classes with probabilities p_1, p_2 ($p_1 + p_2 = 1$), respectively, and it does not depend on other requirements. We firstly assume that each type of requirement forms its own queue. Secondly, if a service is started, it is never interrupted. The studied system operates under the relative priority discipline.

We assume that the system is free of requirements for $t = 0$ and serving lengths are independent random variables equally distributed for requirements of each particular type. The distribution function is $B_{mi}(x)$ and the density is $b_{mi}(x)$ for ith class and mth system , $i = 1, 2$; $\beta_{mi}(s)$ — Laplace-Stieltjes transform of function $b_{mi}(x), i = 1, 2$; β_{mij} – jth moment of random variable with $B_{mi}(x)$ distribution function.

$L(t) = (L_1(t), L_2(t))$ – amount of requirements in a system in time t.

It is known that if $\left(\sum\limits_{i=1}^{N} c_j a_j^{-1}\right)^{-1} \cdot (p_1\beta_{11} + p_2\beta_{21}) < 1$ the non-degenerate limit distribution of stochastic process $L(t)$ exists. In this work, $L_2(t)$ is studied in case while $t \to \infty$ and $\left(\sum\limits_{i=1}^{N} c_{mj} a_{mj}^{-1}\right)^{-1} \cdot (p_{m1}\beta_{m11} + p_{m2}\beta_{m21}) \to 1, m \to \infty$.

We will study the system under the next assumptions:

I) the first and the second moment of service time distribution exist (for each priority class) and

$$\beta_i(s) = 1 - \beta_{i1}s + \frac{\beta_{i2}}{2}s^2 + o_m(s^2), i = 1, 2,$$

where $o_m(s^2)/s^2 \to 0$ while $s \to 0$ uniformly on variable m

II) for each $m \in \{1, 2, \dots\}$: $a_m(p_1\beta_{m11} + p_2\beta_{m21}) < 1$

III) following limits exist $\lim c_i = c_i^*$, $\lim a_i = a_i^*$, $\lim \beta_{ij} = \beta_{ij}^*$, $\lim p_j = p_j^*$, $i = \overline{1, N}$, $j = 1, 2$, where lim denote $\lim\limits_{m\to\infty}$.

The main goal of this study is to find

$$\lim_{m\to\infty} \mathbb{P}\left(\rho^\gamma \cdot L_2\left(\frac{t}{\rho^\alpha}\right) < x\right)$$

where

$$\rho = 1 - a(p_1\beta_{11} + p_2\beta_{21}), \quad a = \left(\sum_{i=1}^{N} \frac{c_j}{a_j}\right)^{-1}, \quad \gamma = \begin{cases} 0.5\alpha, & \alpha \leqslant 2, \\ 1, & \alpha > 2. \end{cases}$$

3 Preliminaries

In [1] expressions which Laplace-Stieltjes transform of generating function of queue length is satisfied has been found. They will be used to find the limit distribution. Let us write them.

Lemma 1. *Equation*

$$(1-p)z \sum_{m=1}^{N} \frac{a_m c_m}{\mu(z) + a_m\,(1-pz)} = 1,$$

has N *continious in domain* $|z| \leqslant 1$ *solutions* $\mu = \mu_k(z),\ k = 1, \dots, N$, *that*:

1) *only one function* $\mu_k(z)$ *is equal to 0 while* $z = 1$;
2) $\Re(\mu_j(z)) < 0$ *for all* $j = 1, \dots, N$ *and* $|z| < 1$;
3) $\mu_i(z) \neq \mu_j(z)$ *while* $i \neq j$.

Denote $\alpha_k(z) = \prod\limits_{j \neq k} [\mu_k(z) - \mu_j(z)]$.

Lemma 2. *For each* $k = 1, \dots, N$ *system of equations*

$$z_1 = \beta_1(s - \mu_k(p_1 z_1 + p_2 z_2)),$$

$$z_2 = \beta_2(s - \mu_k(p_1 z_1 + p_2 z_2)),$$

has a unique solution $z_i = z_{ik}(s)$, *such, that* $|z_{ik}(s)| < 1$ *while* $k = 2, \dots, N$, $\Re s \geq 0$, *and* $z_{i1}(0) = 1$, $|z_{i1}(s)| < 1$ *while* $\Re s > 0$, $i = 1, 2$.

Lemma 3. *Laplace-Stieltjes transform of joint generating function of queue length for the first and the second classes is*

$$\begin{aligned}
p(z_1, z_2, s) &= p_0(s) + \\
&+ \frac{p_1 z_1 + p_2 z_2 - 1}{(1-p)(p_1 z_1 + p_2 z_2)} \times \sum_{k=1}^{N} \frac{1}{\mu_k(p_1 z_1 + p_2 z_2)(s - \mu_k(p_1 z_1 + p_2 z_2))} \times \\
&\times \Big[\gamma_1^{(k)}(z_1, z_2, s)[1 - \beta_1(s - \mu_k(p_1 z_1 + p_2 z_2))] + \\
&\qquad + \gamma_2^{(k)}(z_1, z_2, s)[1 - \beta_2(s - \mu_k(p_1 z_1 + p_2 z_2))] \Big],
\end{aligned}$$

where functions $\gamma_i^{(k)}(z_1, z_2, s),\ i = 1, 2,\ k = \overline{1, N}$ *satisfy*:

$$\gamma_1^{(k)}(z_1, z_2, s)\frac{z_1 - \beta_1(s - \mu_k(p_1 z_1 + p_2 z_2))}{z_1} +$$

$$+ \gamma_2^{(k)}(z_1, z_2, s)\frac{z_2 - \beta_2(s - \mu_k(p_1 z_1 + p_2 z_2))}{z_2} =$$

$$= \frac{(1-p)(p_1 z_1 + p_2 z_2)}{\alpha_k(p_1 z_1 + p_2 z_2)} \prod_{m=1}^{N} [\mu_k(p_1 z_1 + p_2 z_2) + a_m(1 - p(p_1 z_1 + p_2 z_2))] \times$$

$$\times \sum_{j=1}^{N} \frac{c_j a_j f_j(z_1, z_2, s)}{\mu_k(p_1 z_1 + p_2 z_2) + a_j(1 - p(p_1 z_1 + p_2 z_2))};$$

$$f_j(z_1, z_2, s) = 1 - (s + a_j(1 - p(p_1 z_1 + p_2 z_2))c_j^{-1} p_{0j}(s) +$$

$$+ (p_1 z_1 + p_2 z_2)(1-p) \sum_{k=1}^{N} a_k p_{0k}(s),\ j = \overline{1, N},$$

$$\gamma_1^{(k)}(z_1, z_2, s) = \frac{(1-p)(p_1 z_1 + p_2 z_2)}{\alpha_k(p_1 z_1 + p_2 z_2)} \prod_{j=1}^{N} [\mu_k(p_1 z_1 + p_2 z_2) + a_j(1 - p(p_1 z_1 + p_2 z_2))] \times$$

$$\times \sum_{e=1}^{N} \frac{a_e p_{1e}(z_1, z_2, 0, s)}{\mu_k(p_1 z_1 + p_2 z_2) + a_e(1 - p(p_1 z_1 + p_2 z_2))};$$

$$\gamma_2^{(k)}(z_1, z_2, s) = \frac{(1-p)(p_1 z_1 + p_2 z_2)}{\alpha_k(p_1 z_1 + p_2 z_2)} \prod_{j=1}^{N} [\mu_k(p_1 z_1 + p_2 z_2) + a_j(1 - p(p_1 z_1 + p_2 z_2))] \times$$

$$\times \sum_{e=1}^{N} \frac{a_e p_{2e}(z_2, 0, s)}{\mu_k(p_1 z_1 + p_2 z_2) + a_e(1 - p(p_1 z_1 + p_2 z_2))}.$$

functions $p_{0j}(s)$ *might be found from*:

$$p_{0j}(s) = \frac{1}{a_j} \sum_{l=1}^{N} \frac{1 - p(p_1 z_{l1}^* + p_2 z_{l2}^*)}{(1-p)(p_1 z_{l1}^* + p_2 z_{l2}^*)(s - \mu_l^*(s))} \cdot \frac{1}{\prod\limits_{n \neq j} (a_j - a_n)} \times$$

$$\times \left(\frac{\mu_l^*(s)}{1 - p(p_1 z_{l1}^* + p_2 z_{l2}^*)} + a_j \right)^{-1} \times$$

$$\times \prod_{l \neq n} \left(\frac{\mu_l^*(s)}{1 - p(p_1 z_{l1}^* + p_2 z_{l2}^*)} - \frac{\mu_n^*(s)}{1 - p(p_1 z_{n1}^* + p_2 z_{n2}^*)} \right)^{-1} \times$$

$$\times \prod_{k=1}^{N} \frac{(\mu_k^*(s) + a_j(1 - p(p_1 z_{k1}^* + p_2 z_{k2}^*))(\mu_l^*(s) + a_k(1 - p(p_1 z_{l1}^* + p_2 z_{l2}^*))}{(1 - p(p_1 z_{k1}^* + p_2 z_{k2}^*))(1 - p(p_1 z_{l1}^* + p_2 z_{l2}^*))}, \quad (2)$$

where $\mu_k^*(s) = \mu_k(p_1 z_{k1}^* + p_2 z_{k2}^*)$

4 Main Result

To prove the main theorem, some auxiliary expansions in a series are needed, which would be formulated as separate lemmas.

4.1 Auxiliary Expansions in a Series

Lemma 4. *The next asymptotics for* $z(s\rho^\alpha)$ *are true*:

$$z(s\rho^\alpha)-1=\begin{cases}-\sqrt{\dfrac{s\rho^\alpha}{a^2v}}+o(\rho^{\frac{\alpha}{2}}),\ \alpha<2,\\ \rho\cdot\dfrac{1-\sqrt{1+4sv}}{av}+o(\rho),\ \alpha=2,\\ -\dfrac{s\rho^{\alpha-1}}{a}+o(\rho^{\alpha-1}),\ \alpha>2,\end{cases}$$

where

$$v=\frac{a(p_1\beta_{12}+p_2\beta_{22})}{2}+\frac{1}{a(1-p)}\left(a^2\sum_{i=1}^{N}\frac{c_j}{a_j^2}-1\right),$$

$z(s)=p_1z_1(s)+p_2z_2(s)$ *is the solution of equation* $p_1z_1+p_2z_2=p_1\beta_1(s-\mu_1(p_1z_1+p_2z_2))+p_2\beta_2(s-\mu_1(p_1z_1+p_2z_2))$.

Proof. Using assumption I and Lemma 2 it is possible to write

$$z(s\rho^\alpha)=1-(s\rho^\alpha-\mu_1(z(s\rho^\alpha)))\cdot\beta_1+(s\rho^\alpha-\mu_1(z(s\rho^\alpha)))^2\cdot\frac{\beta_2}{2}+o((s\rho^\alpha-\mu_1(z(s\rho^\alpha)))^2),\quad(3)$$

where $\beta_i=p_1\beta_{1i}+p_2\beta_{2i},\ i=1,2.$

Also, we may write next expansion for function $\mu_1(p_1z_1+p_2z_2)$:

$$\mu_1(z(s\rho^\alpha))=\mu_1^{'}(1)(z(s\rho^\alpha)-1)+\frac{\mu_1^{''}(1)}{2}(z(s\rho^\alpha)-1)^2+o((z(s\rho^\alpha)-1)^2).\quad(4)$$

Substitute (3) in (4), after easy manipulations quadratic equation for $z(s\rho^\alpha)-1$ is obtained:

$$av\cdot(z(s\rho^\alpha)-1)^2-\rho\cdot(z(s\rho^\alpha)-1)-\beta_1\cdot s\rho^\alpha+o(\max((z(s\rho^\alpha)-1)^2,\rho\cdot(z(s\rho^\alpha)-1),\rho^\alpha))=0,$$

its solutions:

$$z(s\rho^\alpha)-1=\frac{\rho\pm\sqrt{\rho^2+4\,s\rho^\alpha v}}{2av}+o(z(s\rho^\alpha)-1).$$

from here asymptotics in the lemma statement are obtained.

Corollary 1. Asymptotic expansion for $\mu_1^*(s\rho^\alpha)$ is:

$$\mu_1^*(s\rho^\alpha)=\begin{cases}-\sqrt{\dfrac{s\rho^\alpha}{v}}+o(\rho^{\frac{\alpha}{2}}),\ \alpha<2,\\ -\rho\cdot\dfrac{2\,s}{1+\sqrt{1+4sv}}+o(\rho),\ \alpha=2,\\ -s\rho^{\alpha-1}+o(\rho^{\alpha-1}),\ \alpha>2.\end{cases}$$

This expansion is obtained directly from Lemma 4 and (4)

Lemma 5. *The next asymptotics for $p_{0j}(s\rho^{\alpha})$ are true :*

$$p_{0j}(s\rho^{\alpha}) = \begin{cases} \kappa_j \cdot \sqrt{\frac{v}{s}} \cdot \rho^{-\frac{\alpha}{2}} + o(\rho^{-\frac{\alpha}{2}}), \ \alpha < 2, \\ \kappa_j \cdot \frac{1+\sqrt{1+4sv}}{2\,s} \cdot \rho^{-1} + o(\rho^{-1}), \ \alpha = 2, \\ \kappa_j \frac{\rho^{1-\alpha}}{s} \cdot + o(\rho^{1-\alpha}), \ \alpha > 2, \end{cases}$$

where

$$\kappa_j = \prod_{n \neq j} \frac{a_n}{a_n - a_j} \cdot \prod_{k=2}^{N} \frac{\mu_k^*(0) + a_j(1 - p(p, z_k^*(0)))}{\mu_k^*(0)}.$$

Proof. Since, only $\mu_1(1) = 0$ then from (2):

$$\rho^{\alpha} p_{0j}(s\rho^{\alpha}) = \frac{\rho^{\alpha}}{s\rho^{\alpha} - \mu_1^*(s\rho^{\alpha})} \times \prod_{n \neq j} \frac{a_n}{a_n - a_j} \cdot \prod_{k=2}^{N} \frac{\mu_k^*(0) + a_j(1 - p(p, z_k^*(0)))}{\mu_k^*(0)}.$$

Therefore and from corollary 1, lemma statement is obtained.

Lemma 6. *The next asymptotics for $\mu_1(p_1 z_1^* + p_2 e^{-u\rho^{\gamma}})$ are true:*

$$\mu_1(p_1 z_1^* + p_2 e^{-u\rho^{\gamma}}) = \frac{ap_2 u\rho^{\gamma}}{ap_1\beta_{11} - 1} + \psi\rho^{2\gamma} + o(\rho^{2\gamma}), \ \textit{where}$$

$$\psi = \begin{cases} \frac{ap_1\beta_{11}}{ap_1\beta_{11}-1} s + \frac{ap_2u^2}{2(1-ap_1\beta_{11})} + \frac{\mu_1^{''}(1)}{2} \frac{p_2^2u^2}{(1-ap_1\beta_{11})^3} + \frac{p_1\beta_{12}a^3p_2^2u^2}{2(1-ap_1\beta_{11})^3}, \ \alpha \leqslant 2, \\ \frac{ap_2u^2}{2(1-ap_1\beta_{11})} + \frac{\mu_1^{''}(1)}{2} \frac{p_2^2u^2}{(1-ap_1\beta_{11})^3} + \frac{p_1\beta_{12}a^3p_2^2u^2}{2(1-ap_1\beta_{11})^3}, \ \alpha > 2. \end{cases}$$

Proof. Since $p_1 z_1^* + p_2 e^{-u\rho^{\gamma}} = p_1\beta_1(s\rho^{\alpha} - \mu_1(p_1 z_1^* + p_2 e^{-u\rho^{\gamma}})) + p_2 e^{-u\rho^{\gamma}}$.

Denote: $\tau = p_1 z_1^* + p_2 e^{-u\rho^{\gamma}}$. Using the assumption II we have

$$\tau = p_1 \left(1 - \beta_{11}(s\rho^{\alpha} - \mu_1(\tau)) + \frac{\beta_{12}}{2}((s\rho^{\alpha} - \mu_1(\tau)))^2\right) + \\ + p_2\left(1 - u\rho^{\gamma} + \frac{u^2\rho^{2\gamma}}{2}\right) + o(\rho^{2\gamma}). \quad (5)$$

Using the asymptotics for $\mu_1(\tau)$ and separate the principal part with degree ρ^{γ} we have

$$\tau - 1 = \frac{p_2 u\rho^{\gamma}}{ap_1\beta_{11} - 1} + + \psi\rho^{2\gamma} + o(\rho^{2\gamma}). \quad (6)$$

Substitute (6) in (5), ψ might be found.

Lemma 7. *The next asymptotics are true*:

$$\sum_{j=1}^{N} f_j(z_1^*, e^{-u\rho^\gamma}, s\rho^\alpha)\frac{c_j a_j}{\mu_1(p_1 z_1^* + p_2 e^{-u\rho^\gamma}) + a_j(1 - p(p_1 z_1^* + p_2 e^{-u\rho^\gamma}))} =$$

$$= \frac{1}{1-p} \times \begin{cases} 1 + \dfrac{ap_2 u}{ap_1\beta_{11} - 1}\sqrt{\dfrac{v}{s}} + o(1), \ \alpha < 2, \\ 1 + \dfrac{ap_2 u}{ap_1\beta_{11} - 1}\dfrac{1 + \sqrt{1 + 4sv}}{2\,s} + o(1), \ \alpha = 2, \\ 1 + \dfrac{ap_2 u\rho^{2-\alpha}}{ap_1\beta_{11} - 1}\dfrac{1}{s} + o(\rho^{2-\alpha}), \ \alpha > 2. \end{cases}$$

Proof. Using the definition of $f_j(z_1, z_2, s), \ j = \overline{1, N}$, the investigated expression might be rewritten in the next form:

$$\sum_{j=1}^{N} f_j(z_1^*, e^{-u\rho^\gamma}, s\rho^\alpha)\frac{c_j a_j}{\mu_1(p_1 z_1^* + p_2 e^{-u\rho^\gamma}) + a_j(1 - p(p_1 z_1^* + p_2 e^{-u\rho^\gamma}))} =$$

$$= \frac{1}{(1-p)(p_1 z_1^* + p_2 e^{-u\rho^\gamma})} - (s\rho^\alpha - \mu_1(p_1 z_1^* + p_2 e^{-u\rho^\gamma}))\times$$

$$\times \sum_{j=1}^{N} \frac{a_j p_{0j}(s)}{\mu_1(p_1 z_1^* + p_2 e^{-u\rho^\gamma}) + a_j(1 - p(p_1 z_1^* + p_2 e^{-u\rho^\gamma}))} + o(\rho^{2\gamma}).$$

After using Lemma 5 and the fact that $\sum\limits_{j=1}^{N} \kappa_j = 1$, the lemma statement is obtained.

Lemma 8. *The next asymptotics are true*:

$$e^{-u\rho^\gamma} - \beta_2(s\rho^\alpha - \mu_1(p_1 z_1^* + p_2 e^{-u\rho^\gamma})) =$$

$$\begin{cases} \dfrac{\beta_{21}\rho^{2\gamma}}{ap_1\beta_{11} - 1} \times \left[-s + \dfrac{a^2 p_2^2 v}{(ap_1\beta_{11} - 1)^2} u^2 \right] + o(\rho^{2\gamma}), \ \alpha < 2, \\ \dfrac{\rho^2}{ap_1\beta_{11} - 1} \times \left[u - \beta_{21} s + \dfrac{a^2 p_2^2 \beta_{21} v}{(ap_1\beta_{11} - 1)^2} u^2 \right] + o(\rho^2), \ \alpha = 2, \\ \dfrac{\rho^2}{ap_1\beta_{11} - 1} \times \left[u + \dfrac{a^2 p_2^2 \beta_{21} v}{(ap_1\beta_{11} - 1)^2} u^2 \right] + o(\rho^2), \ \alpha > 2. \end{cases}$$

Proof. These expressions might be obtained directly from Lemma 6 and the following expansion

$$e^{-u\rho^\gamma} - \beta_2(s\rho^\alpha - \mu_1(p_1 z_1^* + p_2 e^{-u\rho^\gamma})) =$$

$$= 1 - u\rho^\gamma + \frac{u^2\rho^{2\gamma}}{2} - 1 + \beta_{21}\left[s\rho^\alpha - \frac{ap_2 u\rho^\gamma}{ap_1\beta_{11} - 1} - \psi\rho^{2\gamma} \right] - \frac{\beta_{22}}{2}\left[\frac{ap_2 u\rho^\gamma}{ap_1\beta_{11} - 1} \right]^2 + o(\rho^{2\gamma}).$$

4.2 Main Theorem

Theorem 1. *While $m \to \infty$ the next limit exists*

$$\lim_{m\to\infty} \mathbb{P}\left(\rho^{\gamma} \cdot L_2\left(\frac{t}{\rho^{\alpha}}\right) < x\right) =$$

$$= \begin{cases} \sqrt{\frac{2}{\pi}} \cdot \int\limits_0^{\sqrt{\frac{v^*}{2t}}wx} e^{-\frac{y^2}{2}}dy, \ \alpha < 2, \\ \\ 1 - \frac{e^{-wx}}{\sqrt{\pi}} \int\limits_{-\sqrt{\frac{t}{4v^*}}+wx\sqrt{\frac{v^*}{4t}}}^{+\infty} e^{-y^2}dy - \frac{1}{\sqrt{\pi}} \int\limits_{\sqrt{\frac{t}{4v^*}}+wx\sqrt{\frac{v^*}{4t}}}^{+\infty} e^{-y^2}dy, \ \alpha = 2, \\ \\ 1 - e^{-wx}, \ \alpha > 2, \end{cases}$$

where

$$w = \frac{1 - a^* p_1^* \beta_{11}^*}{a^* p_2^* v^*}.$$

Proof. Instead of finding $\lim\limits_{m\to\infty} \mathbb{P}\left(\rho^{\gamma} \cdot L_2\left(\frac{t}{\rho^{\alpha}}\right) < x\right)$, we will find $\lim\limits_{\rho\to 0} \rho^{\alpha} \cdot p(1, e^{-u\rho^{\gamma}}, s\rho^{\alpha})$ and then inverse the Laplace transform to obtain the original limit distribution.
Using the expression from Lemma 3, the investigated limit might be rewritten in the next form:

$$\lim_{\rho\to 0}\Bigg\{\rho^{\alpha} \cdot p_0(s\rho^{\alpha}) + \rho^{\alpha} \cdot p_2(e^{-u\rho^{\gamma}} - 1) \cdot \frac{1}{\mu_1(p_1 + p_2 e^{-u\rho^{\gamma}})(s\rho^{\alpha} - \mu_1(p_1 + p_2 e^{-u\rho^{\gamma}}))} \times$$
$$\times \Bigg[\gamma_2^{(1)}(1, e^{-u\rho^{\gamma}}, s\rho^{\alpha}) \frac{1 - e^{-u\rho^{\gamma}}}{(1-p)e^{-u\rho^{\gamma}}} \beta_2(s\rho^{\alpha} - \mu_1(p_1 + p_2 e^{-u\rho^{\gamma}})) +$$
$$+ \frac{\prod\limits_{m=1}^{N} [\mu_1(p_1 + p_2 e^{-u\rho^{\gamma}}) + a_m(1 - p(p_1 + p_2 e^{-u\rho^{\gamma}}))]}{\alpha_1(p_1 + p_2 e^{-u\rho^{\gamma}})} \times$$
$$\times \sum_{j=1}^{N} \frac{c_j a_j f_j(1, e^{-u\rho^{\gamma}}, s\rho^{\alpha})}{\mu_1(p_1 + p_2 e^{-u\rho^{\gamma}}) + a_j(1 - p(p_1 + p_2 e^{-u\rho^{\gamma}}))}\Bigg]\Bigg\}.$$

Considering the fact, that $\lim\limits_{\rho\to 0} \rho^{\alpha} p_0(s\rho^{\alpha}) \to 0$ from lemma 5, $\mu_1(p_1 + p_2 e^{-u\rho^{\gamma}}) = ap_2 u\rho^{\gamma} + o(\rho^{\gamma})$. From Lemma 1 it is possible to find

$$\alpha_1(1) = \prod_{m=2}^{N} (-\mu_m(1)) = \frac{1}{a(1-p)} \prod_{m=1}^{N} (a_m(1-p)),$$

then

$$\lim_{\rho\to 0} \frac{\prod\limits_{m=1}^{N} [\mu_1(p_1 + p_2 e^{-u\rho^{\gamma}}) + a_m(1 - p(p_1 + p_2 e^{-u\rho^{\gamma}}))]}{\alpha_1(p_1 + p_2 e^{-u\rho^{\gamma}})} = \frac{\prod\limits_{m=1}^{N} (a_m(1-p))}{\alpha_1(1)} = a(1-p).$$

So that, the task is equal to finding the next limit:

$$\lim_{\rho\to 0}\left\{\frac{\rho^{\alpha}}{a^2p_2(1-p)(p_1+p_2e^{-u\rho^{\gamma}})}\gamma_2^{(1)}(1,e^{-u\rho^{\gamma}},s\rho^{\alpha})+\right.$$
$$\left.+\frac{\prod_{m=1}^{N}(a_m(1-p))}{\prod_{m=2}^{N}(-\mu_m(1))}\frac{\rho^{\alpha-\gamma}}{a^2p_2u}\sum_{j=1}^{N}\frac{c_ja_jf_j(1,e^{-u\rho^{\gamma}},s\rho^{\alpha})}{\mu_1(p_1+p_2e^{-u\rho^{\gamma}})+a_j(1-p(p_1+p_2e^{-u\rho^{\gamma}}))}\right\}.$$

Similary to the proof of the Lemma 7, we have

$$\lim_{\rho\to 0}\frac{\rho^{\alpha-\gamma}}{a^2p_2u}\sum_{j=1}^{N}f_j(1,e^{-u\rho^{\gamma}},s\rho^{\alpha})\frac{c_ja_j}{\mu_1(p_1+p_2e^{-u\rho^{\gamma}})+a_j(1-p(p_1+p_2e^{-u\rho^{\gamma}}))}=0\ \ \forall\alpha>0.$$

Therefore,

$$\lim_{\rho\to 0}\rho^{\alpha}p(1,e^{-u\rho^{\gamma}},s\rho^{\alpha})=\lim_{\rho\to 0}\frac{\rho^{\alpha}\gamma_2^{(1)}(1,e^{-u\rho^{\gamma}},s\rho^{\alpha})}{a^2p_2(1-p)(p_1+p_2e^{-u\rho^{\gamma}})}=$$
$$=\lim_{\rho\to 0}\frac{\prod_{m=1}^{N}(a_m(1-p))}{\alpha_1(1)(1-p)a^2p_2}\sum_{e=1}^{N}\rho^{\alpha}p_{2e}(e^{-u\rho^{\gamma}},0,s\rho^{\alpha})=$$
$$=\lim_{\rho\to 0}\frac{1}{ap_2}\sum_{e=1}^{N}\rho^{\alpha}p_{2e}(e^{-u\rho^{\gamma}},0,s\rho^{\alpha})=\lim_{\rho\to 0}\frac{1}{p_2a^2(1-p)}\rho^{\alpha}\gamma_2^{(1)}(z_1^*,e^{-u\rho^{\gamma}},s\rho^{\alpha}).$$

From the definition of $\gamma_2^{(1)}(z_1,z_2,s)$, we have

$$\lim_{\rho\to 0}\frac{\rho^{\alpha}\gamma_2^{(1)}(z_1^*,e^{-u\rho^{\gamma}},s\rho^{\alpha})}{p_2a^2(1-p)}=$$
$$=\lim_{\rho\to 0}\frac{(p_1z_1^*+p_2e^{-u\rho^{\gamma}})e^{-u\rho^{\gamma}}\cdot\rho^{\alpha}}{p_2a^2\alpha_1(p_1z_1^*+p_2e^{-u\rho^{\gamma}})\cdot[e^{-u\rho^{\gamma}}-\beta_2(s\rho^{\alpha}-\mu_1(p_1z_1^*+p_2e^{-u\rho^{\gamma}}))]}\times$$
$$\times\prod_{m=1}^{N}[\mu_1(p_1z_1^*+p_2e^{-u\rho^{\gamma}})+a_m(1-p(p_1z_1^*+p_2e^{-u\rho^{\gamma}}))]\times$$
$$\times\sum_{j=1}^{N}\frac{c_ja_jf_j(z_1^*,e^{-u\rho^{\gamma}},s\rho^{\alpha})}{\mu_1(p_1z_1^*+p_2e^{-u\rho^{\gamma}})+a_j(1-p(p_1z_1^*+p_2e^{-u\rho^{\gamma}}))}=$$
$$=\lim_{\rho\to 0}\frac{(1-p)\rho^{\alpha}}{p_2a[e^{-u\rho^{\gamma}}-\beta_2(s\rho^{\alpha}-\mu_1(p_1z_1^*+p_2e^{-u\rho^{\gamma}}))]}\times$$
$$\times\sum_{j=1}^{N}\frac{c_ja_jf_j(z_1^*,e^{-u\rho^{\gamma}},s\rho^{\alpha})}{\mu_1(p_1z_1^*+p_2e^{-u\rho^{\gamma}})+a_j(1-p(p_1z_1^*+p_2e^{-u\rho^{\gamma}}))}.$$

Using Lemmas 7 and 8, we have

$$\lim_{\rho\to 0}\rho^{\alpha}p(1,e^{-u\rho^{\gamma}},s\rho^{\alpha})=\begin{cases}\left[s\cdot\left(1+\dfrac{a^{*}p_2^{*}u}{1-a^{*}p_1^{*}\beta_{11}^{*}}\sqrt{\dfrac{v^{*}}{s}}\right)\right]^{-1}, & \alpha<2,\\ \left[s\cdot\left(1+\dfrac{a^{*}p_2^{*}u}{1-a^{*}p_1^{*}\beta_{11}^{*}}\cdot\dfrac{2v^{*}}{1+\sqrt{1+4sv^{*}}}\right)\right]^{-1}, & \alpha=2,\\ \left[s\cdot\left(1+\dfrac{a^{*}p_2^{*}v^{*}}{1-a^{*}p_1^{*}\beta_{11}^{*}}u\right)\right]^{-1}, & \alpha>2.\end{cases}$$

After inversing the Laplace transform, the theorem statement is obtained.

Corollary 2. *Probability density function of the limit distribution is:*

$$\begin{cases}\sqrt{\dfrac{v^{*}}{\pi t}}w\cdot exp\left\{-\dfrac{v^{*}w^{2}x^{2}}{4t}\right\}, \ \alpha<2,\\[2ex] \dfrac{we^{-wx}}{\sqrt{\pi}}\displaystyle\int_{-\sqrt{\frac{t}{4v^{*}}}+wx\sqrt{\frac{v^{*}}{4t}}}^{+\infty}e^{-y^{2}}dy+w\sqrt{\dfrac{v^{*}}{4\pi t}}\times\\ \times\left(exp\left\{-wx-\left(-\sqrt{\frac{t}{4v^{*}}}+wx\sqrt{\frac{v^{*}}{4t}}\right)^{2}\right\}+exp\left\{\left(\sqrt{\frac{t}{4v^{*}}}+wx\sqrt{\frac{v^{*}}{4t}}\right)^{2}\right\}\right), \ \alpha=2,\\ w\cdot exp\{-wx\}, \ \alpha>2.\end{cases}$$

Corollary 3. *Mathematical expectation of the limit distribution is equal to*

$$\sqrt{\frac{t}{v^{*}\pi}}\cdot\frac{2}{w}, \ if \ \alpha<2,$$

and

$$\frac{1}{w}, \ if \ \alpha>2.$$

Corollary 4. *Variance of the limit distribution is equal to*

$$\frac{(2\pi-4)t}{\pi v^{*}w^{2}}, \ if \ \alpha<2$$

and

$$\frac{1}{w^{2}}, \ if \ \alpha>2.$$

Remark 1. Mathematical expectation and variance in case $\alpha = 2$ should be calculated using numerical methods.

5 Numerical Examples

For visualisation results of this paper let us consider the system with parameters: $n = 2, a_1 = 1, a_2 = 2, c_1 = 0.35, c_2 = 0.65, \beta_{11} = 0.5749, \beta_{21} = 0.775, \beta_{12} = 1, \beta_{22} = 1, p = 0.5, p_1 = 0.5, p_2 = 0.5$. The plots below show the density functions for all different cases for α and for different parameter t.

Also let us show, how mathematical expectation is changing, while parameters p_2 or p are being changed, the main parameters of the system are $n = 2, a_1 = 1, a_2, c_1 = 0.2, c_2 = 0.8$, the other parameters has been chosen such way that ρ is close to 0 (Figs. 1, 2, 3 and Tables 1, 2).

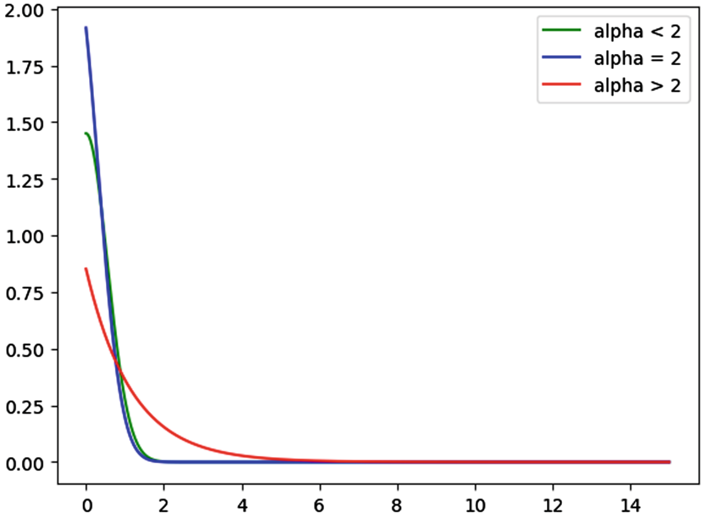


Fig. 1. t = 0.1

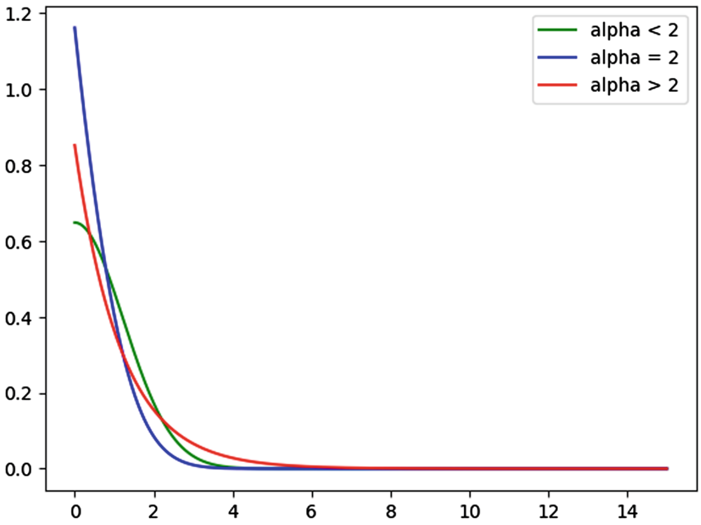


Fig. 2. t = 0.5

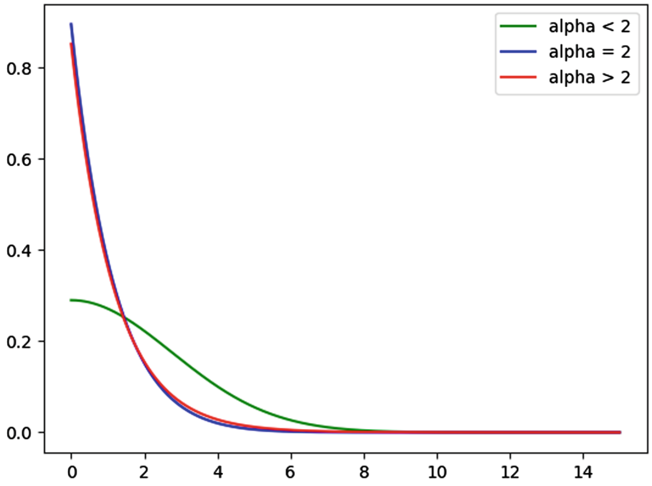


Fig. 3. t = 2.5

Table 1. Mathematical expectation while the probability of going to the lowest class is being changed

p_2	$t = 0.1$			$t = 1$			$t = 10$		
	$\alpha < 2$	$\alpha = 2$	$\alpha > 2$	$\alpha < 2$	$\alpha = 2$	$\alpha > 2$	$\alpha < 2$	$\alpha = 2$	$\alpha > 2$
0.05	0.595	0.522	1.802	1.882	1.242	1.802	5.952	1.785	1.802
0.25	0.605	0.53	1.83	1.912	1.261	1.83	6.046	1.813	1.83
0.5	0.617	0.541	1.867	1.95	1.286	1.867	6.167	1.849	1.867
0.75	0.629	0.552	1.905	1.99	1.313	1.905	6.292	1.887	1.905
0.95	0.64	0.561	1.936	2.023	1.334	1.936	6.397	1.918	1.936

Table 2. Mathematical expectation while the probability of repeating intensity is being changed:

p	$t = 0.1$			$t = 1$			$t = 10$		
	$\alpha < 2$	$\alpha = 2$	$\alpha > 2$	$\alpha < 2$	$\alpha = 2$	$\alpha > 2$	$\alpha < 2$	$\alpha = 2$	$\alpha > 2$
0.05	0.561	0.483	1.493	1.106	1.493	1.493	1.488	1.493	1.493
0.25	0.566	0.489	1.524	1.122	1.524	1.524	1.518	1.524	1.524
0.5	0.58	0.502	1.598	1.161	1.598	1.598	1.59	1.598	1.598
0.75	0.619	0.541	1.818	1.275	1.818	1.818	1.804	1.818	1.818
0.95	0.868	0.789	3.581	2.023	3.581	3.581	3.425	3.581	3.581

6 Conclusion

The main result of this paper is an explicit form of the limit distribution of the queue length for the least priority class, which has been obtained. For each cases ($\alpha < 2,\ \alpha = 2$ and $\alpha > 2$) expressions for the density function are obtained. For cases ($\alpha < 2$ and $\alpha > 2$) mathematical expectation and variance are given in explicit form, in case $\alpha = 2$ these characteristics might be calculated numerically. Numerical examples show us necessity to consider parameter t, since relation among considered cases is changing while t is being changed. Provided theoretical results of this article can be used to analyse real queueing systems in which there is a correlation of the intervals between customer arrivals.

References

1. Bergovin, A.K., Ushakov, V.G.: Discipline-priority queuing systems without serving interruptions. Moscow Univ. Comput. Math. Cybern. **42**(3), 119–125 (2018). https://doi.org/10.3103/S0278641918030032
2. Gusella, R.: Characterizing the variability of arrival processes with indexes of dispersion. IEEE J. Sel. Areas Commun. **2**, 203–211 (1991). https://doi.org/10.1109/49.68448
3. Paxson, V., Floyd, S.: Wide-Area Traffic: the failure of poisson modelling. IEEE/ACM Trans. Netw. (ToN) **3**, 226–244 (1995). https://doi.org/10.1109/90.392383
4. Hwang, G.U., Choi, B.D., Kim, J.-K.: The waiting time analysis of a discrete-time queue with arrivals as an autoregressive process of order 1. J. Appl. Probab. **3**, 619–629 (2002). https://doi.org/10.1239/jap/1034082132
5. Kamoun, F.: The discrete-time queue with autoregressive inputs revisited. Queueing Syst., 185–192 (2006). https://doi.org/10.1007/s11134-006-9591-3
6. Prokhorov, Y.V.: Transitional phenomena in queuing processes. Lith. Math. Collection **3**(1), 199–206 (1963)
7. Kingman, J.F.C.: On queues in heavy traffic. J. Rey. Stat. Soc. B-25, 383–392 (1962). https://doi.org/10.1111/j.2517-6161.1962.tb00465.x
8. Borovkov, A.A.: Asymptotic methods in queuing theory. M., Nauka, 357 (1979)
9. Nazarov, A.A., Moiseeva, S.P.: Method of Asymptotic Analysis in Queuing Theory, p. 112. Tomsk state University, Tomsk (2006)
10. Garayshina, I.R., Moiseeva, S.P., Nazarov, A.A.: Methods for studying correlated flows and special queuing systems Tomsk: NTL Publishing House, p. 206 (2010)
11. Nazarov, A., Paul, S., Lizyura, O.: Asymptotic analysis of Markovian retrial queue with unreliable server and multiple types of outgoing calls. Global Stoch. Anal. **8**(3), 143–149 (2021)
12. Danielyan, E.A.: Description of one class of limit distributions in single-channel priority systems, pp. 48–52. VNIISI, Queuing theory. M. (1981)
13. Ushakov, A.V.: Heavy-traffic analysis for the queueing system with hyper exponential stream. Inform. Appl. **6**(3), 117–121 (2012)
14. Hooke, J.A.: Some heavy-traffic limit theorems for a priority queue with general arrivals. Operat. Res. **20**, 381–388 (1972). https://doi.org/10.1287/opre.20.2.381
15. Hooke, J.A., Prabhy, N.V.: Priority queues in heavy traffic. Opns. Res., 1–9 (1971)
16. Abate, J., Whitt, W.: Asymptotics for M—G—1 low-priority waiting-time tail probabilities. Queueing Syst. **25**, 173–233 (1997)

Markov Decision Process and Artificial Neural Network for Resource Capacity Planning in 5G Network Slicing

Ibram Ghebrial[1], Kseniia Leonteva[1], Irina Kochetkova[1,2](✉), and Sergey Shorgin[2]

[1] RUDN University, 6 Miklukho-Maklaya St, Moscow, Russian Federation
kochetkova-ia@rudn.ru

[2] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilova St, Moscow, Russian Federation

Abstract. The advent of fifth-generation (5G) networks introduces the capability for network slicing, which allows for the creation of multiple, distinct logical network slices on a shared infrastructure, tailored to meet the diverse requirements of different user groups. This study is dedicated to the development and analysis of a resource capacity planning and reallocation model within the context of network slicing, with a particular focus on the dynamics between two service providers managing elastic traffic types such as web browsing. A controller is tasked with assessing the necessity for resource redistribution and devising subsequent capacity planning strategies. To determine an optimal resource capacity for each service provider, we utilize a Markov decision process within a controllable queuing framework. The optimization process is guided by a reward function that adheres to three principles of network slicing: maximum matching for equal resource partitioning, maximum share of signals resulting in resource reallocation, and maximum resource utilization. We employ an artificial neural network (multilayer perceptron) with three layers to find the optimal policy. The state of the system is used as the input layer, and the possible size of the slice for the first provider is the output layer. For the considered scenario of web browsing and group data transfer, we numerically demonstrate the impact of the number of neurons in the hidden layer and the number of training epochs on classification accuracy. We compare the results with those obtained using R. Howard's iteration.

Keywords: 5G · network slicing · capacity planning · resource reallocation · controller · elastic traffic · Markov decision process · queuing system · artificial neural network · multilayer perceptron

This publication has been supported by the RUDN University Scientific Projects Grant System, project No. 025319-2-000.

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 33–47, 2024.
https://doi.org/10.1007/978-3-031-65385-8_3

1 Introduction

The fifth generation (5G) networks offer high-speed data transmission and a wide range of services for multiple users, ensuring high quality of service (QoS) [9,10,13]. These networks support three main scenarios: enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable low-latency communications (URLLC) [7]. Looking toward the future, the sixth generation (6G) network is expected to provide global coverage and improved efficiency. To achieve these goals, new technologies and systems will need to be implemented at the radio interface and in the core network, such as multiple access, cloud/fog computing, and network slicing [5].

Network slicing is the concept of creating logical, independent network resources on a single infrastructure platform [1,6]. This enables the customization of network slices to meet specific requirements. Three main business roles are involved in network slicing: the network slice provider, the network slice service provider, and the network slice service user. Resource capacity planning can be accomplished in two ways: static resource partitioning and dynamic resource reallocation. Static resource partitioning involves allocating resources to network slices that remain unchanged over time. Conversely, dynamic resource reallocation involves modifying resource allocation between slices based on periodic evaluations by the controller. These evaluations consider various mechanisms and criteria, such as slice priorities, weights, resource utilization, fairness, availability, and isolation.

Several recent studies have demonstrated the use of Markov decision processes (MDP) in optimizing resource capacity planning for network slicing tasks. The study in [14] utilizes federated deep reinforcement learning (DRL) with MDP to ensure QoS and manage network load for mobile virtual network operators (MVNOs). In [16], a combination of double deep Q-networks (DDQN), the Dijkstra algorithm, and binary search-assisted gradient descent with MDP is used to minimize the number of accepted service requests with higher priority and overall cost. The paper [8] uses a reinforcement learning (RL) and MDP framework for monitoring and adjusting resource utilization based on network slice and multi-access edge computing (MEC) node statuses. In [4], the authors introduce an exponential weight algorithm and multi-agent DQN with MDP to optimize resource block allocation for URLLC and eMBB slices. Collectively, these studies highlight the versatility of MDP in addressing diverse challenges in resource allocation and optimization within network slicing environments.

The application of artificial neural networks (ANNs) is currently gaining significant traction in various domains, including queuing theory. For instance, Efrosinin et al. [3] study the optimal scheduling problem with parallel queues and a single server by combining modeling methods with neural networks. The paper [2] considers heterogeneous queues where servers differ in terms of service rate and operating cost. Optimal threshold values are calculated using MDP by implementing an iterative algorithm to find the policy. The estimates of the optimal threshold values are obtained using an ANN and compared with the heuristic solution.

This paper focuses on the modeling and analysis of resource capacity planning and reallocation for network slicing between two providers handling elastic traffic, such as that encountered during web browsing sessions. We present a mathematical model in the form of a controllable queuing system based on three principles: maximum matching for equal resource partitioning, maximum share of signals resulting in resource reallocation, and maximum resource utilization. An algorithm inspired by R. Howard's iteration is primarily used for identifying the optimal resource capacity planning policy.

The main contributions of our study are as follows:

- We employ an ANN, specifically a multilayer perceptron (MLP) with three layers, to find the optimal policy for resource reallocation between two network slices in a 5G network. This is a classification problem (supervised learning) – the state of the system is used as the input layer, and the possible size of the slice for the first provider is the output layer.
- For the considered scenario of web browsing and group data transfer, we numerically demonstrate the impact of the number of neurons in the hidden layer and the number of training epochs on classification accuracy. We compare the results with those obtained using R. Howard's iteration method. We employ two optimizers – Adam (adaptive moment estimation) and L-BFGS (limited-memory Broyden-Fletcher-Goldfarb-Shanno). We utilize two libraries of the Python programming language: scikit-learn and PyTorch.

The rest of the paper is organized as follows. Section 2 presents general assumptions and main parameters of the system model. Section 3 describes the continuous-time MDP and the controllable queuing model employed to address optimal resource reallocation. In Sect. 4, we investigate the deployment of an ANN to fine-tune the resource allocation policy. Section 5 presents numerical results, assessing how the number of neurons in the ANN and the training epochs influence the model's accuracy. Conclusions are drawn in the final section.

2 System Model

In this section, we describe the system model for resource reallocation in network slicing, emphasizing the pivotal role of the controller and the principles considered during this reallocation.

2.1 General Assumptions

This paper examines a network slice provider, also known as a network operator, with a fixed total capacity of C bps. This capacity is shared between two network slice service providers, denoted as $K = 2$, each serving its own set of network slice service users or communication service customers. The allocation of resources between these service providers is managed by the controller of the network slice provider, which sends signals to indicate the need for reallocation. The inter-arrival time of these signals follows an exponential distribution

Table 1. Main notation.

Parameter	Description
C	Total bitrate (capacity) for all network slices resources, bps
$K = 2$	Number of service providers, i.e., network slices
δ	Arrival rate of signals from the controller for resource reallocation between service providers, 1/s
w_1	Weight for the maximum matching equal resource partition
w_2	Weight for the maximum share of the signals resulting in resource reallocation
w_3	Weight for the maximum resource utilization
b	Minimal bitrate guarantee for transmitting elastic traffic, bps
$N = \lfloor \frac{C}{b} \rfloor$	Maximum number of all providers users jointly transmitting elastic traffic
R_k	Maximum number of the k-provider users (k-users) waiting for delayed elastic traffic transmission (size of k-buffer)
λ_k	Arrival rate of requests for elastic traffic transmission from k-users, 1/s
μ_k^{-1}	Average volume of elastic traffic transmitted by k-users, bit
ε_k	Abandonment rate due to impatience of k-users from k-buffer, 1/s
m	Maximum number of 1-users jointly transmitting elastic traffic (size of 1-slice)
n_1	Number of 1-users transmitting elastic traffic and waiting in 1-buffer
n_2	Number of 2-users transmitting elastic traffic and waiting in 2-buffer
$\mathbf{s} = (m, n_1, n_2)$	State of the system

with parameter δ. The decision-making process is based on three principles [11]: maximum matching for equal resource partitioning, maximum share of signals resulting in resource reallocation, and maximum resource utilization. Each principle is assigned a weight, represented by w_1, w_2, and w_3 respectively.

The network slice service provider offers a service to its users that guarantees a minimum bitrate of b for transmitting elastic traffic [15]. This implies that the total number of users from all providers who can simultaneously transmit elastic traffic is limited to $N = \lfloor \frac{C}{b} \rfloor$. While a certain level of delay in transmission is deemed acceptable, the number of users from each provider waiting for delayed elastic traffic is constrained by R_k, the size of the provider's buffer. User requests for elastic traffic transmission follow a Poisson process with a rate of λ_k. The volume of elastic traffic transmitted by each user follows an exponential distribution with an average of μ_k^{-1}. Additionally, each user exhibits impatience and will abandon their provider's buffer after an exponentially distributed amount of time with a parameter of ε_k.

The main notations are listed in Table 1, and Fig. 1 illustrates the scheme of the model with these notations as well.

2.2 Resource Capacity Planning

The controller plays a pivotal role in the successful implementation of resource capacity planning through the periodic transmission of signals to assess the need

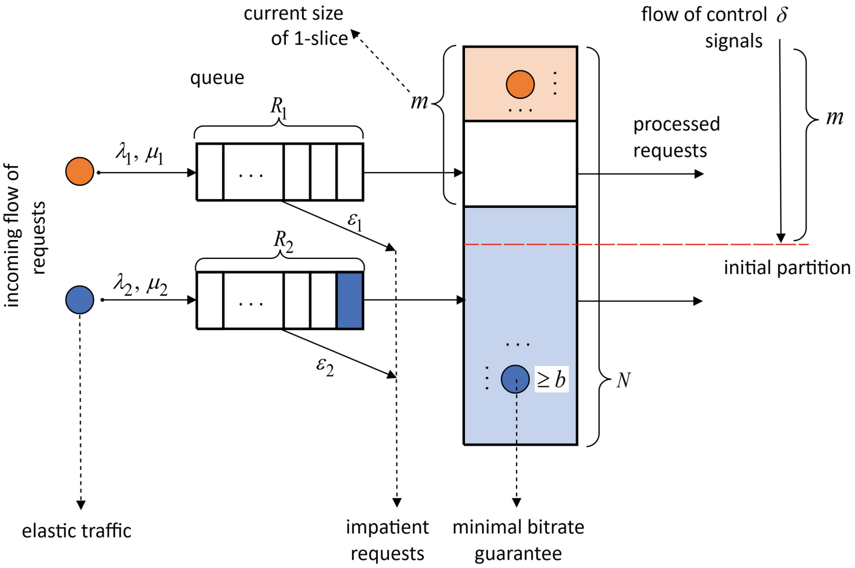


Fig. 1. Scheme of the model.

for resource reallocation. The primary responsibility of the controller is to determine the most efficient approach for scheduling resources, guided by three key principles: maximum matching of equal resource partitions, maximum utilization of signals that result in resource reallocation, and maximum utilization of resources.

Resource reallocation is triggered when the controller transmits a signal, and there are available resources in one slice while another provider has pending requests [12]. This type of signal is denoted as “resulting” and will lead to resource reallocation. However, if all buffers are empty or all slices are currently in use at the time of signal reception, no reallocation will occur.

Resource reallocation will take place in two specific scenarios: (i) when the resources in 1-slice are fully utilized, but there are still users waiting for service, and there are available resources in 2-slice; or (ii) when the resources in 2-slice are fully utilized and there are users waiting for service, but there are available resources in 1-slice. In these cases, idle resources from one slice will be reassigned to the other slice. However, resource reallocation will not occur if any of the following conditions are present: (i) the system is empty, not all resources are occupied, (ii) all resources are occupied but no users are waiting for service, (iii) all resources are occupied and users waiting for service are from only one slice, or (iv) all resources are occupied and there are users waiting for service from both slices.

2.3 Three Principles for Resource Reallocation

The controller operates based on three principles, each assigned a weight w_i, where $i = 1, 2, 3$:

- The first principle accounts for the deviation from the equal resource partition, specifically the number of requests waiting due to an unequal resource partition. This consideration is crucial as the optimal policy should align with the initial partition outlined in the service level agreement (SLA) between the network slice provider and the network slice service providers.
- The second principle assesses the frequency of instances where resource reallocation does not occur upon the arrival of a signal. It is imperative that signals from the controller lead to actual resource reallocation. A frequent occurrence of "non-resulting" signals is considered undesirable as it generates unnecessary signaling messages. This aspect is also significant in radio resource management strategies, as it can affect resource efficiency and utilization.
- The third principle evaluates the amount of available resources while users are waiting. Here, maximizing resource utilization is paramount. Having idle resources is considered undesirable; for example, when a user is waiting for a slice that is currently occupied, while another slice has available resources to serve the user's request.

3 Markov Decision Process and Controllable Queuing Model

In this section, we employ a continuous-time MDP to model resource reallocation within a dynamic network slicing framework.

3.1 Continuous-Time MDP

We define the system behavior using a continuous-time MDP. The states are denoted by $\mathbf{s} = (m, n_1, n_2)$, where m signifies the maximum number of 1-users jointly transmitting elastic traffic (size of 1-slice), and n_k represents the number of k-users transmitting elastic traffic and waiting in k-buffer. To tackle the challenge of optimal resource reallocation, selecting the most appropriate action for determining the volume of resources to be reallocated based on the system's present state is essential. The MDP model is characterized by a 4-tuple $(\mathcal{S}, \mathcal{A}_{\mathbf{s}}, \mathbf{Q}_a, R(\mathbf{s}))$:

1. The set $\mathcal{S}$ of states $\mathbf{s} = (m, n_1, n_2) \in \mathcal{S}$

$$\mathcal{S} = \Big\{\mathbf{s} = (m, n_1, n_2) : \ m = 0, \ldots, N, \\ n_1 = 0, \ldots, m + R_1, \ \ n_2 = 0, \ldots, N - m + R_2\Big\}.$$

2. The set $\mathcal{A}_{\mathbf{s}}$ of actions a available from state $\mathbf{s}$

$$\mathcal{A}_{\mathbf{s}} = \begin{cases} \{m, \dots, n_1\}, & n_1 > m, \;\; n_2 < N - m, \;\; n_1 + n_2 \le N, \\ \{m, \dots, N - n_2\}, & n_1 > m, \;\; n_2 < N - m, \;\; n_1 + n_2 > N, \\ \{N - n_2, \dots, m\}, & n_1 < m, \;\; n_2 > N - m, \;\; n_1 + n_2 \le N, \\ \{n_1, \dots, m\}, & n_1 < m, \;\; n_2 > N - m, \;\; n_1 + n_2 > N, \\ \emptyset, & \text{otherwise}. \end{cases}$$

3. Matrix $\mathbf{Q}_a$ of transition rates under action a

$$q(\mathbf{s}'|\mathbf{s}, a) = \begin{cases} \lambda_1, & \mathbf{s}' = (m, n_1 + 1, n_2), \;\; n_1 + 1 \le R_1, \\ \lambda_2, & \mathbf{s}' = (m, n_1, n_2 + 1), \;\; n_2 + 1 \le R_2, \\ \frac{m}{N} C\mu_1, & \mathbf{s}' = (m, n_1 - 1, n_2), \;\; m > 0, \;\; n_1 > 0, \\ \frac{N-m}{N} C\mu_2, & \mathbf{s}' = (m, n_1, n_2 - 1), \;\; N - m > 0, \;\; n_2 > 0, \\ (n_1 - m)\varepsilon_1, & \mathbf{s}' = (m, n_1 - 1, n_2), \;\; n_1 \ge m, \\ (n_2 - N + m)\varepsilon_2, & \mathbf{s}' = (m, n_1, n_2 - 1), \;\; n_2 \ge N - m, \\ \delta, & \mathbf{s}' = (a, n_1, n_2), \;\; a \in \mathcal{A}_{\mathbf{s}}, \\ & n_1 > m, \;\; n_2 < N - m \;\; \vee \\ & n_1 < m, \;\; n_2 > N - m. \end{cases}$$

4. The reward $R(\mathbf{s})$ received while in state $\mathbf{s}$. It will be described in the next subsection.

3.2 Reward Function

In addition, the slice provider can determine which of the described principles is most important and can customize the ratios between them. Each principle has its own weight; therefore, the coefficients w_i for $i = 1, 2, 3$ are introduced to define the significance of the i-th principle. The reward function is taken with a negative sign as it implies a "penalty" for incorrect resource reallocation. Consequently, the reward function takes the following form

$$R(\mathbf{s}) = -\Big(w_1 \cdot \alpha(\mathbf{s}) + w_2 \cdot \beta(\mathbf{s}) + w_3 \cdot \gamma(\mathbf{s})\Big),$$

where α represents the reward for the maximum matching equal resource partition, β denotes the reward for the maximum share of the signals resulting in resource reallocation, and γ signifies the reward for maximum resource utilization:

$$\alpha(\mathbf{s}) = \begin{cases} \frac{N}{2} - m, & m < \frac{N}{2}, \;\; n_1 > m, \;\; n_1 > \frac{N}{2}, \\ n_1 - m, & m < \frac{N}{2}, \;\; n_1 > m, \;\; n_1 \le \frac{N}{2}, \\ \frac{N}{2} - (N - m), & m > \frac{N}{2}, \;\; n_2 > N - m, \;\; n_2 > \frac{N}{2}, \\ n_2 - (N - m), & m > \frac{N}{2}, \;\; n_2 > N - m, \;\; n_2 \le \frac{N}{2}, \\ 0, & \text{otherwise}; \end{cases}$$

$$\begin{aligned}\beta(\mathbf{s}) =& \delta\Big(\delta + \lambda_1 \cdot \mathbf{1}(n_1 + 1 \le R_1 + m) + \lambda_2 \cdot \mathbf{1}(n_2 + 1 \le R_2 + N - m)\\ &+ \frac{m}{N} C\mu_1 \cdot \mathbf{1}(m > 0,\ \ n_1 > 0) + \frac{N-m}{N} C\mu_2 \cdot \mathbf{1}(N - m > 0,\ \ n_2 > 0)\\ &+ (n_1 - m)\varepsilon_1 \cdot \mathbf{1}(n_1 > m) + (n_2 - N + m)\varepsilon_2 \cdot \mathbf{1}(n_2 > N - m)\Big)^{-1}\\ &\times \mathbf{1}(\mathbf{s} \in \overline{\{n_1 > m,\ \ n_2 < N - m\ \ \vee\ \ n_1 < m,\ \ n_2 > N - m\}});\end{aligned}$$

$$\gamma(\mathbf{s}) = \begin{cases} n_1 - m, & n_1 > m,\ \ n_2 < N - m,\ \ n_1 + n_2 \le N,\\ N - n_2 - m, & n_1 > m,\ \ n_2 < N - m,\ \ n_1 + n_2 > N,\\ m - N + n_2, & n_1 < m,\ \ n_2 > N - m,\ \ n_1 + n_2 \le N,\\ m - n_1, & n_1 < m,\ \ n_2 > N - m,\ \ n_1 + n_2 > N,\\ 0, & \text{otherwise.} \end{cases}$$

3.3 Optimal Policy

The optimal policy aims to maximize the average reward. Let us define the average reward as

$$g^a = \sum_{\mathbf{s} \in \mathcal{S}} R(\mathbf{s})\pi(\mathbf{s}),$$

where $\pi(\mathbf{s})$ is the stationary probability distribution.

The iterative algorithm by R. Howard can be utilized to calculate the optimal policy. This algorithm offers a significant advantage over the straightforward brute force approach, particularly regarding computational complexity and the reduced number of iterations required to ascertain the optimal policy. The computational complexity of the brute force method is contingent on the dimensions of the set of permissible policies for each state within the system. Moreover, the convergence rate of the iterative algorithm is dependent on the judicious selection of the initial solution.

The system of equations for the average reward g^a and estimates $v_a(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$ for the iterative solution method is given as:

$$v_a(\mathbf{s}) = \frac{R(\mathbf{s}) + \sum\limits_{\mathbf{s}' \in \mathcal{S} \setminus \mathbf{s}} q(\mathbf{s}'|\mathbf{s}, a) v_a(\mathbf{s}') - g^a}{\sum\limits_{\mathbf{s}' \in \mathcal{S} \setminus \mathbf{s}} q(\mathbf{s}'|\mathbf{s}, a)}, \quad \mathbf{s} \in \mathcal{S}.$$

The objective function for improving the control policy is calculated using the formula:

$$a(\mathbf{s}) = \arg\max_{a \in \mathcal{A}_{\mathbf{s}}} v_a(a, n_1, n_2), \ \ \mathbf{s} \in \mathcal{S}.$$

For the initial step, let us adopt the maximum resource utilization principle. Upon the arrival of a signal at rate δ, the system transitions to state (m', n_1, n_2),

where

$$m' = \begin{cases} n_1, & n_1 > m, \ n_2 < N - m, \ n_1 + n_2 \leq N, \\ N - n_2, & n_1 > m, \ n_2 < N - m, \ n_1 + n_2 > N, \\ m - n_2 - N + m, & n_1 < m, \ n_2 > N - m, \ n_1 + n_2 \leq N, \\ n_1, & n_1 < m, \ n_2 > N - m, \ n_1 + n_2 > N. \end{cases}$$

4 Artificial Neural Network for Optimal Policy Computing

In this section, we discuss an artificial neural network designed to compute the optimal policy for capacity planning within our model.

4.1 ANN Architecture

To determine the optimal policy, we employed an artificial neural network approach. The architecture of the neural network constructed is depicted in Fig. 2. It comprises an input layer, one hidden layer, and an output layer. The input layer contains three neurons, each corresponding to a component of the system's state $\mathbf{s} = (m, n_1, n_2)$. The hidden layer is composed of n neurons, which are connected to the input layer neurons as follows:

$$\mathbf{y} = \mathbf{W}^1 \mathbf{s} + \mathbf{b}^1,$$

where

$$\mathbf{W}^1 = \begin{bmatrix} w^1_{00} & w^1_{01} & w^1_{02} \\ w^1_{10} & w^1_{11} & w^1_{12} \\ \vdots & \vdots & \vdots \\ w^1_{n-1,0} & w^1_{n-1,1} & w^1_{n-1,2} \end{bmatrix}, \quad \mathbf{b}^1 = \begin{bmatrix} b^1_0 \\ b^1_1 \\ \vdots \\ b^1_{n-1} \end{bmatrix},$$

where $\mathbf{W}^1$ are weights and $\mathbf{b}^1$ are biases.

Subsequently, the Rectified Linear Unit (ReLU) activation function is applied to the neurons of the hidden layer. This function introduces nonlinearity to the processed values, enabling the neural network to make more accurate predictions. The use of ReLU also helps prevent the exponential growth in computation required for operating the neural network. Since ReLU is characterized by non-saturating properties, the data must be normalized between 0 and 1 prior to application.

$$y'_i = \mathrm{ReLu}(y_i) = \max(0, y_i), \quad i = 0, \ldots, n-1.$$

The output layer consists of k neurons, corresponding to the number of possible values for m'. The transition from the hidden layer to the output layer is facilitated by a system of linear equations:

$$\mathbf{z} = \mathbf{W}^2 \mathbf{y}' + \mathbf{b}^2,$$

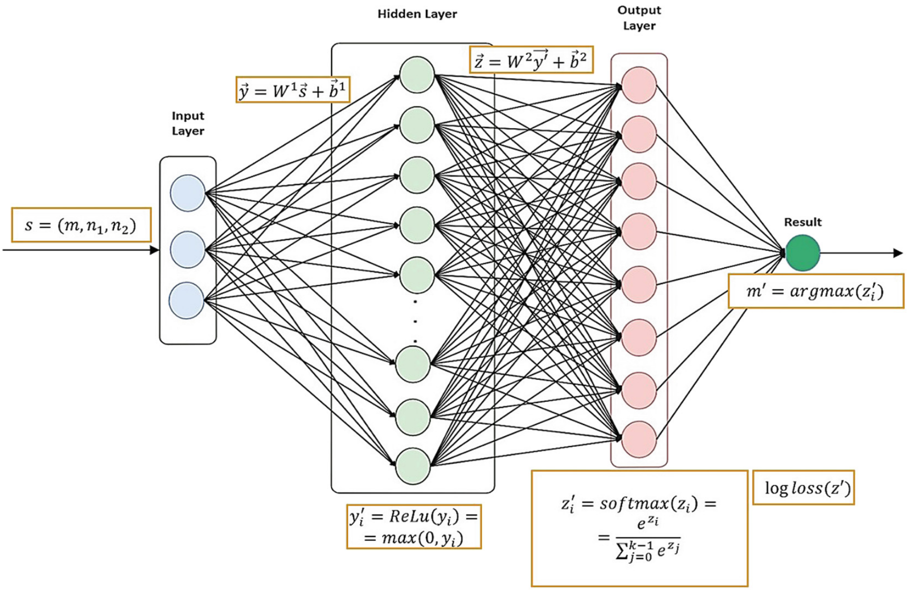


Fig. 2. Used multilayer perceptron architecture.

where

$$\mathbf{W}^2 = \begin{bmatrix} w_{00}^2 & w_{01}^2 & \cdots & w_{0,n-1}^2 \\ w_{10}^2 & w_{11}^2 & \cdots & w_{1,n-1}^2 \\ \vdots & \vdots & \ddots & \vdots \\ w_{k-1,0}^2 & w_{k-1,1}^2 & \cdots & w_{k-1,n-1}^2 \end{bmatrix}, \quad \mathbf{b}^2 = \begin{bmatrix} b_0^2 \\ b_1^2 \\ \vdots \\ b_{k-1}^2 \end{bmatrix},$$

where $\mathbf{W}^2$ are weights and $\mathbf{b^2}$) are biases.

Next, we apply the Softmax activation function:

$$z_i' = \mathrm{softmax}(z_i) = \frac{\exp(z_i)}{\sum_{j=0}^{k-1} \exp(z_j)}, \quad i = 0, \ldots, k-1.$$

The neural network's output, m', is determined by identifying the index of the neuron in the output layer with the highest value.

$$m' = \arg\max(z_i').$$

4.2 Loss Function and Evaluation Metric

We utilized two distinct optimizers to modify the weights of the network: the Adam algorithm and the L-BFGS algorithm. The former, a stochastic gradient-based optimizer, computes the average of the first and second moments of the gradients. It is widely used in neural network training and is considered highly effective. The latter approximates the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, which is among the quasi-Newton methods. As will be demonstrated subsequently, for our dataset, the L-BFGS algorithm proved most suitable, converging more rapidly and performing optimally on smaller datasets (up to 1000).

We use cross-entropy loss function. Cross-entropy loss, typically used in multi-class classification with multiple labels, aims to quantify the agreement between predicted probabilities and true class labels. This loss function is utilized to adjust model weights during training. Our goal is to minimize losses – a lower loss indicates a more accurate model, which is why we employ this function. For an ideal model, the cross-entropy loss would be zero.

$$L(\mathbf{z}, j) = -\log\left(\frac{\exp(z_j)}{\sum_{i=0}^{k-1}\exp(z_i)}\right), \quad j = 0, \ldots, k-1.$$

The primary metric under investigation was the accuracy of the results, calculated using the formula:

$$\text{Accuracy}(m', \hat{m}') = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \mathbf{1}(\hat{m}'_i = m'_i).$$

5 Numerical Results

In this section, we present the outcomes of our numerical analyses, which provide insights into various aspects of our study.

5.1 Considered Scenario

We focus on evaluating two specific services, namely web-browsing and bulk data transfer. The numerical parameters for each service are detailed in Table 2. We use the recommended delay time t_{rk} and allowable delay time t_{ak} for data transfer to estimate the minimum bitrate guarantee b and the abandonment rate due to user impatience from buffering ε_k. The total bitrate for all network slice resources is calculated based on the selected bandwidth, modulation and coding scheme (MCS), and multiple input multiple output (MIMO) scheme.

The neural network was implemented using two Python programming language libraries: scikit-learn and PyTorch. In scikit-learn, we employed the MLP-Classifier (multi-layer perceptron), which internally applies the logistic loss function, also known as cross-entropy. This loss function incorporates the softmax mechanism as the neural network's output function, ensuring accurate probability calculations for each class.

5.2 Impact of Number of Neurons

In our research, we analyzed the impact of varying the number of neurons in our neural network on its accuracy when addressing problems using the Adam and

Table 2. Parameters for numerical example.

Parameter	Description	Value
B	Bandwidth	5 MHz
–	MCS	QPSK
–	MIMO scheme	2×2
C	Total bitrate for two network slices	10 Mbps
δ	Arrival rate of signals from the controller	0.000001 1/s
(w_1, w_2, w_3)	Weights for the reallocation principles	$(1, 1, 1)$
t_{r1}, t_{r2}	Recommended delay time	15, 2 s
t_{a1}, t_{a2}	Allowable delay time	60, 4 s
b	Minimal bitrate guarantee	1.067 Mbps
R_1, R_2	Size of 1-buffer and 2-buffer	5, 5
λ_1, λ_2	Arrival rate of requests from 1-users and 2-users	0.03, 0.6 1/s
μ_1, μ_2	Average volume of traffic transmitted by 1-users and 2-users	0.125, 0.937 Mb
ε_1, ε_2	Abandonment rate due to impatience of 1-users from 1-buffer and 2-users from 2-buffer	0.01, 0.25 1/s

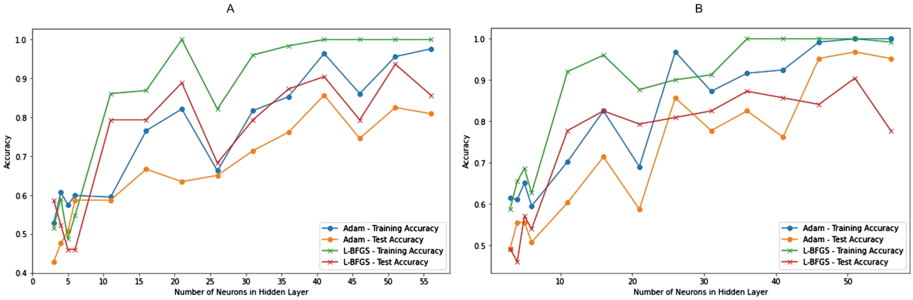


Fig. 3. Accuracy vs number of neurons for Adam and L-BFGS optimizers: (a) scikit-learn Python library, (b) PyTorch Python library.

L-BFGS learning algorithms within the scikit-learn library. Figure 3(a) demonstrates that an increase in the number of neurons generally enhances the accuracy of the training sample. However, beyond a certain threshold, this enhancement may result in overtraining and a subsequent decrease in test sample accuracy. Notably, in some instances, the L-BFGS algorithm yielded higher test sample accuracy compared to Adam. Our findings suggest that the optimal number of hidden neurons for our task is within a moderate range, which balances high test data accuracy and minimizes the risk of overtraining.

In parallel analyses conducted with PyTorch, Fig. 3(b) shows that with the Adam optimizer, learning accuracy improves as the number of neurons increases, achieving 100% with 51 neurons, while test sample accuracy plateaus at approx-

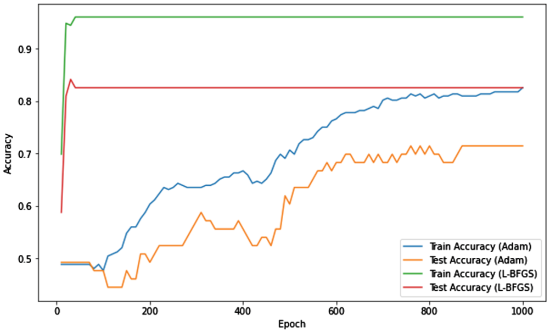


Fig. 4. Accuracy vs number of epochs for Adam and L-BFGS optimizers with 16 hidden neurons in the PyTorch Python library.

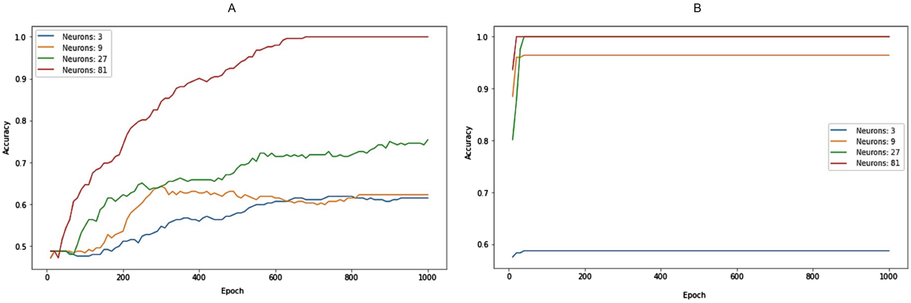


Fig. 5. Accuracy vs number of epochs for different numbers of hidden neurons in the PyTorch Python library: (a) Adam optimizer; (b) L-BFGS optimizer.

imately 97%. Conversely, with the L-BFGS optimizer, learning accuracy also improves with an increasing neuron count, reaching around 100% with 36 neurons. Importantly, test dataset accuracy continued to show enhancement, maintaining a high level of around 90% with 51 neurons.

5.3 Impact of Number of Epochs

Figure 4 depicts that the training sample exhibited a significant rise in accuracy from 48.81% to 81.75% over 1000 epochs using Adam. Analysis of the test sample reveals a steady increase in accuracy from 49.21% to 71.43%. When applying L-BFGS to the training sample, we observed an initial high accuracy level of 69.84%, which stabilized at 96%. The test sample also showed remarkable results, starting at 58.73% and rapidly achieving 82.54%, underscoring the speed and stability advantages of the L-BFGS method.

Figure 5 presents similar results but with varying numbers of neurons in the hidden layer. Our analysis highlights that L-BFGS outperforms in terms of providing consistently high accuracy across fewer epochs for both training and test samples; meanwhile, Adam attains respectable accuracy levels albeit with some variability.

6 Conclusions

Today's 5G networks offer unprecedented opportunities for data transmission, enabling innovative services and applications. Efficient resource allocation is critical to ensuring high-performance and reliable data transmission. This paper addresses the resource scheduling challenge within network slicing, exploring the application of Markov decision processes and artificial neural networks to optimize 5G network resources.

In this paper, we investigate the technology of network slicing using a controllable queuing system with elastic traffic, motivated by the desire to utilize network resources more efficiently. A system model is developed for a network slice provider and two network slice service providers. We establish the guiding principles for resource reallocation, which encompass the maximum matching equal resource partition, the highest share of signals leading to resource reallocation, and the utmost resource utilization. Additionally, we propose an approach that combines Markov decision processes and artificial neural networks to automate and optimize resource management in network slicing. We aim to develop adaptive strategies for optimal resource allocation based on the network's current state and user requirements.

In future research, the application of neural networks will be extended to larger datasets, and their capabilities will be leveraged to estimate performance metrics. Additionally, the exploration of alternative reward functions will be conducted to further enhance our understanding of the system dynamics.

Acknowledgements. This publication has been supported by the RUDN University Scientific Projects Grant System, project No. 025319-2-000.

References

1. Dangi, R., Jadhav, A., Choudhary, G., Dragoni, N., Mishra, M.K., Lalwani, P.: ML-based 5G network slicing security: a comprehensive survey. Future Internet **14**(4) (2022). https://doi.org/10.3390/fi14040116
2. Efrosinin, D., Stepanova, N.: Estimation of the optimal threshold policy in a queue with heterogeneous servers using a heuristic solution and artificial neural networks. Mathematics **9**(11) (2021). https://doi.org/10.3390/math9111267
3. Efrosinin, D., Vishnevsky, V., Stepanova, N.: Optimal scheduling in general multi-queue system by combining simulation and neural network techniques. Sensors **23**(12) (2023). https://doi.org/10.3390/s23125479
4. Filali, A., Mlika, Z., Cherkaoui, S., Kobbane, A.: Dynamic SDN-based radio access network slicing with deep reinforcement learning for URLLC and eMBB services. IEEE Trans. Network Sci. Eng. **9**(4), 2174–2187 (2022). https://doi.org/10.1109/TNSE.2022.3157274

5. Giordani, M., Polese, M., Mezzavilla, M., Rangan, S., Zorzi, M.: Toward 6G networks: use cases and technologies. IEEE Commun. Mag. **58**(3), 55–61 (2020). https://doi.org/10.1109/MCOM.001.1900411
6. Hu, Y., Gong, L., Li, X., Li, H., Zhang, R., Gu, R.: A carrying method for 5G network slicing in smart grid communication services based on neural network. Future Internet **15**(7) (2023). https://doi.org/10.3390/fi15070247
7. (ITU-T), I.T.U.T.S.S.: Framework for the support of network slicing in the IMT-2020 network. Recommendation ITU-T Y.3112 (2018). https://www.itu.int/rec/T-REC-Y.3112-201812-I
8. Kim, Y., Lim, H.: Multi-agent reinforcement learning-based resource management for end-to-end network slicing. IEEE Access **9**, 56178–56190 (2021). https://doi.org/10.1109/ACCESS.2021.3072435
9. Kochetkov, D., Almaganbetov, M.: Using patent landscapes for technology benchmarking: a case of 5G networks. Adv. Syst. Sci. Appl. **21**(2), 20–28 (2021). https://doi.org/10.25728/assa.2021.21.2.988
10. Kochetkov, D., Vuković, D., Sadekov, N., Levkiv, H.: Smart cities and 5G networks: an emerging technological area? J. Geograph. Instit. Jovan Cvijic SASA **69**(3), 289–295 (2019). https://doi.org/10.2298/IJGI1903289K
11. Kochetkova, I., Vlaskina, A., Burtseva, S., Savich, V., Hosek, J.: Analyzing the effectiveness of dynamic network slicing procedure in 5g network by queuing and simulation models. In: Galinina, O., Andreev, S., Balandin, S., Koucheryavy, Y. (eds.) NEW2AN ruSMART 2020. LNCS, pp. 71–85. Springer, Heidelberg (2020). https://doi.org/10.1007/978-3-030-65726-0_7
12. Kochetkova, I., Vlaskina, A., Vu, N., Shorgin, V.: Queuing system with signals for dynamic resource allocation for analyzing network slicing in 5G networks. Informatika i ee Primeneniya **15**(3), 91–97 (2021). https://doi.org/10.14357/19922264210312, in Russian
13. Moltchanov, D., Sopin, E., Begishev, V., Samuylov, A., Koucheryavy, Y., Samouylov, K.: A tutorial on mathematical modeling of 5G/6G millimeter wave and terahertz cellular systems. IEEE Commun. Surv. Tutorials **24**(2), 1072–1116 (2022). https://doi.org/10.1109/COMST.2022.3156207
14. Ou, R., Sun, G., Ayepah-Mensah, D., Boateng, G.O., Liu, G.: Two-tier resource allocation for multitenant network slicing: a federated deep reinforcement learning approach. IEEE Internet Things J. **10**(22), 20174–20187 (2023). https://doi.org/10.1109/JIOT.2023.3283553
15. Vlaskina, A., Polyakov, N., Gudkova, I.: Modeling and performance analysis of elastic traffic with minimum rate guarantee transmission under network slicing. In: Galinina, O., Andreev, S., Balandin, S., Koucheryavy, Y. (eds.) NEW2AN ruSMART 2019. LNCS, vol. 11660, pp. 621–634. Springer, Heidelberg (2019). https://doi.org/10.1007/978-3-030-30859-9_54
16. Xiao, D., Chen, S., Ni, W., Zhang, J., Zhang, A., Liu, R.: A sub-action aided deep reinforcement learning framework for latency-sensitive network slicing. Comput. Netw. **217** (2022). https://doi.org/10.1016/j.comnet.2022.109279

A Stochastic Model of a Passenger Transport Hub Operation Based on Queueing Networks

Alexander Kazakov[1], Giang Vu[2], and Maxim Zharkov[1](✉)

[1] Matrosov Institute for System Dynamics and Control Theory of Siberian Branch of Russian Academy of Sciences (IDSTU SB RAS), Irkutsk, Russia
zharkm@mail.ru

[2] Irkutsk National Research Technical University, Irkutsk, Russia
http://idstu.irk.ru,https://www.istu.edu/

Abstract. The paper concerns a methodology of mathematical modeling of passenger transport hub operation, which are significant elements of the transport infrastructure of a megalopolis. We use non-stationary and non-ordinary flows to describe the arrival of passengers on various modes of transport. We model the movement of passengers through the system using an open queueing network. The nodes' service parameters are time-dependent. Thus, the model considers the characteristics of passenger traffic from different transport modes, the hierarchical structure of the system, several traffic routes within it, and the fluctuation of transport schedules the day. To apply the methodology, we select two objects located in the capitals of Russia and Vietnam. We construct mathematical models, perform scenario simulations, and then estimate the current and maximum capacity and provide recommendations for improving performance based on the numerical results.

Keywords: mathematical model · queuing theory · BMAP · simulation · passenger transport hubs · passenger traffic

1 Introduction

Passenger transport hubs (further hubs) are important components of a megalopolis transport infrastructure. The efficiency of their operation determines the comfort and speed of transfers and, consequently, the overall quality of public transport services [1,2]. The average transfer time is a commonly used quantitative indicator of the hub's efficiency: less is better. Agent-based methods are predominant among a variety of modelling techniques for its evaluation. Dashamirov F. and Javadli U. [3] present the model of passenger movement in a hub. Yu J. (et al.) [4] describe the model of boarding and disembarking from a metro train. Lee E. (et al.) [5] evaluate the efficiency of transfers between bus routes within the hub's system. Article [6] investigates the capacity of public transport networks following failures. Agent-based modeling provides high-level

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 48–62, 2024.
https://doi.org/10.1007/978-3-031-65385-8_4

accuracy, details, and scalability for the resulting models. However, the process of studying the appropriate software and constructing models of a particular object is time-consuming. At the same time, the existing models are not always applicable to systems having different structures, and it is necessary to design new models.

One of the well-known alternative approaches to modeling the operation of hubs and their parts is the Queueing Networks (QNs) [7]. This mathematical apparatus can significantly reduce the complexity of model development and is efficient and versatile enough to describe different objects. Queueing Network (QN) models are used in [8,9] to evaluate the capacity of urban railway stations, in [10] to analyze passenger movements in the hub, and in [11] to investigate the passenger service process in the airport terminal before departure. Liu J. (et al.) [12] develop a model of the rail transport system (metro network and urban rail) based on QNs and identify the distribution of passenger delays in the hubs system.

We have successfully used queueing theory in our studies of the operation of transport systems. We present a railway terminal model in [13,14], and models for hubs based on metro stations in [15,16]. In contrast to the methods proposed in papers [8–11], we use QNs with BMAP flows [17]. It allows us to describe in a single model the arrival of passengers from different directions and to set parameters that depend on the mode of transport, particularly the distribution of passenger group sizes.

In this study, we improve the proposed methodology for modeling the operation of the hubs. We use QNs with non-stationary and non-ordinary flows that take into account the fluctuations of passenger arrival rates, and the service parameters of transport are time-dependent. Section 2 presents the subject description of the hub concept. Section 3 presents the methodology for the modeling of its operation. In Sect. 4, we construct operation models of two objects in Russia and Vietnam. Finally, in Sect. 5, we carry out the numerical experiment and discuss its results.

2 Subject Description

A passenger transport hub is a terminal that allows the distribution of passenger traffic between different modes and directions to reduce transfer times and increase the comfort and convenience of public transport [1,2].

Passenger traffic. The mode of transport determines the entry point of passengers into the system, the size of incoming passenger groups, and the time distribution between their arrivals. Metro trains run according to a schedule and can accommodate large groups. Buses run according to the traffic situation, and their timetable is not guaranteed. Private transport (cars) and taxis do not follow a schedule. Passengers may also walk from the surrounding area.

The structure of the system. A hub consists of public transport stations, a park-and-ride facility (parking), and a terminal for moving between the different stations and parking in a comfortable environment. The subsystems are autonomous and have distinct technical characteristics, including capacity, the number of limiting elements (such as turnstiles, doors, ticket offices, and lifts), the size of the passenger groups served, and the duration of the service.

The number of passenger routes depends on the number of stations in the PTH. We assume that the purpose of passengers is to transfer between different modes of transport via the terminal. Therefore, movements limited to one station of the “bus-bus” type are not taken into account.

3 Methodology

The mathematical model of the hub operation is designed in three stages. In the first stage, we describe the arrival of passengers at the stations. Second stage models the operation of the subsystems. Finally, we consider the passenger routes within the system.

Description of incoming passenger traffic. We consider transport as an external source of passenger flows and their destinations. We simulate the arrival of passengers from trains by a non-ordinary deterministic flow, where the time between arrival groups obeys the schedule. In other cases, we use a non-ordinary Poisson point process. The time between arrivals obeys the exponential distribution $exp(\lambda_i(t)/n_i(t))$, where $\lambda_i(t)$ is the intensity and $n_i(t)$ is the average group size.

If all incoming passenger flows travel in the same direction, they can be combined using a modified version of the BMAP, which is construct as follows. 1) Using statistical data, we determine the number of passenger flows W, the sum intensity of passenger group arrivals $\lambda(t)$ at time t, the distribution of arriving group sizes $f_v(x)$, which parameters are calculated by the method of moments, and the probability of arrival of a group $p(v)$ from the flow v. We assume that groups from different flows arrive independently, so $p(v)$ could be found as the ratio of the volume of passengers in the flow v to the total volume of all passenger flows. 2) The data obtained are applied for constructing a BMAP, which consists of a set of matrices $D_k(t), k = \overline{0, V}$, where V is the maximum size of the incoming passenger groups.

$$\begin{aligned} (D_0)_{v,v}(t) = -\lambda(t), (D_0)_{v,v'}(t) = \lambda(t)p(v')f_v(0), \\ (D_k)_{v,v'}(t) = \lambda(t)p(v')f_v(k), v, v' = \overline{1, W}, k = \overline{1, V}, \end{aligned} \tag{1}$$

Remark: If in (1), we replace $\lambda(t)$ with λ, $p(v')$ with $p(v, v')$ – the probability of a group of passengers arriving from the flow v' given the previous group arriving from the flow v, and $p(v, v')f_v(k) = p_k(v, v')$, then the formulas (1) coincide with those in [17] (see p. 65).

Description of the stations and the terminal operation. An open QN is used to simulate the operations of the hub subsystems. The QN consists of a finite number of S queueing systems (QS) or nodes. Requests enter the QN from an external source, which refers to a dummy node with index 0. The routes of the requests are determined by a stochastic matrix $P = ||P_{ij}||$, where P_{ij} is the probability of a transfer request from node i to node $j (i, j = \overline{0, S})$ [7].

In general, stations and terminals are described by several nodes, which are determined by the number of different passenger routes and limiting elements. Each node contains channels that simulate the operation of separate limiting elements. The maximum queue length is calculated based on the available space in front of this element, with a capacity of 2 people per square meter.

Description of passenger routes. Arriving requests are accepted according to the discipline of complete admission [17]. The route matrix P describes the itineraries, with transfer probabilities between nodes defined as relative frequencies. This ratio reflects the passenger flow entering the selected element (node) in relation to the total passenger traffic in that direction. The movement time between nodes is constant and calculated as the average time that it takes for passengers to move between the corresponding limiting elements. Its value is added to the average time the request (passenger) stays in the system. There is also feedback between the QN nodes, which consists of temporarily blocking the channels of the previous QS until enough space appears in the next QS for request receiving.

The purpose of the modeling is to determine the performance indicators, including the loss probability, the average sojourn time of a request in each node and in the system, the time of channel blocking, and some others. On the basis of these indicators, we draw conclusions about the efficiency of the operation.

4 Mathematical Models

Let us apply the methodology to describe the operation of two hubs. The first is located in Russia, and the other is in Vietnam. These hubs are selected for the following reasons. Firstly, they are typical and have similar structures. Secondly, statistics on passenger traffic and operations are available in open sources, and it is possible to carry out a comprehensive survey.

Sokolinaya Gora (SG) PTH, Moscow, Russia. The first level includes Moscow Central Circle station (MCC), the bus stop, and the parking. The transition between them is possible only through the second level, where the terminal and the above-ground passage are located (see Fig. 1).

The average daily passenger flow is 5.8 thousand people from the metro station and 6.1 thousand people in the opposite direction [19]. The hub's operating hours run from 5.30 a.m. to 01.00 a.m., with rush hours from 7.30 a.m. to 11.30 a.m. and 4 p.m. to 9 p.m. Table 1 shows the distribution of passengers in the system during the day, where t is the middle of the time intervals (one hour), and $d(t)$ is part of the total passenger flow.

Description of incoming passenger traffic. Table 2 shows the parameters of passenger flows obtained from field observations and open data. Further, $\mathrm{B}(a; b)$ represents the binomial distribution, a is the number of trials, and b is success probability.

Here is a brief overview of the key parameters in Table 2. The average size of the passenger groups arriving per hour is determined from the data in Table 1 using the formula $V_1(t) = 2900d(t)/n(t)$, where $n(t) = 15$ is the number of trains

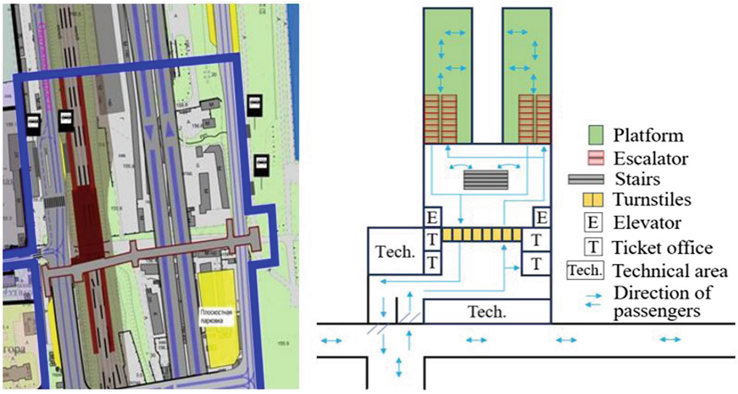


Fig. 1. Map and terminal scheme of the SG hub

Table 1. Distribution of the daily passenger traffic of the SG hub

t	**6:00**	**7:00**	**8:00**	**9:00**	**10:00**	**11:00 - 15:00**		**16:00**
d(t)	0.02	0.06	0.11	0.08	0.05	0.05 per hour		0.06
t	**17:00**	**18:00**	**19:00**	**20:00**	**21:00**	**22:00**	**23:00**	**23:30 - 01:00**
d(t)	0.085	0.11	0.07	0.05	0.02	0.02	0.01	0.005

in one hour during the rush period and $n(t) = 7,5$ for the rest of the time. The average number of incoming pedestrians per day equals $11900 \times 0.7 \times 0.35 = 2916$ people, where 11900 is the population density, 0.7 km is the walking distance, and 0.35 is the population who use a metro (35%, see [13]). The average size of the groups is 1.2 people, obtained from video camera data. Also, 238 cars enter the parking lot per day, with an average group size of 1.34 people (see [15]). These parameters and data from Table 1 are used in $\lambda_3(t)$ and $\lambda_4(t)$.

We model the passenger traffic from the MCC by non-ordinary deterministic flows and from the bus stops, parking, and pedestrians by the BMAP flow. Table 2 shows its parameters. The matrices are 3×3 in size, with a total number of 46, but are not shown.

Table 2. Models of incoming passenger flows at the SG hub

Flow	Types of transport	Rush hour	Left time	Group size
$D^X_{1.1}; D^X_{1.2}$	Subway (2 directions)	4 min	8 min	$\mathrm{B}(2V_1(t); 0.5)$
$BMAP_1$	Bus (per h.)	$\lambda_1 = 6.8$	$\lambda_2 = 7.7$	$\mathrm{B}(30; 0.66)$
	Pedestrians (per h.)	$\lambda_3(t) = 2916d(t)/1.2$		$\mathrm{B}(7; 0.179)$
	Parking (per h.)	$\lambda_4(t) = 238d(t)/1.34$		$\mathrm{B}(7; 0.195)$

Description of the system's operation. We consider the bus stop and the parking only as external sources of passenger traffic since they are located far (300 m or more) from the terminal, and passengers can depart from them freely. Additionally, we ignore the above-ground passage as it does not restrict the movement of passengers.

The MCC station consists of two separate platforms with an area of 1100 m.^{2}, from which trains depart in opposite directions (see 1). Passengers enter (and exit) the terminal from the platforms by escalator, staircase, or lift. We do not consider stairs as they do not restrict the movement of passengers. We allocate 50% of the platform capacity to passengers boarding the train, 40% and 10% to passengers entering the terminal by escalator and lift, respectively.

There are two ticket offices and two ticket vending machines in the terminal; six turnstiles to the MCC and two at the back; two doors each to the entrance and exit of the above-ground passage. We assume that the capacity of the above-ground passage at the terminal entrance is 500 people. We allocate 40% of the terminal hall's area (460 m.^{2}) to passengers traveling to and from trains and 20% to the ticket office. We split the paid area (340 m.^{2}) of the terminal into five sections: 30% in front of the turnstiles leading to the city, 20% each in front of the two escalators, and 15% each in front of the two lifts. Table 3 shows the average travel times between the limiting elements of the hub.

Table 3. Transfer time between limit elements at the SG hub

Elements of the hub	Mean transfer time ($\mathbf{T_d}$)
MCC train - escalator / elevator	0.5 min. / 0.75 min.
Escalator / elevator - turnstiles	0.33 min. / 0.25 min.
Turnstiles / ticket office - terminal doors	0.42 min.
Terminal doors - tunnel (to the city)	1.5 min

We model the operation of the SG hub using a QN with 15 nodes and three flows (QN 1). Nodes 1–6 simulate trains, escalators, and lifts on the MCC platform. Nodes 7–15 describe the limiting elements in the terminal. We model each lift by two nodes: 5 and 12, 6 and 13. The service time in them is doubled compared to the one-way travel time. The sizes of the groups of requests ($\boldsymbol{X}$) and the distributions of their service times in all nodes are obtained from the analysis of the field data. Table 4 shows this information and the formal description of nodes in terms of Queueing theory. Figure 2 shows the non-zero elements of the routing matrix P_1 as graph weights on the QN 1 scheme.

Hub is based at Thuong Dinh metro station (TD hub) in Hanoi, Vietnam. The structure of urban transport in Vietnam differs from that in Russia. Motorbikes and bicycles are the main types of vehicles in Hanoi. Therefore, 49% of passengers arrive at the hub by it. Pedestrians account for 38%, while buses and cars make up 10% and 3% of arrivals, respectively [21].

Table 4. Parameters of the QN 1

Nodes	Elements	Model	Service time	X
1 & 2	MCC trains	$*/D^X/1/1100$	8 min.; 4 min during rush hour	150
3 & 4	Escalators from MCC	$D^X/D/1/880$	0.1 min.	1
5 & 6	Elevators to the terminal	$D^X/D^X/1/220$	1 min.	8
7	Ticket offices	$*/M/4/90$	$exp(2)$ min.	1
8	Turnstiles to MCC	$*/M/6/185$	$exp(6)$ min.	1
9	Turnstiles from MCC	$*/M/2/100$	$exp(6)$ min.	1
10 & 11	Escalators to MCC	$*/D/1/70$	0.1 min	1
12 & 13	Elevators to MCC	$*/D^X/1/50$	1 min	8
14	Doors from the terminal	$*/M/2/185$	$exp(6)$ min.	1
15	Doors to the terminal	$BMAP/M/2/500$	$exp(6)$ min.	1

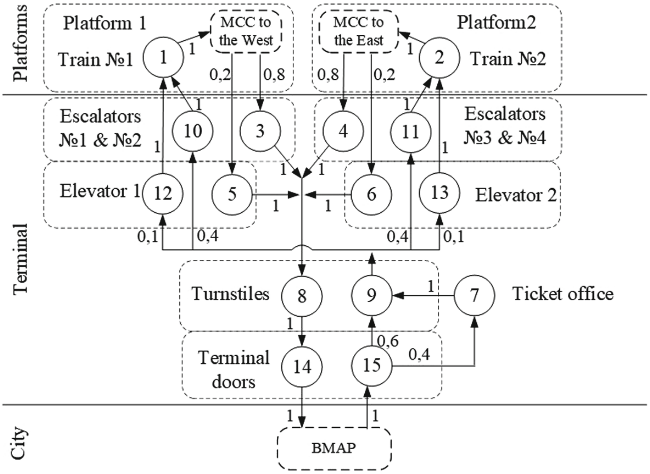


Fig. 2. Scheme of QN 1

The first level of the hub includes a bus stop. However, the pavements on both sides of the roads in Hanoi tend to turn into spontaneous motorbike parks. The largest of them form near this hub in particular. The second level is the terminal, and the third level is the metro station (see Fig. 3). The system operates between 5.30 a.m. and 10.00 p.m., with rush times from 7.00 a.m. to 8.30 a.m. and 4.30 p.m. to 6.00 p.m.

The average daily passenger flow is 2.1 thousand people from the metro station and 1.7 thousand people in the opposite direction. The data received was based on an analysis of field data. Table 5 shows ***t***, the mean of time intervals (one hour) and ***d(t)***, part of the total passenger flow.

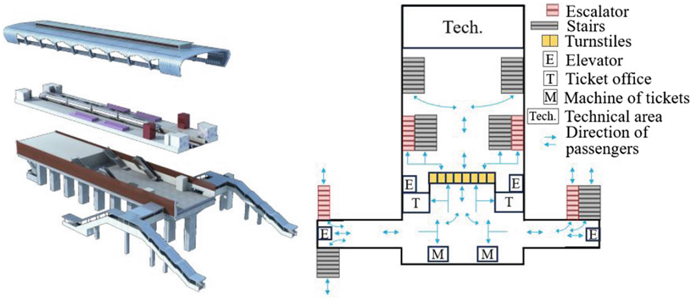


Fig. 3. Map and terminal scheme of the TD hub

Table 5. Distribution of the daily passenger traffic of the TD hub

t	6:00	7:00	8:00	9:00	10:00	11:00–14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00
d(t)	0.02	0.1	0.15	0.07	0.05	0.04	0.05	0.06	0.1	0.11	0.05	0.04	0.03	0.01

Parking is not fixed and occurs spontaneously, and the least number of passengers arrive from buses, so we consider these elements only as sources of passenger flows. Next, we construct the TD hub operation model by drawing parallels with the description of the SG hub.

Modeling incoming passenger traffic. Table 6 shows the data for passenger arrivals. We describe the arrival of passengers from the MCC as non-ordinary deterministic flows and from buses, parking, and pedestrians as a BMAP flow.

Table 6. Models of incoming passenger flows at the TD hub

Flow	Type of transport	Rush hour	Left time	Group size
$D^X_{2.1}; D^X_{2.2}$	Subway (2 directions)	6 min	10 min	B$(3400d(t)/n(t); 0.5)$
$BMAP_2$	Bus (per h.)	$\lambda_5 = 172 * d(t)/3$		B$(20; 0.1)$
	Pedestrians (per h.)	$\lambda_6(t) = 0.382 * 1700d(t)/3$		B$(10; 0.3)$
	Parking (per h.)	$\lambda_7(t) = 0.517 * 1700d(t)$		B$(3; 0.25)$

Description of the system's operation. We implicitly include the stairs in the model. It's the average travel time added to the transfer time at the metro station or the terminal. Table 7 shows the travel times connecting the limiting elements of the hub.

We describe the operation of trains and limiting elements by QN with 17 nodes (QN 2). Table. 8 shows its parameters in terms of Queueing theory, where $\boldsymbol{X}$ represents the sizes of the served request groups. Figure 4 shows the request transitions' probabilities between the nodes of QN 2 as graph weights.

Table 7. Transfer time between limit elements at the TD hub

Elements of the hub	Mean transfer time ($\mathbf{T_d}$)
Subway train - escalator (elevator) / stairs	0.5 min. / 1 min.
Escalator / elevator / stairs - turnstiles	0.17 min.
Turnstiles / ticket office - terminal doors	0.58 min

Table 8. Parameters of QN 2

Nodes	Elements	Model	Service time	X
1 & 2	Subway trains	$*/D^X/1/400$	10 min.; 6 min during rush hour	96
3 & 4	Elevators to the terminal	$D^X/D^X/1/400$	1 min.	8
5 & 6	Elevators to subway	$*/D^X/1/72$	1 min.	8
7 & 8	Escalators to subway	$*/D/1/96$	0.1 min.	1
9	Turnstiles from subway	$*/M/6/88$	$exp(6)$ min.	1
10	Turnstiles to subway	$*/M/6/144$	$exp(6)$ min.	1
11	Ticket offices	$*/M/4/44$	$exp(2)$ min.	1
12 & 13	Elevators to the city	$*/D^X/1/44$	1 min.	8
14 & 15	Elevators to the terminal	$BMAP/D^X/1/500$	1 min.	8
16 & 17	Escalators to the terminal	$BMAP/D/1/500$	0.1 min.	1

5 Computational Experiment

We numerically investigate QNs using a simulation model [15] based on the discrete-event simulation approach and Monte Carlo methods. The software is created in the Object Pascal programming language and can calculate performance indicators of a QN with up to 100 nodes and 20 request flows, including BMAP flows. We can set the inactivity time when requests do not arrive. It is necessary to take into account the operation time of the hub. We conducted experiments with various flow parameters for each obtained QN. The tables in this section show the average results for ten software runs. The virtual simulation time (run duration) is 24 h.

The tables below use the following notation. $\mathbf{T_{sys}}$ is the average transfer time within a hub; $\mathbf{T_c}$ and $\mathbf{T_m}$ are the average transfer times between the train and road in both directions; $\mathbf{T_d}$ is the average time at movement between adjacent hub elements; $\mathbf{T_l}$ refers to waiting in a queue; $\mathbf{T_n}$ refers to remaining at a node; $\mathbf{T_r} = \mathbf{T_d} + \mathbf{T_n}$ is movement to the next element of the hub along the route and maintenance in it, except for nodes 1 and 2 where $\mathbf{T_r} = \mathbf{T_d} + \mathbf{T_l}$. $\mathbf{T_d}$ is employed to determine $\mathbf{T_{sys}}$, $\mathbf{T_c}$, and $\mathbf{T_m}$ consistently. System performance indicators: $\boldsymbol{K}$ is the average number of load channels; $\boldsymbol{L}$ is the average length of the queue; $\mathbf{V_n}$ is the average number of requests arriving at the node in a day; $\boldsymbol{V(t)}$ is the

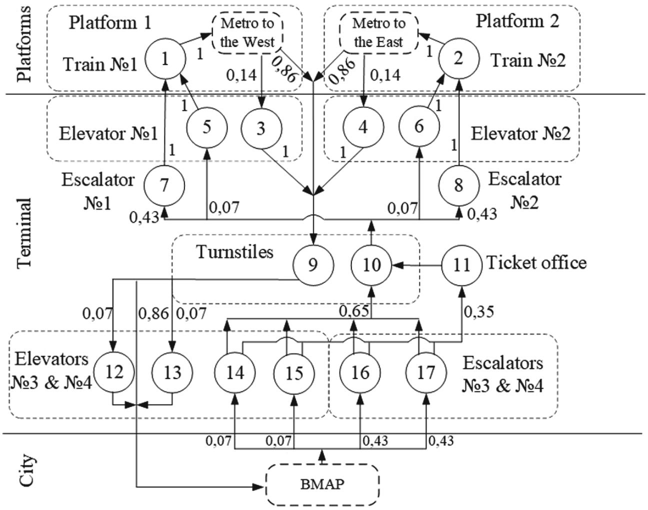


Fig. 4. Scheme of QN 2

number of all incoming requests per hour in the hub, $\boldsymbol{t}$ is the midpoints of the time intervals.

Experiment 1. A study of the Sokolinaya Gora hub model (Table 3). Tables 9 and 10 and Fig. 5 show the simulation results when the request flows correspond to the current daily passenger traffic. Table 11 shows the average performance indicators of QN 1 with increases in request volume of 20%, 30%, 40%, and 50%, corresponding to 14.4, 15.7, 16.8, and 19.5 thousand people per day.

Table 9. Number of incoming requests to the QN 1 per day

$\boldsymbol{t}$	**6:00**	**7:00**	**8:00**	**9:00**	**10:00**	**11:00**	**12:00**	**13:00**	**14:00**	**15:00**
$\boldsymbol{V(t)}$	246.4	663.5	1292.0	937.17	617.15	604.49	583.93	587.08	592.10	606.15
$\boldsymbol{t}$	**16:00**	**17:00**	**18:00**	**19:00**	**20:00**	**21:00**	**22:00**	**23:00**	**00:00**	**Sum**
$\boldsymbol{V(t)}$	700.0	989.15	1336.01	816.03	632.99	252.43	240.78	129.77	69.65	12021.0

Verification of the model. The average relative deviation $\varepsilon = \boldsymbol{(V(t)/12021 - h(t))/h(t)}$ is equal to 3.9%, where $\boldsymbol{V(t)/12021}$ is in Table 9 and $\boldsymbol{h(t)}$ is in Table 1. At 11:00, $\mathbf{T_c}$ and $\mathbf{T_m}$ differ from the field-observed transfer time, which is 5.5 min in the MCC-street direction and 8.1 min in the opposite direction, by 1% and 3.1%, respectively. Thus, the maximum relative deviation does not exceed 4%.

Interpretation of the modeling results (see Table 10 and Fig. 5). A transfer at the SG hub is considered comfortable if it does not exceed 15 min. The average

Table 10. Performance indicators of QN 1

$\mathbf{T_{sys}}$	$\mathbf{T_c}$	$\mathbf{T_m}$	Arrived	$D^X_{1.1}$	$D^X_{1.2}$	$BMAP_1$
6.70	5.56	7.85	**Requests**	2921.4	2903.6	6196.0
Parameters	$\boldsymbol{K}$	$\boldsymbol{L}$	$\mathbf{T_l}$	$\mathbf{T_n}$	$\mathbf{T_r}$	$\mathbf{V_n}$
Node 1	0.97	5.75	2.67	8.46	2.67	3093.48
Node 2	0.97	5.76	2.66	8.46	2.66	3101.74
Node 3	0.21	1.60	0.99	0.95	1.45	2478.64
Node 4	0.21	1.67	1.02	0.96	1.46	2468.82
Node 5	0.06	0.15	1.07	1.44	2.19	442.76
Node 6	0.06	0.15	1.06	1.39	2.14	434.82
Node 7	1.08	0.08	0.27	0.68	1.09	2479.28
Node 8	0.86	0.04	0.15	0.23	0.54	5824.56
Node 9	0.90	2.40	0.80	0.63	1.05	6195.46
Node 10	0.21	0.05	0.08	0.17	0.42	2472.10
Node 11	0.21	0.05	0.08	0.17	0.42	2482.06
Node 12	0.42	0.14	0.51	1.45	1.65	621.40
Node 13	0.42	0.14	0.51	1.44	1.64	619.78
Node 14	0.85	4.37	1.46	0.95	1.37	5824.50
Node 15	0.90	8.69	2.31	1.53	3.03	6195.98

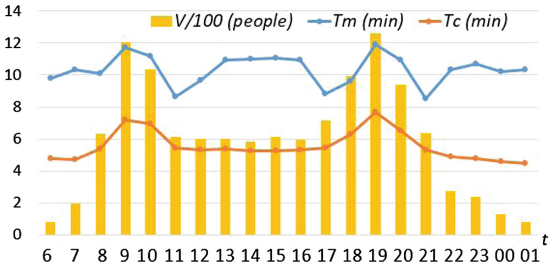


Fig. 5. Changes in $\mathbf{T_c}$ and $\mathbf{T_m}$ over the day (hourly)

Table 11. Performance indicators of QN 1 with an increase in the volume of the request flow

V	$\mathbf{T_c}$	$\mathbf{T_m}$	$\mathbf{T_{sys}}$	$\mathrm{T_c}$(rush)	$\mathrm{T_m}$(rush)	Nodes with max $\mathrm{T_n}$
+20%	8.51	14.06	11.28	14.19	19.55	**Nodes 14 & 15**
+30%	9.79	16.38	13.08	16.26	25.22	**Nodes 14 & 15**
+40%	13.51	19.83	16.67	24.03	30.36	**Nodes 14 & 15**
+50%	16.43	26.68	21.55	27.58	38.27	**Nodes 14 & 15**

transfer time is below this characteristic for the current passenger traffic. Passengers wait an average of 2.67 min ($\mathbf{T_l}$) at the MCC station (nodes 1 and 2). Crossing the terminal is comfortable and takes 3-4 min. If excluding nodes 1 and 2, then nodes 5 and 6 (elevators) and nodes 14 and 15 (terminal doors) have the highest values of $\mathbf{T_n}$. Elevators receive 10% of the total passenger traffic, so their capacity affects insignificantly on the system's efficiency as a whole. The terminal doors are the bottleneck due to the high average queue length in front of them.

When passenger numbers increase by up to 30% (as shown in Table 11), the average transfer time of day ($\mathbf{T_{sys}}$) is below 15 min. However, during rush hours, the transfer becomes uncomfortable. If passenger numbers increase by up to 50%, $\mathbf{T_c}$ is 28.9 min, $\mathbf{T_m}$ is 30.4 min, and the loss probability is non-zero (0.003). Across all cases, we observe the maximum values of average queue length in front of the terminal doors (nodes 14 and 15).

Experiment 2. A study of the Thuong Dinh hub model (see Table 3). Table 12 shows the performance indicators of QN 2 for one working day. Figure 6 shows their change during the day (per hour). Table 13 shows the performance indi-

Table 12. Performance indicators of QN 2

$\mathbf{T_{sys}}$	$\mathbf{T_c}$	$\mathbf{T_m}$	Arrived	$D^X_{2.1}$	$D^X_{2.2}$	$BMAP_2$
6.29	3.81	8.77	**Requests**	1054.24	1036.64	1673.12
Parameters	$\boldsymbol{K}$	$\boldsymbol{L}$	$\mathbf{T_l}$	$\mathbf{T_n}$	$\mathbf{T_r}$	$\mathbf{V_n}$
Node 1	0.98	3.64	4.32	12.06	4.82	836.14
Node 2	0.98	3.64	4.33	12.10	4.83	836.92
Node 3	0.03	0.04	0.99	1.01	1.51	153.68
Node 4	0.02	0.03	0.96	0.93	1.43	139.66
Node 5	0.11	0.01	0.53	1.21	1.38	114.64
Node 6	0.12	0.01	0.52	1.20	1.36	116.84
Node 7	0.07	0.01	0.07	0.15	0.72	721.52
Node 8	0.07	0.01	0.07	0.15	0.72	720.10
Node 9	0.35	0.09	0.10	0.27	1.15	2090.88
Node 10	0.29	0.00	0.04	0.17	0.75	1673.12
Node 11	0.29	0.00	0.15	0.52	1.10	580.28
Node 12	0.13	0.04	0.67	1.53	2.12	146.24
Node 13	0.13	0.04	0.66	1.54	2.12	149.84
Node 14	0.08	0.01	0.58	1.06	1.06	113.18
Node 15	0.08	0.01	0.52	1.07	1.07	118.84
Node 16	0.07	0.07	0.27	0.31	0.31	717.40
Node 17	0.07	0.07	0.27	0.32	0.32	723.70

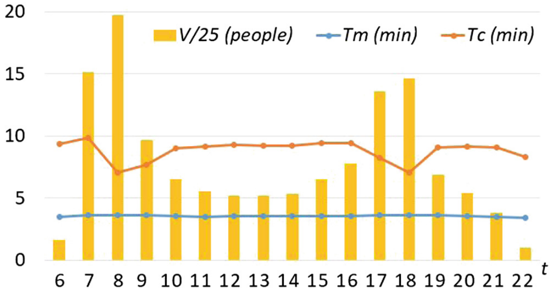


Fig. 6. Changes in $\mathbf{T_c}$ and $\mathbf{T_m}$ over the day (hourly)

Table 13. Performance indicators of QN 2

V	$\mathbf{T_c}$	$\mathbf{T_m}$	$\mathbf{T_{sys}}$	T_c(rush)	T_m(rush)	Nodes with max T_n
+50%	3.85	11.77	7.81	3.86	14.02	**Nodes 1 & 2**
+100%	3.87	12.16	8.02	3.88	15.89	**Nodes 1 & 2**
+200%	3.91	13.05	8.48	3.92	18.23	**Nodes 1 & 2**

cators of the model with a 50%, 100%, and 200% increase in application flows, corresponding to the receipt of 5.6, 7.5, and 11.3 thousand people per day.

Experiment 2 results interpretation. Passengers spend the most time moving from the street into the metro. Half of the time is spent by waiting for trains on the platform. The queue length does not exceed four people at any node, there is no blockage of channels, and the loss probability is zero. As a result, passengers move through the terminal without any hindrance.

As the daily number of passengers increases, we observe the following: First, $\mathbf{T_c}$ remains practically unchanged. The reason is that the most passengers use stairs, which are wide enough and do not restrict movement. When people move from the metro to the street, there is only one limiting element: six turnstiles (node 9), whose width is also sufficient. Secondly, the transfer is still comfortable when the number of passengers quadruples. In addition, passengers spend most of their time on the metro platform (nodes 1 and 2), waiting for the train.

Thus, both systems have a capacity margin. Passengers can comfortably transfer to Sokolinaya Gora hub with an increase in traffic of up to 30% and to Thuong Dinh hub with an increase of up to 200%. However, the train intervals in the Vietnamese metro are much longer than those in Russia and can be reduced in the future. In this scenario, the Thuong Dinh hub's capacity will be equal that of the Sokolinaya Gora hub and the transfer time will be less due to fewer limiting elements.

6 Conclusion

This paper continues the authors' previous research devoted to mathematical modeling and simulation of the hubs. We present the methodology for the operation modeling of such systems based on a specific type of Queuing Network. We use the modified BMAP flows to describe the arrival of passengers from different modes of transport, taking into account fluctuations in their volumes throughout the day. We also consider the departure of passengers in groups and the changes in transportation parameters. Therefore, the resulting models generalize our previous models and $M/G/C/C$ and $PH/PH/C/C$ queuing network models [9–11].

We consider the operation of the hubs in Russia and Vietnam. Both systems have a multi-tiered structure and a comparable design of the terminals and the metro stations, but differ in the incoming urban transport and the volume of passenger traffic. Through numerical modeling, we determined the current and maximum capacity for those systems. We also found bottlenecks in their structures and concluded about the efficiency of both hubs.

Further research can be related to studying a network of transport hubs. We have used a similar approach to model railway network operation [22,23], which proves to be effective in assessing capacity and long-term forecasting.

The research was funded by the Ministry of Education and Science of the Russian Federation within the framework of the project "Theoretical foundations, methods and high-performance algorithms for continuous and discrete optimization to support interdisciplinary research" (No. of state registration: 121041300065-9).

References

1. Williams, M., Seggerman, K.: Multimodal Transportation Best Practices and Model Element. National Center for Transit Research, Florida (2014)
2. Carmo, L.P.R.: Multimodal transport hubs. Good practice guidelines. AFD reprography service, France (2020)
3. Dashdamirov, F., Javadli, U.: Development of a methodology for creating an agent based model of transport hubs in suburban area. In: International Conference on Problems of Logistics, Management and Operation in the East-West Transport Corridor (PLMO), pp. 153–156 (2021)
4. Yu, J., Ji, Y., Gao, L., Gao, Q.: Optimization of metro passenger organizing of alighting and boarding processes: simulated evidence from the metro station in Nanjing. Sustainability **11**, 3682 (2019). https://doi.org/10.3390/su11133682
5. Lee, E., Patwary, A.U.Z., Huang, W., Lo, H.K.: Transit interchange discount optimization using an agent-based simulation model. Procedia Comput. Sci. **170**, 702–707 (2020). https://doi.org/10.1016/j.procs.2020.03.168
6. Thibaut, B., Amine, N.W., Changtao, Y., Juste, R.: An agent-based model for modal shift in public transport. Transp. Res. Procedia **62**, 711–718 (2022). https://doi.org/10.1016/j.trpro.2022.02.088
7. Bolch, G., Greiner, S., de Meer, H., Trivedi, K.S.: Queueing networks and Markov chains: modeling and performance evaluation with computer science applications. John Wiley & Sons Inc., New York (2006). https://doi.org/10.1002/0471791571

8. Zhu, J., Hu, L., Jiang, Y., Khattak, A.: Circulation network design for urban rail transit station using a PH(n)/PH(n)/C/C queuing network model. Eur. J. Oper. Res. **260**(3), 1043–1068 (2017). https://doi.org/10.1016/j.ejor.2017.01.030
9. Pan, H., Liu, Z.: A queuing network based optimization model for calculating capacity of subway station. Discr. Dyn. Nat. Soc. **2017**, 1–7 (2017). https://doi.org/10.1155/2017/4825802
10. Khattak, A., Hussain, A.: Hybrid DES-PSO framework for the design of commuters' circulation space at multimodal transport interchange. Math. Comput. Simul. **180**, 205-229 (2021). https://doi.org/10.1016/j.matcom.2020.08.025
11. Jarah, N.B., Saleh, S.S.: Model of series queues networks for passenger at Basra International Airport. Turkish J. Comput. Math. Educ. **12**(10), 3429–3435 (2021)
12. Liu, J., Hu, L., Xu, X.: A queuing network simulation optimization method for coordination control of passenger flow in urban rail transit stations. Neural Comput. Applic. **33**, 10935–10959 (2021). https://doi.org/10.1007/s00521-020-05580-5
13. Bychkov, I.V., Kazakov, A.L., Lempert, A.A., Bukharov, D.S., Stolbov, A.B.: An intelligent management system for the development of a regional transport logistics infrastructure. Autom. Remote. Control. **2**, 332–343 (2016)
14. Bychkov, I., Kazakov, A., Lempert, A., Zharkov, M.: Modeling of railway stations based on queuing networks. Appl. Sci. **11**, 2425 (2021). https://doi.org/10.3390/app11052425
15. Kazakov, A.L., Lempert, A.A., Zharkov, M.L.: A stochastic model of a transport hub and multi-phase queueing systems. Adv. Intell. Syst. Res. **158**, 117–123 (2018)
16. Zharkov, M.L., Kazakov, A.L., Lempert, A.A.: Transient process modeling in micrologistic transport systems. IOP Conf. Ser. Earth Environ. Sci. **629**, 012–023 (2021). https://doi.org/10.1088/1755-1315/629/1/012023
17. Dudin, A.N., Klimenok, V.I., Vishnevsky, V.M.: The Theory of Queuing Systems with Correlated Flows. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-32072-0
18. Alfa, N., Siti, M., Bertha, M.S.: Distribution pattern of public transport passenger in Yogyakarta. AIP Conf. Proc. **1941**, 020053 (2018). https://doi.org/10.1063/1.5028111
19. Passenger traffic at Moscow Metro stations. https://data.mos.ru/opendata/62743. Accessed 15 Dec 2023
20. Statistics in the St. Petersburg metro. https://kommet.ru/stats. Accessed 15 Dec 2023
21. Evaluating the level of satisfaction of passengers traveling by Cat Linh - Ha Dong train. https://tapchicongthuong.com.vn/bai-viet/danh-gia-muc-do-hai-long-cua-hanh-khach-tham-gia-giao-thong-bang-tau-dien-cat-linh-ha-dong-99759.htm. Accessed 15 Dec 2023
22. Kazakov, A., Lempert, A., Zharkov, M.: An approach to railway network sections modeling based on queuing networks. J. Rail Transport Planning Manage. **27**, 100404 (2023). https://doi.org/10.1016/j.jrtpm.2023.100404
23. Kazakov, A., Lempert, A., Zharkov, M.: Modeling a section of a single-track railway network based on queuing networks. In: Dudin, A., Nazarov, A. (eds) ITMM 2022. CCIS, vol. 1803, pp. 266–277 (2023). https://doi.org/10.1007/978-3-031-32990-6_4

Stochastic Analysis of a Discrete-Time Production Inventory Model Incorporating Perishable Items and Positive Service Time

Jijo Joy[1] and K. P. Jose[2](✉)

[1] St. Aloysius College Edathua, Alappuzha, Kerala 689 573, India
[2] St. Peter's College Kolenchery, Ernakulam, Kerala 682 311, India
kpjspc@gmail.com

Abstract. In this article, we examine a discrete time (s, S) production inventory system with positive service time and perishable products. We define a reasonable cost function for the model by assuming that the demand process is a Bernoulli process and that production and service time are geometric. This model is examined as a level independent quasi-birth-death process.We investigate the model using the Matrix Analytic Method. Furthermore, we use numerical experiments to determine the optimal (s, S) pair for the model across a range of parameter values.

Keywords: Bernoulli Process · Perishable Inventory · Production Process · Matrix Analytic Method

1 Introduction

Whether a company is a manufacturing one or a service one, inventory is an integral aspect of both. All business must maintain some inventory to keep things running smoothly. It is impossible for any company to assert that they do not maintain any inventory. Deteriorating items are widespread in our daily lives. Deteriorating items are those that deteriorate over time and become rotted, damaged, expired, devalued, and so on. Fruits, vegetables, and dairy products deteriorate with time due to variables such as moisture, temperature, and bacterial development. Consuming expired or damaged food might be hazardous to one's health.

Perishable inventory models are mathematical models used to optimise inventory management for perishable products that have a limited shelf life and can deteriorate or expire over time. These models help businesses determine how much perishable inventory to order, when to order it, and how to correctly

The authors acknowledge the financial support provided by FIST Program, Dept. of Science and Technology, Govt. of India, to the PG Research Dept. of Mathematics through SR/FST/College-2018-XA 276 (C).

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 63–75, 2024.
https://doi.org/10.1007/978-3-031-65385-8_5

assign it to avoid waste and increase profitability. Recent work by Munyaka and Yadavalli (2022) [9] emphasises the necessity of inventory management concepts and implementations in the face of increasingly demanding human needs. In the paper by Balagopal N. et al. (2021) [1] two discrete time models with positive service time and lead time were studied. By considering (s, S) policy and (s, Q) policy respectively they analyzed the system and derived the conditions for stability. Selvakumar et al. (2020) [11] addressed a discrete time service facility system with Bernoulli arrival under (s, S) policy. By assuming lead time and service time to follow geometric distributions they used Markov Decision Process to obtain the optimal policy to be implemented. Tan and Weng (2012) [14] studied a discrete time inventory control system where the deterioration and customer demand were in a constant rate. By allowing baclogs they succeeded in deriving a closed form solution for the optimal solution. Jose and Anilkumar(2020) [5] investigated a discrete time Geo/Geo/1 production inventory model with service time and local purchasing. They achieved a closed form solution for the steady state by setting the lead time to zero.

Inventory models with positive service time have first been examined by Sigman and Simchi-Levi(1992) [13]. Dhanya Shajin et al. (2018) [12] analyzed two discrete queueing inventory models with positive service time. By assuming that when inventory drops to zero, new customers are not allowed and the service process of queued customers also stops. They derived analytical relations for the basic stationary performance characteristics of the system.

Perishable-Inventory Control Models explicitly include perishability aspect in inventory management choices. It optimises order quantity and timing by taking into account parameters like perishability rates, shelf life, and demand patterns. The goal is to reduce both holding and shortage costs while minimising waste. One of the earliest works on inventory model with exponential decay was by Ghare and Schrader (1963) [3]. For perishable inventory models, Lian and Liu (1999) [7] used a discrete time model to approximate the corresponding continuous time counterpart. An inventory model of perishable products with constant deterioration rate and demand as a periodic function of time was proposed by Patil and Michra (2017) [10]. They succeeded in obtaining the average total cost per unit time of the system for two models by allowing shortages and the other without shortage.

Krishnamoorthy and Viswanath (2011) [6] studied a (s, S) production inventory system with positive service time and service interruption. They derived an explicit equation for the necessary and sufficient criteria for the stability of the system under consideration, assuming arrival to be a Poisson process, service time to be an Erlang distribution, and service interruption to be an exponential distribution. Nan Li et al. (2017) [8] developed an economic production quantity (EPQ) model with deteriorating production processes and deteriorating inventory. By allowing backlogs and finding the optimal total production and backlog quantity, they succeeded in minimizing the expected total cost per unit product. Chowdhury and Ghosh (2022) [2] investigated a production-inventory model with exponential demand whereas production rate and holding costs are linear functions of demand and time respectively. They attempted to find optimal production switch off and switch on time to avoid shortages. In the continuous review case Jose and Reshmi (2021) [4] considered a production inventory model with perishable items and retrial of customers. After constructing a suitable cost function they obtained the optimum control policy numerically. Yue and Qin (2019) [15] analysed an (s, S) production inventory with production vacation and service time. By assuming that arriving customers will be lost during stock out period, they derived the joint stationary distribution in product form of queue length and inventory level.

When dealing with perishable items, these models assist firms in making decisions about inventory management, ordering procedures, pricing tactics, and overall profitability. These models can reduce waste, improve customer service levels, and increase profitability for organisations in perishable industries by optimising inventory allocation and replenishment.

The rest of the article is arranged in the following way: Sect. 2 is on Mathematical Modelling and Analysis. Section 3 discusses stability and steady-state analysis. Section 4 dealt with performance measures and cost analysis. Section 5 included numerical and graphical illustrations.

2 Mathematical Modelling and Analysis

We investigate a discrete-time (s, S) production inventory model with perishable items and positive service time in this model. It is assumed that the arrival is a Bernoulli process with parameter a. Both service and manufacturing times are assumed to be geometric, with parameters b and c. Perishability is distributed geometrically with a parameter d. The consumer is not permitted to join the system when the inventory level is zero.

We examine a single-server approach in which each client receives only one inventory after the service is completed. Production begins when the inventory level falls to s owing to demand. When the inventory level hits S, the production process will be halted. The inventory level varies from 0 to $S-1$ during the manufacturing process.

Notations

$$X(m) : \text{Number of customers in the queue at an epoch } m.$$

$$Y(m) : \text{Inventory level at the epoch } m.$$

$$Z(m) : \text{The production status at an epoch } m.$$

$$\mathbf{e} : (1, 1, 1, ..., 1)', \text{column vector of 1's of size } 2S - s.$$

Then $\Psi = \{(X(m), Y(m), Z(m)); m = 0, 1, 2, 3, ..\}$ is a Quasi-Birth-Death Process on the state space $E = \{(i, j), i \geq 0, 0 \leq j \leq s\} \cup \{(i, j, k), i \geq 0, s+1 \leq j \leq S-1, k = 0, 1\} \cup \{(i, S), i \geq 0\}$.

Now the transition probability matrix of the process is

$$\mathcal{P} = \begin{array}{c|cccccc} & 0 & 1 & 2 & 3 & \cdots\cdot & \\ 0 & D_1 & D_0 & & & & \\ 1 & A_2 & A_1 & A_0 & & & \\ 2 & & A_2 & A_1 & A_0 & & \\ 3 & & & A_2 & A_1 & A_0 & \\ \vdots & & & & \ddots & \ddots & \ddots \end{array}$$

where the blocks D_0, D_1, A_0, A_1, A_2 are square matrices whose $(j,k)^{th}$ element with phase i is given below.

$$(D_0)_{jk} = \begin{cases} a\bar{c}d, & \text{for } 2 \le j \le s+1, k = j-1, \\ a(cd+\bar{c}\bar{d}), & \text{for } 2 \le j \le s+1, k = j, \\ ac\bar{d}, & \text{for } 2 \le j \le s, k = j+1, \\ ac\bar{d}, & \text{for } j = s+1, s+3, \cdots, 2S-s-3, k = j+2, \\ ac\bar{d}, & \text{for } j = 2S-s-1, k = 2S-s, \\ ad, & \text{for } j = s+2, k = s+1, \\ a\bar{c}d, & \text{for } j = s+3, k = s+1, \\ a\bar{d}, & \text{for } j = s+2, s+4, \cdots, 2S-s, k = j, \\ a(cd+\bar{c}\bar{d}), & \text{for } j = s+3, s+5, \cdots, 2S-s-1, k = j, \\ ad, & \text{for } j = s+4, s+6, \cdots, 2S-s, k = j-2, \\ a\bar{c}d, & \text{for } j = s+5, s+7, \cdots, 2S-s-1, k = j-2, \\ 0, & \text{otherwise.} \end{cases}$$

$$(D_1)_{jk} = \begin{cases} \bar{c}, & \text{for } j = 1, k = 1, \\ c, & \text{for } j = 1, k = 2, \\ \bar{a}\bar{c}d, & \text{for } 2 \le j \le s+1, k = j-1, \\ \bar{a}(cd+\bar{c}\bar{d}), & \text{for } 2 \le j \le s+1, k = j, \\ \bar{a}c\bar{d}, & \text{for } 2 \le j \le s, k = j+1, \\ \bar{a}c\bar{d}, & \text{for } j = s+1, s+3, \cdots, 2S-s-3, k = j+2, \\ \bar{a}c\bar{d}, & \text{for } j = 2S-s-1, k = 2S-s, \\ \bar{a}d, & \text{for } j = s+2, k = s+1, \\ \bar{a}\bar{c}d, & \text{for } j = s+3, k = s+1, \\ \bar{a}\bar{d}, & \text{for } j = s+2, s+4, \cdots, 2S-s, k = j, \\ \bar{a}(cd+\bar{c}\bar{d}), & \text{for } j = s+3, s+5, \cdots, 2S-s-1, k = j, \\ \bar{a}d, & \text{for } j = s+4, s+6, \cdots, 2S-s, k = j-2, \\ \bar{a}\bar{c}d, & \text{for } j = s+5, s+7, \cdots, 2S-s-1, k = j-2, \\ 0, & \text{otherwise.} \end{cases}$$

$$(A_0)jk = \begin{cases} a\bar{c}, & \text{for } j = 1, k = 1, \\ ac, & \text{for } j = 1, k = 2, \\ a\bar{b}\bar{c}d, & \text{for } 3 \le j \le s+1, k = j-1, \\ a\bar{b}(cd + \bar{c}\bar{d}), & \text{for } 2 \le j \le s+1, k = j, \\ a\bar{b}c\bar{d}, & \text{for } 2 \le j \le s, k = j+1, \\ a\bar{b}c\bar{d}, & \text{for } j = s+1, s+3, \cdots, 2S-s-3, k = j+2, \\ a\bar{b}c\bar{d}, & \text{for } j = 2S-s-1, k = 2S-s, \\ a\bar{b}d, & \text{for } j = s+2, k = s+1, \\ a\bar{b}\bar{c}d, & \text{for } j = s+3, k = s+1, \\ a\bar{b}\bar{d}, & \text{for } j = s+2, s+4, \cdots, 2S-s, k = j, \\ a\bar{b}(cd + \bar{c}\bar{d}), & \text{for } j = s+3, s+5, \cdots, 2S-s-1, k = j, \\ a\bar{b}d, & \text{for } j = s+4, s+6, \cdots, 2S-s, k = j-2, \\ a\bar{b}\bar{c}d, & \text{for } j = s+5, s+7, \cdots, 2S-s-1, k = j-2, \\ 0, & \text{otherwise.} \end{cases}$$

$$(A_2)_{jk} = \begin{cases} \bar{a}b\bar{c}d, & \text{for } 3 \le j \le s+1, k = j-2, \\ \bar{a}b(cd + \bar{c}\bar{d}), & \text{for } 2 \le j \le s+1, k = j-1, \\ \bar{a}bc\bar{d}, & \text{for } 2 \le j \le s, k = j+1, \\ \bar{a}bd, & \text{for } j = s+2, k = s, \\ \bar{a}bd, & \text{for } j = s+4, k = s+1, \\ \bar{a}b\bar{c}d, & \text{for } j = s+3, k = s, \\ \bar{a}b\bar{c}d, & \text{for } j = s+5, k = s+1, \\ \bar{a}b\bar{d}, & \text{for } j = s+2, k = s+1, \\ \bar{a}b(cd + \bar{c}\bar{d}), & \text{for } j = s+3, k = s+1, \\ \bar{a}bc\bar{d}, & \text{for } j = s+3, s+5, \cdots, 2S-s-3, k = j, \\ \bar{a}b\bar{d}, & \text{for } j = s+4, s+6, \cdots, 2S-s, k = j-2, \\ \bar{a}b(cd + \bar{c}\bar{d}), & \text{for } j = s+5, s+7, \cdots, 2S-s-1, k = j-2, \\ \bar{a}bd, & \text{for } j = s+6, s+8, \cdots, 2S-s, k = j-4, \\ \bar{a}b\bar{c}d, & \text{for } j = s+7, s+9, \cdots, 2S-s-1, k = j-4, \\ 0, & \text{otherwise.} \end{cases}$$

$$(A_1)_{jk} = \begin{cases} \bar{a}\bar{c}, & \text{for } j = 1, k = 1, \\ \bar{a}c, & \text{for } j = 1, k = 2, \\ ab\bar{c}d, & \text{for } 3 \leq j \leq s+1, k = j-2, \\ ab(cd + \bar{c}\bar{d}) + \bar{c}d, & \text{for } j = 2, k = 1, \\ ab(cd + \bar{c}\bar{d}) + \bar{a}\bar{b}\bar{c}d, & \text{for } 3 \leq j \leq s+1, k = j-1, \\ \bar{a}\bar{b}(cd + \bar{c}\bar{d}) + abc\bar{d}, & \text{for } 2 \leq j \leq s+1, k = j, \\ \bar{a}\bar{b}c\bar{d}, & \text{for } 2 \leq j \leq s, k = j+1, \\ \bar{a}\bar{b}c\bar{d}, & \text{for } j = s+1, s+3, \cdots, 2S-s-3, k = j+2, \\ \bar{a}\bar{b}c\bar{d}, & \text{for } j = 2S-s-1, k = 2S-s, \\ abd, & \text{for } j = s+2, k = s, \\ abd, & \text{for } j = s+4, k = s+1, \\ ab\bar{c}d, & \text{for } j = s+3, k = s, \\ ab\bar{c}d, & \text{for } j = s+5, k = s+1, \\ ab\bar{d} + \bar{a}\bar{b}d, & \text{for } j = s+2, k = s+1, \\ ab(cd + \bar{c}\bar{d}) + \bar{a}\bar{b}\bar{c}d, & \text{for } j = s+3, k = s+1, \\ \bar{a}\bar{b}\bar{d}, & \text{for } j = s+2, s+4, \cdots, 2S-s, k = j, \\ \bar{a}\bar{b}(cd + \bar{c}\bar{d}) + abc\bar{d}, & \text{for } j = s+3, s+5, \cdots, 2S-s-1, k = j, \\ ab\bar{d} + \bar{a}\bar{b}d, & \text{for } j = s+4, s+6, \cdots, 2S-s, k = j-2, \\ ab(cd + \bar{c}\bar{d}) + \bar{a}\bar{b}\bar{c}d, & \text{for } j = s+5, s+7, \cdots, 2S-s-1, k = j-2, \\ abd, & \text{for } j = s+6, s+8, \cdots, 2S-s, k = j-4, \\ ab\bar{c}d, & \text{for } j = s+7, s+9, \cdots, 2S-s-1, k = j-4, \\ 0, & \text{otherwise.} \end{cases}$$

3 Stability and Steady-State Analysis

When the inventory is 0 in this model, no customer is allowed to join the system. As a result, the inventory level is unaffected by queue length. The stability of the proposed queueing inventory model is the same as that of a regular discrete-time Geo/Geo/1 queue, with inter-arrival time and service time geometrically distributed with parameters a and b, respectively. If $a < b$, the system is stable.

Assume that the QBD is aperiodic and positive recurrent. Denote by π its stationary probability vector. It is the unique solution of the system $\pi P = \pi$ and $\pi \mathbf{e} = 1$, where $\mathbf{e}$ is a column vector of ones of appropriate order.

Let π be partitioned by levels as $\pi = (\pi_{\mathbf{0}}, \pi_{\mathbf{1}}, \pi_{\mathbf{2}}, \pi_{\mathbf{3}} \ldots,)$. Then π_i has the matrix geometric form $\pi_i = \pi_1 R^{i-1}, i \geq 2$. Where R is the minimal non negative solution of the matrix quadratic equation

$$R^2 A_2 + RA_1 + A_0 = R$$

. Where π_0 and π_1 are obtained by solving the equations

$$\pi_0(D_1 - I) + \pi_1 A_2 = 0$$

and

$$\pi_0 D_0 + \pi_1(A_1 + RA_2 - I) = 0$$

with the normalizing condition

$$\pi_{\mathbf{0}}\mathbf{e} + \pi_1(I - R)^{-1}\mathbf{e} = 1$$

where $\mathbf{e}$ is a column vector of 1's of size $(2S - s)$.

4 Performance Measures and Cost Analysis

We, partition the components of π as $\pi = (\pi_0, \pi_1, \pi_2 \ldots)and\pi_{\mathbf{i}} = (\pi_{i,0}, \pi_{i,1} \ldots, \pi_{i,s}, \pi_{i,s+1,0}, \pi_{i,s+1,1} \ldots, \pi_{i,S-1,0}, \pi_{i,S-1,1}, \pi_{i,S})$. The performance measures of the system under steady state are

1. Expected number of customers in the system, ENC, is given by

$$ENC = \sum_{i=1}^{\infty} i\pi_{\mathbf{i}}\mathbf{e}$$

2. Expected inventory level, EIL, is given by

$$EIL = \sum_{i=0}^{\infty}\sum_{j=0}^{s} j\pi_{i,j} + \sum_{i=0}^{\infty}\sum_{j=s+1}^{S-1} j[\pi_{i,j,0} + \pi_{i,j,1}] + \sum_{i=0}^{\infty} S\pi_{i,S}$$

3. Expected production rate, EPR, is given by

$$EPR = c\left[\sum_{i=0}^{\infty}\sum_{j=0}^{s} \pi_{i,j} + \sum_{i=0}^{\infty}\sum_{j=s+1}^{S-1} \pi_{i,j,1}\right]$$

4. Expected loss rate of customers when the inventory level is zero, ELR, is given by

$$ELR = a\sum_{i=0}^{\infty} \pi_{i0}$$

5. Expected rate at which production is switched on, EON, is given by

$$EON = (b\bar{d} + d)\sum_{i=0}^{\infty} \pi_{i,s+1,0} + bd\sum_{i=0}^{\infty} \pi_{i,s+2,0}$$

6. Expected rate of departure after completing the service,ERD, is given by

$$ERD = b\left[\sum_{i=1}^{\infty}\sum_{j=1}^{s}\pi_{i,j} + \sum_{i=1}^{\infty}\sum_{j=s+1}^{S-1}\left[\pi_{i,j,0}+\pi_{i,j,1}\right] + \sum_{i=1}^{\infty}\pi_{iS}\right]$$

7. Expected perishability rate, ERP, is given by

$$ERP = d\left[\sum_{i=0}^{\infty}\sum_{j=1}^{s}\pi_{i,j} + \sum_{i=0}^{\infty}\sum_{j=s+1}^{S-1}\left[\pi_{i,j,0}+\pi_{i,j,1}\right] + \sum_{i=0}^{\infty}\pi_{iS}\right]$$

8. Expected perishable Quantity, EPQ, is given by

$$EPQ = d\left[\sum_{i=0}^{\infty}\sum_{j=1}^{s}j\pi_{i,j} + \sum_{i=0}^{\infty}\sum_{j=s+1}^{S-1}j\left[\pi_{i,j,0}+\pi_{i,j,1}\right] + \sum_{i=0}^{\infty}S\pi_{iS}\right]$$

Define the expected total cost of the system per unit time as

$$ETC = c_1(EON) + c_2(EIL) + c_3(EPR) + c_4(ELR) + c_5(EPQ)$$

where,

c_1 : fixed cost for starting a production run
c_2 : holding cost of inventory/unit/unit time
c_3 : cost of production per inventory
c_4 : cost incurred due to loss of customers
c_5 : cost due to perishability of item

5 Numerical and Graphical Illustrations

The tables below show the numerical findings obtained for various performance measures in relation to the various parameters studied. The following diagram depicts the relationship between several performance measures and parameters.

5.1 Effect of *a* on Various Performane Measures

When $b = 0.55, c = 0.4, d = 0.5, s = 2, S = 6, c_1 = 1, c_2 = 4.5, c_3 = 2, c_4 = 2, c_5 = 2$, Table 1 indicates that when the expected arrival rate rises, so does the expected production rate and customer loss rate, but the expected production switch-on rate, inventory level, and perishable quantity decline.

5.2 Effect of *b* on Various Performane Measures

When $a = 0.25, c = 0.3, d = 0.4, s = 4, S = 8, c_1 = 5, c_2 = 1, c_3 = 1, c_4 = 14, c_5 = 1$, Table 2 demonstrates that when the expected service rate rises, so does the expected production rate and customer loss rate, but the expected production switch-on rate, inventory level, and perishable quantity fall.

Table 1. Effect of a on the model.$b = 0.55, c = 0.4, d = 0.5, s = 2, S = 6, c_1 = 1, c_2 = 4.5, c_3 = 2, c_4 = 2, c_5 = 2$

a	EON	EIL	EPR	ELR	EPQ	ETC
0.21	0.0007821	0.53376	0.39882	0.11684	0.26688	3.9678
0.22	0.0006897	0.53089	0.39896	0.12259	0.26544	3.96337
0.23	0.0005389	0.52738	0.39919	0.12837	0.26369	3.9562
0.24	0.0003838	0.52409	0.39942	0.13413	0.26204	3.9500
0.25	0.0002567	0.52150	0.39962	0.13986	0.26075	**3.9475**
0.26	0.0001662	0.51971	0.39975	0.14555	0.25986	3.9492
0.27	0.0001077	0.51857	0.39984	0.15121	0.25928	3.9543
0.28	0.0000724	0.51789	0.39989	0.15685	0.25894	3.9619
0.29	0.0000521	0.5175	0.39992	0.16248	0.25875	3.9711

Table 2. Effect of b on the model.$a = 0.25, c = 0.3, d = 0.4, s = 4, S = 8, c_1 = 5, c_2 = 1, c_3 = 1, c_4 = 14, c_5 = 1$

b	EON	EIL	EPR	ELR	EPQ	ETC
0.26	0.0000309	0.70372	0.299945	0.12689	0.28149	3.06178
0.27	0.0000266	0.68768	0.299953	0.12838	0.27507	3.06010
0.28	0.0000228	0.67239	0.299960	0.12982	0.26896	3.05892
0.29	0.0000196	0.65780	0.299966	0.13123	0.26312	3.058172
0.30	0.0000169	0.64387	0.299971	0.13260	0.25755	**3.05782**
0.31	0.0000145	0.63054	0.299976	0.13393	0.25222	3.05783
0.32	0.0000125	0.61779	0.299979	0.13523	0.24711	3.05816
0.33	0.0000108	0.60556	0.299982	0.13650	0.24223	3.05879
0.34	0.0000093	0.59384	0.299985	0.13773	0.23754	3.05968
0.35	0.0000081	0.58259	0.299987	0.13894	0.23304	3.06081

5.3 Effect of c on Various Performane Measures

When $a = 0.25, b = 0.55, d = 0.4, s = 4, S = 8, c_1 = 5, c_2 = 1, c_3 = 1, c_4 = 14, c_5 = 1$, Table 3 demonstrates that when the production rate rises, the expected customer loss rate falls, but the expected production switch-on rate, inventory level, and perishable quantity rise.

We generate ETC graphs by altering the parameters a, b, and c while holding the other parameters constant. These graphs show that the company's expected total cost will be as stated in the graphs in the long run. The projected total cost and its related values of a, b, and c can be determined. Figures 1, 2, and 3 have minimum values of $(0.25, 3.9475)$, $(0.30, 3.05782)$, and $(0.36, 3.0883)$, respectively.

Table 3. Effect of c on the model.$a = 0.25, b = 0.55, d = 0.4, s = 4, S = 8, c_1 = 5, c_2 = 1, c_3 = 1, c_4 = 14, c_5 = 1$

c	EON	EIL	EPR	ELR	EPQ	ETC
0.32	0.0000538	0.45823	0.31993	0.15263	0.18329	3.0985
0.33	0.0000874	0.47552	0.32988	0.14987	0.19021	3.0942
0.34	0.0001363	0.49367	0.33981	0.14709	0.19747	3.0909
0.35	0.0002037	0.51285	0.3497	0.1443	0.20514	3.0889
0.36	0.0002912	0.53319	0.35956	0.14149	0.21327	**3.0883**
0.37	0.0003977	0.55471	0.36939	0.13865	0.22188	3.0891
0.38	0.0005175	0.57728	0.37918	0.1358	0.23091	3.0911
0.39	0.0006407	0.60060	0.38896	0.13293	0.24024	3.0940
0.40	0.0007536	0.62422	0.39874	0.13006	0.24969	3.0972
0.41	0.0008416	0.64764	0.40856	0.1272	0.25905	3.1003

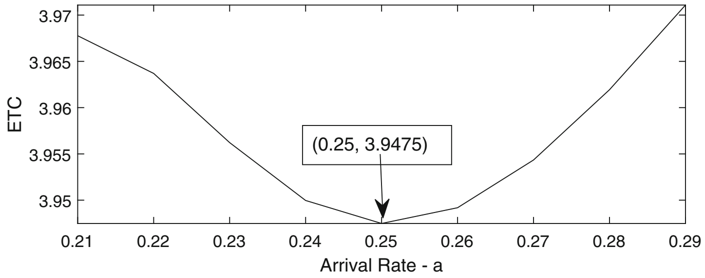


Fig. 1. a vs ETC, $b = 0.55, c = 0.4s = 5, S = 10, m = 15$

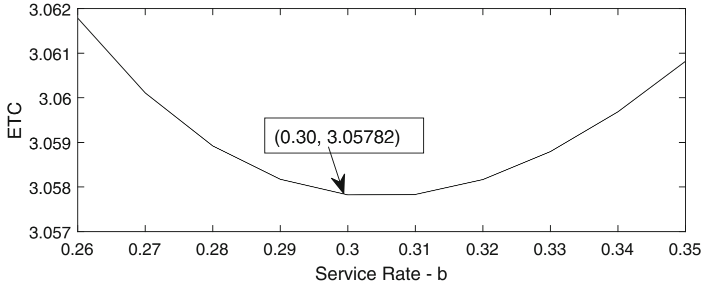


Fig. 2. b vs ETC, $a = 0.4, c = 0.4s = 5, S = 10, m = 15$

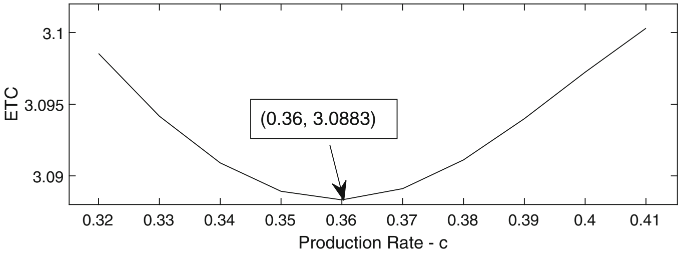


Fig. 3. c vs ETC, $a = 0.4, b = 0.5s = 5, S = 10, m = 15$

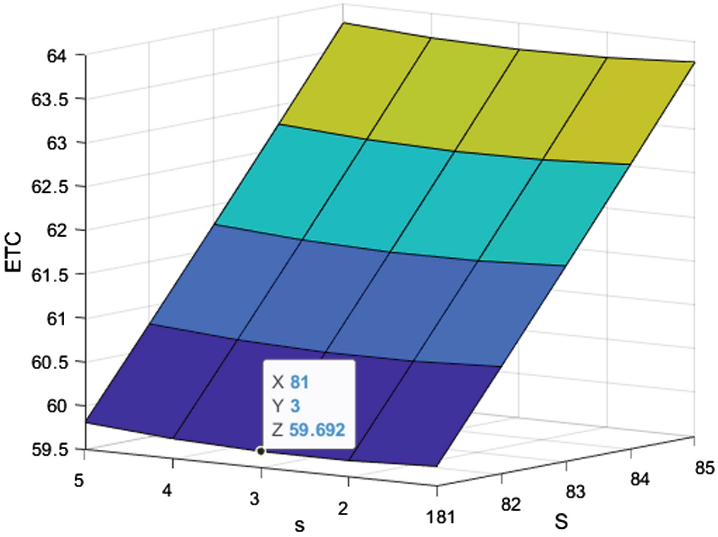


Fig. 4. a vs ETC, $b = 0.55, c = 0.4s = 5, S = 10, m = 15$

5.4 (s, S) Pair

We find different ETC values by altering parameter values. By adjusting the reorder point s and maximum inventory S, one can discover the (s, S) pair of the total cost. The Mathlab programming language is used to do the numerical analysis of the current model. The optimal (s, S) pair for the fixed parameter values indicated below is presented in table below (Fig. 4 and Table 4).

Table 4. (s, S) pair of the model for fixed parameter values $a = 0.35, b = 0.65, c = 0.45, d = 0.15, c_1 = 150, c_2 = 12.5, c_3 = 1, c_4 = 36.5, c_5 = 1$

s S	81	82	83	84	85
1	59.7334	60.7256	61.7286	62.7424	63.7669
2	59.6922	60.6812	61.6810	62.6916	63.7130
3	**59.6920**	60.6779	61.6746	62.6820	63.7001
4	59.7318	60.7146	61.7081	62.7123	63.7271
5	59.8111	60.7907	61.7810	62.7820	63.7936

6 Concluding Remarks

In this article, we studied a discrete (s, S) perishable inventory system with positive service times and production of items. The primary demand constitutes a geometric process. Service time follows a geometric distribution. We analyzed the system using the Matrix Analytic Method. Some important system performance measures are derived. The expression for expected total cost is also obtained. One can modify this model by allowing backlog when the inventory level is zero.

References

1. Balagopal, N., et al.: Comparison of discrete time inventory systems with positive service time and lead time. Korean J. Math **29**(2), 371–386 (2021)
2. Chowdhary, R.R., Ghosh, S.K.: A production-inventory model for perishable items with demand dependent production rate, shortages and variable holding cost. Int. J. Procure. Manage. **15**(3), 424–446 (2022)
3. Ghare, P.M., Schrader, G.P.: A model for an exponentially decaying inventory. J. Indust. Eng. **14**, 238–243 (1963)
4. Jose, K.P., Reshmi, P.S.: A production inventory model with deteriorating items and retrial demands. Opsearch **58**(1), 71–82 (2021)
5. K. P., Jose, M. P. Anilkumar: Stochastic optimization of local purchase quantities in a Geo/Geo/1 production inventory system. DCCN **CCIS 1337**, 206–220 (2020)
6. Krishnamoorthy, A., Viswanath, C.N.: Production inventory with service time and vacation to the server. IMA J. Manag. Math. **22**(1), 33–45 (2011)
7. Lian, Z., Liu, N.: A discrete time model for perishable inventory systems. Ann. Oper. Res. **87**, 103–116 (1999)
8. Li, Nan., et al.: A stochastic production inventory model in a two-state production system with inventory deterioration, rework process, and backordering. Trans. Syst., Man Cybern. Syst. **47**(6), 1–11 (2017)
9. Munyaa, J.B., Yadavalli, V.S.S.: Inventory management concepts and implementations: a systematic review. S. Afr. J. Ind. Eng. **33**(2), 15–36 (2022)
10. Patil, P., Mishra, P. N.: Inventory model of perishable products with periodic demand. IJESC **7**(11) (2017)
11. Selvakumar, C., et al.: A discrete markov decision process-inventory control in a service facility system. Revista Investigacion Operacional **41**(6), 872–881 (2020)

12. Shajin, D., et al.: Discrete product inventory control system with positive service time and two operation modes. Autom. Remote. Control. **79**(9), 1593–1608 (2018)
13. Sigman, K.: Simchi-Levi D: light traffic heuristic for an M/G/1 queue with limited inventory. Ann. Oper. Res. **40**(1), 371–380 (1992)
14. Tan, Y., Weng, M.X.: A discrete-in-time inventory model with deterioration and backlog. Int. J. Serv. Oper. Inf. **7**(1), 37–51 (2012)
15. Yue, D., Qin, Y.: A production inventory system with service time and production vacations. J. Syst. Sci. Syst. Eng. **28**(1), 168–180 (2019)

Identification of Network Traffic Using Neural Networks

Daria Salimzyanova(✉) and Ekaterina Lisovskaya

National Research Tomsk State University, Tomsk, Russia
bugakovadaria@inbox.ru, ekaterina_lisovs@mail.ru

Abstract. The solution of the problem of identifying network traffic will allow operators and infrastructure owners to make decisions about the strategy of its service, as well as predict its behavior in the future. This solution will help in the design of virtual or physical networks. In this paper, the problem of identifying several processes (stationary Poisson, MMPP, renewal) is solved using neural networks. First, the classification subtask is solved, which allows to tell which process model the input data correspond to. And, secondly, the subtask of parameter estimation is solved, which allows to tell which values of its parameters better describe the input data. Fully connected and recurrent neural networks are considered as tools. Predictive ability was assessed using classical metrics of regression and classification tasks, as well as using additional metrics.

Keywords: Stochastic process · Network traffic · Neural networks

1 Introduction

The motivation of this project is to help mathematical scientists and engineers join strengths in the study of communication networks. We see that, on the one hand, mathematicians (we mean mainly specialists in queue theory) are extensively researching complex models of queueing systems and queueing networks, obtaining huge formulas that, to be honest, could simplify the work of engineers. On the other hand, engineers who know all the features of communication networks and can describe existing traffic service strategies or propose new ones, but cannot work them out thoroughly, since the problem of traffic identification has not been resolved yet. We mean choosing a good mathematical model of the process and evaluating its parameters, i.e. such data, which need to describe queue. Therefore, currently we can see a lot of papers on the study of hard queue and any network technologies that use the Poisson process and exponential distributions. Moreover, the Erlang model is still often used.

The tasks of identifying network traffic have been exciting the minds of scientists for many years. The fact is that it is still unclear how to identify real traffic. That is, which mathematical process model should be used in each specific case and with what parameters. We managed to find several papers that are studied

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 76–90, 2024.
https://doi.org/10.1007/978-3-031-65385-8_6

to traffic identification, but in a slightly different sense: in all papers, the authors classify packets by type (IoT, video, etc.). Next, we will talk a little about this.

Due to the widespread use of neural networks, scientists began trying to use them to solve problems in telecommunications networks. For example, in papers [1–3], the authors use deep networks to solve problems of classifying network traffic by type (audio, email, etc.). This can lead to more efficient routing later on and therefore minimize latency.

The paper [4] uses real data from RedIRIS (the Spanish academic and research backbone network providing advanced communication services to the scientific community and national universities). Convolutional (CNN) and Recurrent (RNN) Neural Networks have been used as neural network models for traffic classification. The uniqueness of this work lies in the successful use of CNN to classify Internet traffic, also the work demonstrates the performance of a combination of CNN and RNN networks. Also convolutional neural network architectures in the traffic classification task have been extensively reviewed in the works of W. Wang [5,6]. The paper [7] compares various popular CNN architectures: LeNet-5 [8], AlexNet [9], GoogleNet [10].

A method based on the use of ensembles of neural networks and decision trees were used in the paper [11]. The dataset for the study was collected by browsing the most visited HTTPs sites twice a day in the browsers Google Chrome and Mozilla Firefox [12]. Similar models have been used in [13–15]. In the paper [16] the classification of two types of traffic (IoT- and video-traffic) is considered, recurrent neural networks (LSTM and RNN) are used. The approach, which is based on the use of unsupervised learning methods, was outlined in the paper [17]. Author uses a k-means clustering algorithm that allows to identify applications and group them into a new cluster.

The paper [18] gives a big overview of the research that is carried out within the framework of telecommunications networks using neural networks. Of course, at now this paper is slightly outdated, nevertheless, it gives us an absolutely clear understanding that many tasks of telecommunication networks are of huge practical importance, and can be solved using neural networks.

So, the goal of a large project is to develop a framework that will allow to select the best stochastic process model in real time by the packets arrive moments, as well as evaluate its parameters. This will allow real-time management of network traffic service. In addition, it will allow engineers not to simplify their models when researching any new technologies. This paper is part of a large project and demonstrates the currently results.

The idea of the framework is as follows: (i) take a set number of event time moments (packets arriving) as input, (ii) use the classification subtask to determine the mathematical model of the traffic, (iii) use the regression subtask to estimate the parameters of the corresponding distributions.

In short, we taught the neural network to classify three process models on simulated data (Poisson, MMPP, renewal) and evaluate its parameters. We will have to work on the metrics, as the existing ones cannot visualize the results well. Of course, we need to add the ability to differentiate between other process models as well. And then we need to test the results on real data.

2 Data Set

The samples for training are obtained by simulation [19]. Its sizes and modelling parameters are given in Table 1.

Table 1. Parameter values

Parameter name	Parameter value, classification	Parameter value, parameters estimation
Poisson		
Number of moments	500	1000
Number of samples	5000	10000
Intensity values, λ	$\lambda = unif(0, 10000)$	$\lambda = unif(0, 10000)$
MMPP		
Number of moments	500	1500
Number of samples	5000	10000
Number of states, K	$K = unif(2, 10)$	$K = unif(2, 10)$
Values of $\mathbf{Q}$	$q_{i,j} = unif(0, 10000)$	$q_{i,j} = unif(0, 10000)$
Values of $\mathbf{\Lambda}$	$\lambda_i = unif(0, 10000)$	$\lambda_i = unif(0, 10000)$
Renewal		
Number of moments	500	3000
Number of samples	5000	10000
Shape values, α	$\alpha = unif(0, 10000)$	$\alpha = exp(100)$
Rate values, β	$\beta = unif(s - sq, s + sq)$, $s = \frac{1}{\alpha * 5000}$, $sq = \frac{1}{max(\alpha) * 5000}$	$\beta = unif(0, 100)$

For the *classification* subtask:

- the dataset contain the event time moments (t_i) and process labels (stationary Poisson/MMPP/renewal),
- the dataset was splitted into train and test datasets in a ratio of 70/30 (10500/4500),
- from the train dataset, 25% of the samples (2625) were selected for validation during training.

For the *parameters estimation* subtask:

- the dataset contain the lengths of intervals between the event time moments ($\tau_i = t_i - t_{i-1}$) and parameters of processes ($\lambda/K, \mathbf{Q}, \mathbf{\Lambda}/\alpha, \beta$),
- the datasets were splitted to train and test datasets in the ratio 80/20 (8000/2000),
- from the train datasets, 20% of the samples (1600) were selected for validation during training.

3 Quality Metrics

3.1 Classical Metrics of Classification

The following metrics are usually used for the classification problem: *accuracy* (1), *recall* (2), *precision* (3), F_1 (4), *AUC–ROC*.

The definitions of metrics are based on the following four concepts: *false positive* (FP), when there is actually no detection, but the model concludes that there is one, *false negative* (FN), when the model concludes that there is no detection, but actually there is one, *true positive* (TP), when model concludes that the detection there is, and it's actually true, and *true negative* (TN), when model concludes that there is no detection, and it's actually true.

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN}, \tag{1}$$

$$recall = \frac{TP}{TP + FN}, \tag{2}$$

$$precision = \frac{TP}{TP + FP}, \tag{3}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \tag{4}$$

AUC–ROC is a metric showing the correlation between *true positive rate* (TPR) and *false positive rate* (FPR):

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}.$$

3.2 Classical Metrics of Regression

For a regression task (in our case, parameter estimation), the following metrics are usually used:

- *mean absolute error* (MAE) is a linear estimator, hence all errors are weighted equally on average

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \bar{y_i}|,$$

where n is the number of samples, y_i is the actual value of the variable, $\bar{y_i}$ is the predicted (*estimated*) value.

- *coefficient of determination* (R^2) is the metric that shows how well a given model performs better than a "naive" model

$$R^2 = 1 - \frac{\sum\limits_{i=1}^{n}(y_i - \bar{y}_i)^2}{\sum\limits_{i=1}^{n}(y_i - \hat{y})^2},$$

  where $\hat{y}$ is the sample mean y_i. For a model with perfect predictive ability, the coefficient of determination is 1.

3.3 Other Metrics of Regression

Other metrics also were proposed. Let us illustrate the motivation for their use with the following examples. Suppose we estimate the parameters of the MMPP arrival process: the number of states, the intensities and the generator matrix. Let the true value of the number of states be 3 (the number of parameters: $3+6$), but the model was able to detect 4 states (number of parameters: $4+12$). Then the usual machine learning metrics cannot be applied.

Or let the true intensity of the events in the first state be much higher than the others, but during the estimation the model numbered this state as the third. This will not affect the quality of the approximation in any way, but it may significantly worsen the metric.

We suggest using metrics based on the following intuition.

Comparison of the Proximity of Two Curves. To evaluate the quality of the approximation, we generate a new process with the estimated parameters and compare the proximity of the two curves: the true process and the one generated by the estimated parameters (Fig. 1).

Below, are the metrics proposed for assessing the proximity of two curves.

Average maximum of modulus of the difference between the curves

$$\bar{m}_{max} = M\left\{\left(\max_{1\le i\le 2\,m}|n_{true}(t_i) - n_{estimated}(t_i)|\right)_n\right\},$$

where i is the index of the event time moment of one of the processes occurred, m is the number of event time moments in one process, n is the size of the test dataset.

Average relative error of the average length of the intervals

$$\bar{m}_{av} = M\left\{\left(\frac{M\{(\tau_{true})_m\} - M\{(\tau_{estimated})_m\}}{M\{(\tau_{true})_m\}}\right)_n\right\},$$

where τ is the length of intervals between event time moments in the process.

Median relative error of the average length of the intervals

$$\bar{m}_{median} = M_e\left\{\left(\frac{M\{(\tau_{true})_m\} - M\{(\tau_{estimated})_m\}}{M\{(\tau_{true})_m\}}\right)_n\right\}.$$

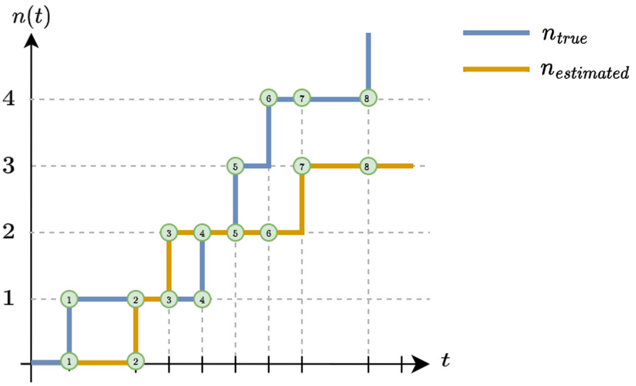


Fig. 1. Illustration of a true and estimated processes

4 Classification

In this paper, more than two type of processes are considered, so the classification task is multi-class.

In order to verify that the classification of processes does not depend on the intensity value of this process, the intensities $\lambda_{poisson}$, λ_{mmpp} (5), $\lambda_{renewal}$ (6) were calculated:

$$\lambda_{mmpp} = \mathbf{r}\mathbf{\Lambda}\mathbf{e}, \tag{5}$$

where $\mathbf{\Lambda}$ is the matrix of conditional intensities, $\mathbf{e}$ is the unit column vector, $\mathbf{Q}$ is the generator matrix, $\mathbf{r}$ is the vector of the stationary probability distribution, which is found using the matrix equations system

$$\begin{cases} \mathbf{rQ} = \mathbf{0}, \\ \mathbf{re} = 1. \end{cases}$$

$$\lambda_{renewal} = \frac{1}{\alpha * \beta} \tag{6}$$

It is found that the measures of the central tendency of the intensities for different process are close, and thus classification by intensity is excluded (Fig. 2).

Next, we consider two neural network architectures to solve the classification problem: a fully connected neural network (FCNN) [20], and an Long short-term memory neural network (LSTM) [21].

Tuning of FCNN parameters does Bayesian Optimisation Algorithm [22] from the Keras Tuner [23] library. The Table 2 lists the hyperparameters and their configurable values.

For LSTM network, the variable *lookback* is initially equal to 10. Consider the neural network architectures used for classification (Table 3).

Looking at the training graph of these architectures (Fig. 3), the value of the accuracy metric for the FCNN is much lower than that of the architecture with LSTM layer. On the validation sample, the value reaches almost 0.9.

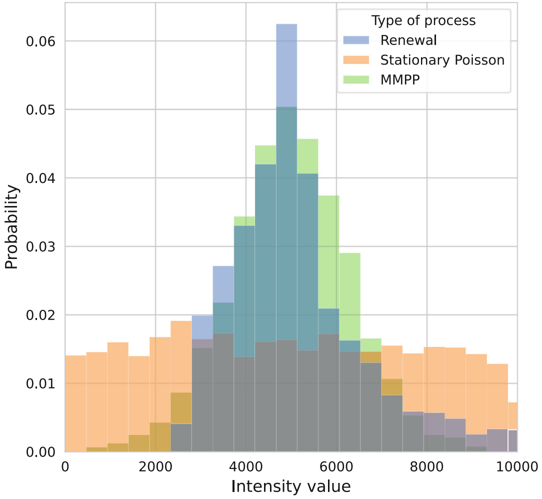


Fig. 2. Histogram of the considered processes intensities

Table 2. Hyperparameters and its configurable values

Hyperparameter	Configurable values
Number of hidden layers	$[1, 2, 3, 4, 5]$
Number of neurons in the hidden layer	$[32; 512]$
Activation function (AF)	ReLU, tanh, sigmoid, SeLU
Learning rate	$[10^{-4}, 10^{-1}]$
Dropout	$[0.1; 0.2; 0.3; 0.4]$

To achieve better predictive ability, LSTM model was trained with different number of epochs and with the value of the variable $lookback = 30$, see the Table 4. Figure 4 compares the confusion matrices for networks with 10 epochs, $lookback = 10$ and 50 epochs, $lookback = 30$.

A ROC curve (Fig. 5) was also plotted for each class of LSTM architecture with the number of epochs 10, $lookback = 10$ and the number of epochs 50, $lookback = 30$. As can be seen from the Fig. 5a, class 0 (Poisson process) is best classified by the model (AUC-ROC = 0.999989), class 1 (MMPP arrival process) (AUC-ROC = 0.999940), in the Fig. 5b, all types of flows are classified by the model with high accuracy.

The degenerate case, where $\lambda_1 = ... = \lambda_k = \lambda$ (the MMPP process is the same as the Poisson process) was considered separately. In this case, the classification model sees no difference between the types of process.

Finally, we can conclude that the neural network with LSTM layers has better predictive ability. It is also affected from the number of epochs in training and

Table 3. The used architectures of the FCNN and LSTM

Network parameters	FCNN	LSTM
Layer 1	Input layer (500)	
Layer 2	Hidden layer (114) Dropout (0.3)	LSTM (256)
Layer 3	Hidden layer (67) Dropout (0.3)	Hidden layer (32)
Layer 4	Hidden layer (124) Dropout (0.3)	–
Layer 5	Hidden layer (32) Dropout (0.3)	–
Layer 6	Output layer (3)	
Hidden layer AF	Sigmoid	ReLU
Optimizer	Adam	Adam
Learning rate	0.005	0.001
Number of epochs	30	30

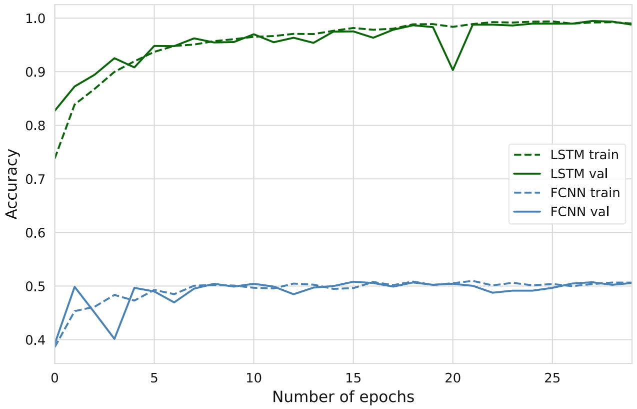
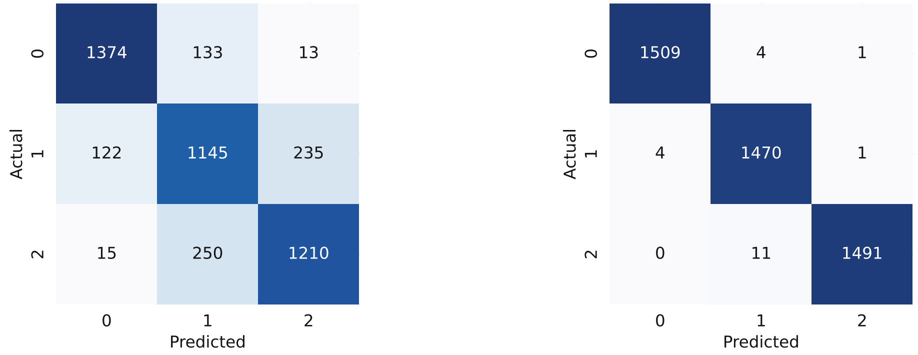


Fig. 3. Comparison of accuracy for a FCNN and LSTM

Table 4. Values of LSTM metrics with different number of epochs and lookback

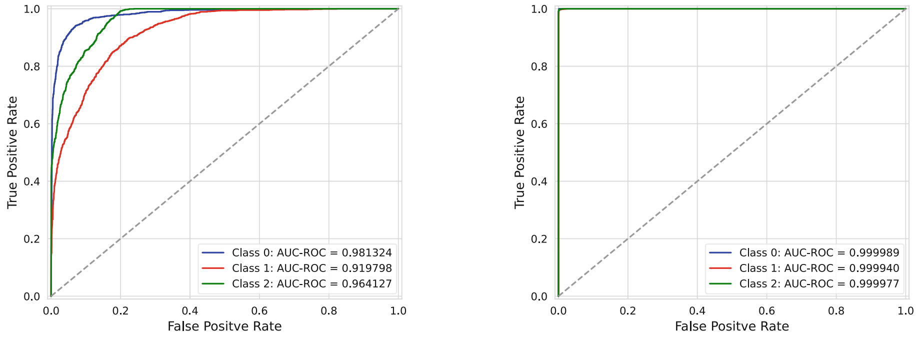
Lookback value	Number of epochs	Accuracy	Recall	Precision	F_1
10	10	0.829219	0.828868	0.829527	0.829167
10	30	0.864354	0.864013	0.865805	0.864661
30	10	0.971499	0.971354	0.971512	0.971378
30	30	0.986862	0.986841	0.986825	0.986823
30	50	0.995323	0.995328	0.995305	0.995311



a) 10 epochs, 10 lookback

b) 50 epochs, 30 lookback

Fig. 4. Confusion matrix for each class (0 – Poisson, 1 – MMPP, 2 – renewal)



a) 10 epochs, 10 lookback

b) 50 epochs, 30 lookback

Fig. 5. ROC curves for each class (0 – Poisson, 1 – MMPP, 2 – renewal)

the value of lookback variable, the more these parameters the better the metrics values and generalization ability of the model.

5 Parameters Estimation

5.1 Stationary Poisson Process

For a stationary Poisson process, the target variable is the estimated value of the intensity parameter λ, so we solve the problem of univariate regression (in terms of neural networks). For this subtask we used FCNN.

At the beginning of model building, the **event time moments** t_i were used as input data. The first models were fitted using a Bayesian optimization algorithm, where hyperparameters were selected from the Table 2, except for the number of neurons in the hidden layer:

- $[32; 1024]$ for Architecture 1
- and $[32; 2048]$ for Architecture 2.

The metric R^2 was calculated for Architecture 1 ($R^2 = -0.059957$) and Architecture 2 ($R^2 = 0.113776$). As we can see, Architecture 2 works a little better, because it is a deeper network. To improve the predictive ability we could have deepened the network, but this would have increased the number of trained parameters. Therefore, it was decided to try to use the intervals between events time moments τ_i as input data.

As a first approximation, the model parameters (Architecture 3) were selected using a Bayesian optimization algorithm. Architectures 1, 2 and 3 are presented in the table 5.

Table 5. The used architectures of the stationary Poisson process

Network parameters	Architecture 1	Architecture 2	Architecture 3
Layer 1	Input layer		
Layer 2	Hidden layer (698)	Hidden layer (1797)	Hidden layer (368)
Layer 3	Hidden layer (603)	Hidden layer (1816)	Hidden layer (64)
Layer 4	Hidden layer (32)	–	Hidden layer (742)
Layer 5	–	–	Hidden layer (106)
Layer 6	Output layer (1)		
Hidden layer AF	ReLU	ReLU	ReLU
Optimizer	Adam	Adam	Adam
Learning rate	0.001	0.001	0.001
Number of epochs	50	50	50

The metric R^2 for Architecture 3 is significantly better than previous results ($R^2 = 0.892156$). Further models were trained based on Architecture 3 with a different number of epochs, and the learning results are shown in the Table 6.

Table 6. Metrics for modifications of Architecture 3

Number of epochs	MAE	R^2
50	749.899196	0.892156
70	729.681985	0.903051
100	623.516307	0.916881

When training for 150 epochs, the maximum value of the metric R^2 was achieved, but an overfitting effect was observed (Fig. 6). So we decided to make

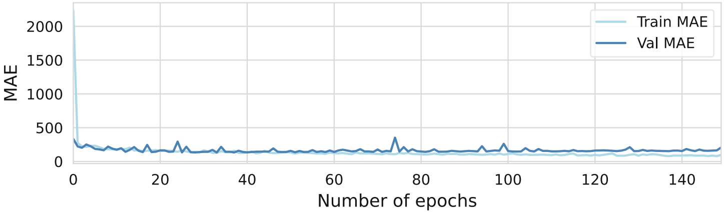


Fig. 6. Architecture 3 training plot for 150 epochs

the model simpler (Architecture 4: 2 hidden layers of 80 neurons each) and see how it behaves.

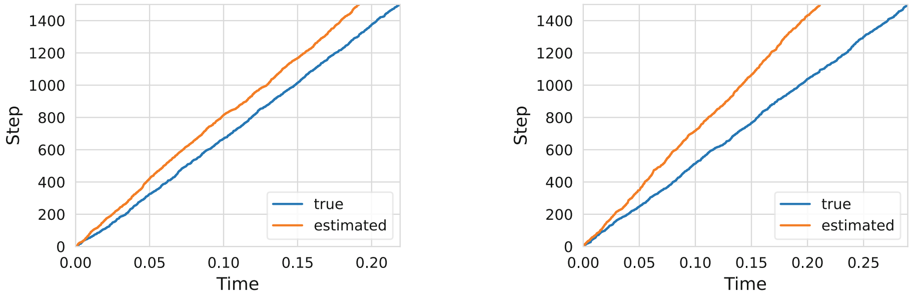
The predictive ability of the model improved by 0.06, $R^2 = 0.978362$. Then we tried to train a model with fewer epochs (18). This is the number of epochs that satisfy the condition that the error value on the i^{th} iteration is less than the error value on the $(i-1)^{\text{th}}$ iteration. This increased the predictive ability $MAE = 226.270934$, $R^2 = 0.989362$.

5.2 MMPP

For MMPP, the target variables are the qenerator matrix $\mathbf{Q}$ and the matrix of conditional intensities $\mathbf{\Lambda}$. Therefore, we solve the multivariate regression. The number of output features will be equal to $K^2 + K$, where K is the number of states. Generally speaking, the subtask of estimating the number of states K is a separate problem that should be solved in the future, but within the framework of this work we believe that K is already known. We also used FCNN for this subtask. The intervals between event time moments τ_i were chosen as input data. The architecture was selected using automatic hyperparameter selection (hidden layer AF: ReLU, optimizer: Adam, learning rate: 0.003, number of epochs: 50). The Table 7 indicates the number and types of layers.

Table 7. The used architecture of the MMPP

Network parameters	Architecture
Layer 1	Input layer (1500)
Layer 2	Hidden layer (713) Dropout (0.1)
Layer 3	Hidden layer (266) Dropout (0.1)
Layer 4	Hidden layer (854) Dropout (0.1)
Layer 5	Hidden layer (853) Dropout (0.1)
Layer 6	Output layer (110)



a) $\bar{m}_{max} = 153$, $\bar{m}_{av} = 0.998814$

b) $\bar{m}_{max} = 372$, $\bar{m}_{av} = 0.998993$

Fig. 7. True and estimated curves

The quality of the model was assessed using the metrics introduced in paragraph 2.3. To do this, curves were constructed, the Fig. 7 shows the appearance of two randomly selected examples, and the values of the metrics for each example.

The model was also trained on 100 epochs, the model became more accurate in estimating the parameters ($\bar{m}_{max} = 350.961831$), and this model was also trained with different hidden layer activation functions. It was found that with the same other parameters, the model performed best when trained with the ReLU activation function (see Table 8).

The results obtained show that the proposed metrics are still poorly measurable for such intense processes, with a small error in estimating the parameters, the metrics instantly deteriorate significantly. We will try other metrics in the future.

Table 8. Metrics values for MMPP

Number of epochs	Activation function	$\bar{m}_{max}$	$\bar{m}_{av}$	$\bar{m}_{median}$
50	ReLU	425.790304	0.999044	0.999021
100	ReLU	350.961831	0.998965	0.998968
50	SELU	441.293103	0.998608	0.998513
50	sigmod	520.495963	0.998593	0.998591
50	tanh	464.273782	0.998631	0.998634

5.3 Renewal Process

To estimate the parameters of the renewal arrival process, data different from those used in the classification task were chosen, since for the classification task the parameters were specifically chosen to obtain the desired intensity. In the regression task, models trained on such data have poor generalization ability. To solve this problem, the parameters α and β were modelled from two distributions, and the number of time points considered in one observation is equal 1000.

The first neural network architectures (based on FCNN) for further training was selected using a Bayesian optimization algorithm (hidden layer AF: ReLU, optimizer: Adam, learning rate: 0.004, number of epochs: 150). The Table 9 shows the number and types of layers. As in the MMPP parameter estimation subtask, metrics calculated from two curves were used to determine the predictive ability of the model.

Table 9. The used architectures of the renewal process

Network parameters	Architecture 1	Architecture 2	Architecture 3
Layer 1	Input layer (1000)	Input layer (1000)	Input layer (3000)
Layer 2	Hidden layer (614) Dropout (0.1)	Hidden layer (614) Dropout (0.2)	Hidden layer (567) Dropout (0.2)
Layer 3	Hidden layer (224) Dropout (0.1)	Hidden layer (224) Dropout (0.2)	Hidden layer (1006) Dropout (0.2)
Layer 4	–	–	Hidden layer (39) Dropout (0.2)
Layer 5	–	–	Hidden layer (474) Dropout (0.2)
Layer 6	Output layer (2)		

The Architecture 1 was overfitted with 150 epochs (Fig. 8), so the value of dropout layer was increased in Architecture 2, as can be seen from the Table 10, this had a positive effect on the metrics. The Architecture 3 with more input data (3000) was trained on fewer epochs as it contains many parameters, however it performed the best based on the metrics. An assumption can be made that the more data the model receives as input, the better it performs.

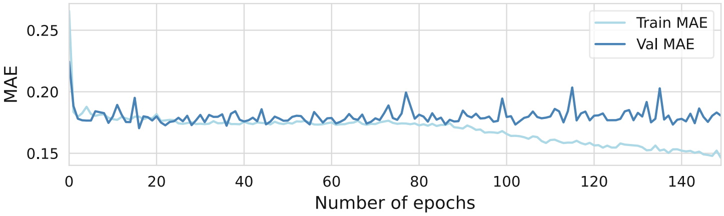


Fig. 8. Overfitting of the Architecture 1

Table 10. Metrics values for renewal process

Architecture number	Number of epochs	$\bar{m}_{max}$	$\bar{m}_{av}$	$\bar{m}_{median}$
1	150	354.070344	0.996098	0.999009
2	150	331.036943	0.996044	0.997374
3	100	263.391194	0.996474	0.997478

6 Conclusion

As a result of this work, stationary Poisson, MMPP and renewal process were simulated. The neural network technology was used to classify the process (AUC ROC = 0.99), and it was also possible to obtain excellent predictive ability in predicting the parameter λ for the stationary Poisson process ($R^2 = 0.989362$). A solution to the multivariate regression problem in predicting the parameters $\mathbf{Q}$ and $\mathbf{\Lambda}$ for the MMPP, as well as the parameters α and β of the gamma distribution of the renewal process, was obtained. To evaluate the predictive ability of the considered architectures, we used metrics widely known in machine learning, and also proposed metrics that are calculated on the basis of two curves: the true curve and the curve simulated by the estimated parameters.

In the interim results, we see that the problem is solved quite well, and can provide a good tool for engineers and mathematicians in the task of traffic identification. However, there are some difficulties with visualizing these results, and we will have to work with new metrics in the future. And also need to add other process models to this framework, other distributions for the renewal process. And our global goal is to make software that will receive traffic data in real time, find a process model, its parameters, and propose a strategy for its service in networks.

References

1. Rezaei, S., Liu, X.: Deep learning for encrypted traffic classification: an overview. IEEE Commun. Mag. **57**, 76–81 (2018)
2. Šmít, D., Millar, K., Page, C., Cheng, A., Chew, H., Lim, C.: Looking deeper: using deep learning to identify internet communications traffic (2017)
3. Lim, H., Kim, J., Heo, J., Kim, K., Hong, Y., Han, Y.: Packet-based network traffic classification using deep learning. In: International Conference on Artificial Intelligence in Information and Communication 2019, ICAIIC, Okinawa, Japan, pp. 46–51 (2019). https://doi.org/10.1109/ICAIIC.2019.8669045
4. Lopez-Martin, M., Carro, B., Sánchez-Esguevillas, A.J., Lloret, J.: Network traffic classifier with convolutional and recurrent neural networks for internet of things. IEEE Access **5**, 18042–18050 (2017)
5. Wang, W., Zhu, M., Wang, J., Zeng, X., Yang, Z.: End-to-end encrypted traffic classification with one-dimensional convolution neural networks. In: IEEE International Conference on Intelligence and Security Informatics 2017, ISI, Beijing, China, pp. 43–48 (2017). https://doi.org/10.1109/ISI.2017.8004872

6. Wang, W., Zhu, M., Zeng, X., Ye, X., Sheng, Y.: Malware traffic classification using convolutional neural network for representation learning. International Conference on Information Networking 2017, ICOIN, Da Nang, Vietnam, pp. 712–717 (2017). https://doi.org/10.1109/ICOIN.2017.7899588
7. Salman, O., Elhajj, I.H., Chehab, A., Kayssi, A.I.: A multi-level internet traffic classifier using deep learning. In: 9th International Conference on the Network of the Future 2018, NOF, Poznan, Poland, pp. 68–75 (2018). https://doi.org/10.1109/NOF.2018.8598055
8. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM **60**, 84–90 (2012)
10. Szegedy, C., et al.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition 2015, CVPR, Boston, MA, USA, pp. 1–9 (2015). https://doi.org/10.1109/CVPR.2015.7298594
11. Bayat, N., Jackson, W.J., Liu, D.: Deep Learning for Network Traffic Classification. ArXiv (2021)
12. HTTPS Websites Dataset (2016). http://betternet.lhs.loria.fr/datasets/https/
13. Wang, C., Xu, T., Qin, X.: network traffic classification with improved random forest. In: 11th International Conference on Computational Intelligence and Security 2015, CIS, Shenzhen, China, pp. 78–81 (2015). https://doi.org/10.1109/CIS.2015.27
14. Perera, P., Tian, YC., Fidge, C., Kelly, W.: A comparison of supervised machine learning algorithms for classification of communications network traffic. In: Liu, D., Xie, S., Li, Y., Zhao, D., El-Alfy, E.S. (eds.) ICONIP. vol. 10634, pp. 445–454. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70087-8_47
15. Shafiq, M., Yu, X., Wang, D.: Network traffic classification using machine learning algorithms. In: Xhafa, F., Patnaik, S., Zomaya, A. (eds.) Advances in Intelligent Systems and Interactive Applications 2017, IISA, vol. 686, pp. 621–627. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69096-4_87
16. Volkov, A.: Issledovanie i razrabotka metodov postroeniya infrastruktury i predostavlenie uslug setej svyazi na osnove tekhnologii iskusstvennogo intellekta [Research and development of methods for building infrastructure and providing services of communication networks based on artificial intelligence technologies] (2021)
17. Shikhaliyev, R.: Analysing and classifying network traffic in computer networks. Probl. Inf. Technol. **1**(2), 15–23 (2010)
18. Fadlullah, Z.M., et al.: State-of-the-art deep learning: evolving machine intelligence toward tomorrow's intelligent network traffic control systems. IEEE Commun. Surv. Tutor. **19**, 2432–2455 (2017)
19. Margolis, Yu.: Imitacionnoe modelirovanie [Simulation modelling]. Publishing House of Tomsk State University, Tomsk (2015). (in Russian)
20. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation (1986)
21. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–1780 (1997)
22. Bayesian Optimisation Algorithm. https://keras.io/api/keras_tuner/tuners/bayesian/
23. Keras Tuner. https://keras.io/api/keras_tuner/

Stochastic Analysis of a Production Inventory System with Deteriorating Items, Unreliable Server and Retrial of Customers

Bobina J. Mattam and K. P. Jose(✉)

Post Graduate and Research Department of Mathematics, St. Peter's College,
Kolenchery-682311, Kerala, India
kpjspc@gmail.com

Abstract. This paper considers a production inventory system with perishable items where the customers arrive according to a homogeneous Poisson process. The time interval between the production of each item is exponentially distributed. This is a single server system, where the service time follows exponential distribution. The server may become breakdown according to a Poisson process in which case the service restarts after an exponentially distributed time. On arrival, a customer leads to service if the server is available and the inventory is non-empty. The arriving customer on finding the server busy or breakdown goes to a waiting place(orbit) of infinite capacity with pre-determined probability or exits the system with complementary probability. Each customer in the orbit retries to access the service facility in an exponentially distributed time interval. If the retrial is unsuccessful, the customer returns to the orbit with a pre-allotted probability or is lost forever with probability equal to its complement. An algorithmic solution to the problem is obtained using Matrix Analytic Method. Average inventory level in the system, mean number of customers in the orbit, expected rate at which production is switched ON, the mean number of customer loss before and after entering the system, the rate of successful retrials among overall retrials, average breakdown rate and repair rate and some other performance measures of the system are derived. The impacts of system parameters on different measures are numerically studied. A suitable cost function is constructed to find the optimum cost.

Keywords: Perishable Inventory · Unreliable Server · Retrials of Customers · Production

The authors acknowledge the financial support provided by FIST Program, Dept. of Science and Technology, Govt. of India, to the P. G. & Research Dept. of Mathematics, through SR/FST/College-2018-XA 276(C).
The authors would like to express sincere gratitude to the referees for their insightful comments and valuable suggestions, which greatly improved the quality of this paper.

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 91–105, 2024.
https://doi.org/10.1007/978-3-031-65385-8_7

1 Introduction

A glance through the literature on perishable inventory gives us an understanding that majority of the existing perishable inventory models in the literature assume that the system does not have a production unit and the items are purchased from outside sources. This can be detrimental to industries and firms if there are no goods or items on hand. Due to this, industries/firms would lose their business. Normally, a firm must be prepared to deliver the items on demand. To make it easier for the system to deliver items without delay, we explore a deteriorating item inventory model with a production facility. Another possible difficulty that industries may face is server failure during service, which may take random time to repair. In this work, we propose a production inventory model of perishable items with an unreliable server.

An inventory system with positive service time and retrial of customers has received a small scale of attention in the literature. Artalejo et al. [1] were the first to publish a work on retrial inventory where the authors considered a continuous review (s, S) inventory system in which primary customers who arrive when the inventory level is zero, leave the server and retry after some random time. Sivakumar [9] considered a continuous review (s, Q) inventory system with perishable items with a finite number of demands. The lifetime of each item and the lead time of the system are assumed to be exponentially distributed. Reshmi and Jose [7] studied a queueing inventory system with perishable items and all underlying processes are assumed as exponential. Items in the inventory perish at a linear rate. Periyasamy [6] analysed a continuous review perishable inventory system with a single server and zero lead time. If the demand occurs during a busy period, it is directed to an obit and may retry from there. Krishnamoorthy and Islam [3] introduced perishability in a retrial inventory model with one production unit. When the inventory level reaches zero, arriving demands are sent to the orbit which has finite capacity and try for their luck. Customers, who find the orbit full and inventory level zero, may even leave the system.

Jose and Reshmi [2] studied a production inventory system with perishable items. Unsatisfied customers are provided with a retrial facility and all the underlying distributions are exponential. Ushakumari [10] analyzed a retrial inventory system with an unreliable server. The server may undergo a breakdown. Here, an interrupted customer moves to orbit and retries to access the free server. Reshmi and Jose [8] also considered a MAP/PH/1 perishable inventory model with retrials. In this retrial customers re-enter the orbit according to inventory dependent probabilities.

2 Mathematical Model

Consider a single server production inventory system with an (s, S) control policy. The items under consideration are perishable and we also assume that the system has a retrial facility. The lifetime of an item in the inventory follows an exponential distribution with parameter $j\omega$, provided there are j items in the inventory. Arrival of customers follows the Poisson process with parameter λ and each customer is served a single item. If the server is available at the arrival epoch of a customer and the inventory is non-empty then that customer is immediately allowed to enter the service. The service time duration follows an exponential distribution with parameter μ. The production process follows an exponential distribution with parameter β and it adds one unit to the inventory at a time. When the on-hand inventory level drops to s, the production is switched ON and it continues until the inventory level S is reached.

The server may experience a breakdown while in service and it follows a Poisson distribution with parameter δ_1 and its repair times are exponentially distributed with parameter δ_2. Any customer who arrives when the inventory level is zero, or server is busy or the server breakdown, is offered the choice of either joining a waiting space (orbit) of infinite capacity with probability γ or exiting the system with probability $1-\gamma$. All customers who enter the orbit, independently generate requests for service at exponentially distributed time intervals with mean $\frac{1}{\theta}$. The retrial customers who find the inventory out of stock or server busy or breakdown, return to the orbit with probability δ or exit the system with complementary probability $1-\delta$.

Let $N(t)$ and $I(t)$ denote the number of customers in the orbit at time t and the inventory level at time t respectively. Further, let

$$P(t)=\begin{cases}0, \text{if the production is OFF at time } t\\ 1, \text{if the production is ON at time } t\end{cases}$$

and

$$C(t)=\begin{cases}0, \text{if the server is idle at time } t\\ 1, \text{if the server is busy at time } t\\ 2, \text{if the server is breakdown at time } t\end{cases}$$

denote the status of the server and the production unit at the epoch t.

Now, $\mathbf{X}=\{(N(t),C(t),P(t),I(t))|t\geq 0\}$ constitute a continuous time Markov chain with state space $G_1\cup G_2\cup G_3$, where

$$\begin{aligned}
G_1&=\{(i,k,0,j)|i\geq 0; k=0,1,2; s+1\leq j\leq S\}\\
G_2&=\{(i,k,1,j)|i\geq 0; k=0,1; k\leq j\leq S-1\} \text{ and}\\
G_3&=\{(i,2,1,j)|i\geq 0; 1\leq j\leq S-1\}.
\end{aligned}$$

The generator matrix of the process is

$$Q=\begin{bmatrix} A_{10} & A_0 & & & \\ A_{21} & A_{11} & A_0 & & \\ & A_{22} & A_{12} & A_0 & \\ & & A_{23} & A_{13} & A_0 \\ & & & \ddots & \ddots & \ddots \end{bmatrix}$$

where each element in Q is a square matrix of order $(6S-3s-2)$.

Transitions of A_0

- $(i,0,1,0) \xrightarrow{\lambda\gamma} (i+1,0,1,0); i \geq 0$
- $(i,k,0,j) \xrightarrow{\lambda\gamma} (i+1,k,0,j); i \geq 0, s+1 \leq j \leq S, k=1,2$
- $(i,k,1,j) \xrightarrow{\lambda\gamma} (i+1,k,1,j); i \geq 0, 1 \leq j \leq S-1, k=1,2$

Transitions of A_{2i}

- $(i,0,0,j) \xrightarrow{i\theta} (i-1,1,0,j); i \geq 1, s+1 \leq j \leq S$
- $(i,0,1,0) \xrightarrow{i\theta(1-\delta)} (i-1,0,1,0)$
- $(i,0,1,j) \xrightarrow{i\theta} (i-1,1,1,j); i \geq 1, 1 \leq j \leq S-1$
- $(i,k,0,j) \xrightarrow{i\theta(1-\delta)} (i-1,k,0,j); i \geq 1, s+1 \leq j \leq S, k=1,2$
- $(i,k,1,j) \xrightarrow{i\theta(1-\delta)} (i-1,k,1,j); i \geq 1, 1 \leq j \leq S-1, k=1,2$

Transitions of A_{1i}

- $(i,0,0,j) \xrightarrow{\lambda} (i,1,0,j); s+1 \leq j \leq S$
- $(i,0,1,j) \xrightarrow{\lambda} (i,1,1,j); 1 \leq j \leq S-1$
- $(i,1,0,j) \xrightarrow{\mu} (i,0,1,j-1); j = s+1$
- $(i,1,0,j) \xrightarrow{\mu} (i,0,0,j-1); s+2 \leq j \leq S$
- $(i,1,1,j) \xrightarrow{\mu} (i,0,1,j-1); 1 \leq j \leq S-1$
- $(i,0,1,j) \xrightarrow{\beta} (i,0,1,j+1); 0 \leq j \leq S-2$
- $(i,0,1,j) \xrightarrow{\beta} (i,0,0,j+1); j = S-1$
- $(i,1,1,j) \xrightarrow{\beta} (i,1,1,j+1); 1 \leq j \leq S-2$
- $(i,1,1,j) \xrightarrow{\beta} (i,1,0,j+1); j = S-1$
- $(i,2,1,j) \xrightarrow{\beta} (i,2,1,j+1); 1 \leq j \leq S-2$
- $(i,2,1,j) \xrightarrow{\beta} (i,2,0,j+1); j = S-1$
- $(i,0,0,j) \xrightarrow{j\omega} (i,0,0,j-1); s+2 \leq j \leq S$
- $(i,0,0,j) \xrightarrow{(s+1)\omega} (i,0,1,j-1); j = s+1$
- $(i,0,1,j) \xrightarrow{j\omega} (i,0,1,j-1); 1 \leq j \leq S-1$
- $(i,1,0,j) \xrightarrow{(s+1)\omega} (i,1,1,j-1); j = s+1$
- $(i,1,0,j) \xrightarrow{j\omega} (i,1,0,j-1); s+2 \leq j \leq S$
- $(i,1,1,j) \xrightarrow{j\omega} (i,1,1,j-1); 2 \leq j \leq S-1$
- $(i,2,0,j) \xrightarrow{j\omega} (i,2,0,j-1); s+2 \leq j \leq S$
- $(i,2,0,j) \xrightarrow{(s+1)\omega} (i,2,1,j-1); j = s+1$
- $(i,2,1,j) \xrightarrow{j\omega} (i,2,1,j-1); 2 \leq j \leq S-1$

- $(i,1,0,j) \xrightarrow{\delta_1} (i,2,0,j); s+1 \leq j \leq S$
- $(i,1,1,j) \xrightarrow{\delta_1} (i,2,1,j); 1 \leq j \leq S-1$
- $(i,2,0,j) \xrightarrow{\delta_2} (i,1,0,j); s+1 \leq j \leq S$
- $(i,2,1,j) \xrightarrow{\delta_2} (i,1,1,j); 1 \leq j \leq S-1$
- $(i,0,0,j) \xrightarrow{\alpha_j} (i,0,0,j);$
  $\alpha_j = -\lambda - j\omega - i\theta; s+1 \leq j \leq S$
- $(i,0,1,j) \xrightarrow{\tau_j} (i,0,1,j);$
  $$\tau_j = \begin{cases} -\lambda\gamma - \beta - i\theta(1-\delta); j = 0 \\ -\lambda - \beta - j\omega - i\theta; 1 \leq j \leq S-1 \end{cases}$$
- $(i,1,0,j) \xrightarrow{\chi_j} (i,1,0,j);$
  $$\chi_j = -\lambda\gamma - \mu - j\omega - i\theta(1-\delta) - \delta_1; s+1 \leq j \leq S$$
- $(i,1,1,j) \xrightarrow{\epsilon_j} (i,1,1,j);$
  $$\epsilon_j = \begin{cases} -\lambda\gamma - \beta - \mu - \delta_1 - i\theta(1-\delta); j = 1 \\ -\lambda\gamma - \beta - \mu - \delta_1 - j\omega - i\theta(1-\delta); 2 \leq j \leq S-1 \end{cases}$$
- $(i,2,0,j) \xrightarrow{\psi_j} (i,2,0,j);$
  $$\psi_j = -\lambda\gamma - \delta_2 - i\theta(1-\delta); s+1 \leq j \leq S$$
- $(i,2,1,j) \xrightarrow{\phi_j} (i,2,1,j);$
  $$\phi_j = \begin{cases} -\lambda\gamma - \beta - \delta_2 - i\theta(1-\delta); j = 1 \\ -\lambda\gamma - \beta - \delta_2 - j\omega - i\theta(1-\delta); 2 \leq j \leq S-1 \end{cases}$$

3 System Stability

The system under consideration is stable if and only if,

$$\lim_{N\to\infty} \left(\frac{\pi_N A_0 e}{\pi_N A_{2N} e}\right) < 1$$

where,

$$\begin{aligned} \pi_N A_0 e &= \pi_{01}^N(\lambda\gamma c_S(1) \otimes r_S(1))e + ((\pi_{10}^N + \pi_{11}^N + \pi_{20}^N + \pi_{21}^N)\lambda\gamma I_{S-s})e, \\ \pi_N A_{2N} e &= (\pi_{00}^N N\theta I_{S-s})e + \pi_{01}^N(N\theta(1-\delta)c_S(1) \otimes r_S(1))e \\ &\quad + ((\pi_{10}^N + \pi_{20}^N)N\theta(1-\delta)I_{S-s})e + ((\pi_{11}^N + \pi_{21}^N)N\theta(1-\delta)I_{S-1})e. \end{aligned}$$

For obtaining this, we apply the Neuts-Rao truncation [5] by assuming $A_{1i} = A_{1N}$ and $A_{2i} = A_{2N}$ for all $i \geq N$. When the number of customers in the orbit are sufficiently large, the bulk of them fail to access the server and do not alter the state of the system. In this scenario, the change in the steady state probability vector is negligible if the number of customers in the orbit is limited to a suitably chosen N. The infinitesimal generator Q of the truncated system will be

$$Q = \begin{bmatrix} A_{10} & A_0 & & & & & & \\ A_{21} & A_{11} & A_0 & & & & & \\ & A_{22} & A_{12} & A_0 & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & A_{2N} & A_{1N} & A_0 & & \\ & & & & A_{2N} & A_{1N} & A_0 & \\ & & & & & \ddots & \ddots & \ddots \end{bmatrix}.$$

Define $A_N = A_0 + A_{1N} + A_{2N}$; then

$$A_N = \begin{bmatrix} H_{11} & H_{12} & H_{13} & 0 & 0 & 0 \\ H_{21} & H_{22} & 0 & H_{24} & 0 & 0 \\ H_{31} & H_{32} & H_{33} & H_{34} & H_{35} & 0 \\ 0 & H_{42} & H_{43} & H_{44} & 0 & H_{46} \\ 0 & 0 & H_{53} & 0 & H_{55} & H_{56} \\ 0 & 0 & 0 & H_{64} & H_{65} & H_{66} \end{bmatrix},$$

We introduce some notations to describe the terms in the above matrix.

1. I_m denote an identity matrix of order m.
2. E_m denote a matrix of order m defined as $E_m(i, j) = 1; j = i + 1$ and zero elsewhere.
3. F_m denote a matrix of order m defined as $F_m(i, j) = 1; i = j + 1$ and zero elsewhere
4. $r_m(i)$ denote a $1 \times m$ row matrix whose ith entry is 1 and all other entries are zeros.
5. $c_m(i)$ denotes the transpose of $(r_m(i))$.
6. $\otimes$ denotes Kronecker product of matrices.

Thus the entries given in the matrix A_N are as follows.

$$\begin{aligned} H_{11} &= (-\lambda - N\theta)I_{(S-s)} + \Sigma_{j=1}^{S-s} j\omega(c_{S-s}(j) \otimes r_{S-s}(j)) \\ &\quad + \Sigma_{j=2}^{S-s}(s + j)\omega(c_{S-s}(j) \otimes r_{S-s}(j - 1)), \\ H_{12} &= (s + 1)\omega(c_{S-s}(1) \otimes r_S(s + 1)), \\ H_{13} &= \lambda I_{S-s}, H_{21} = \beta c_{S-1}(S - 1) \otimes r_{S-s}(S - s), \\ H_{22} &= \beta E_S + \Sigma_{j=2}^{S-1}(j - 1)\omega(c_S(j) \otimes r_S(j - 1)) - \beta I_S \\ &\quad - (\lambda + N\theta)\Sigma_{j=2}^{S}(j - 1)\omega(c_S(j) \otimes r_S(j)), \end{aligned}$$

$$\begin{aligned}
H_{24} &= (\lambda + N\theta)\Sigma_{j=2}^{S}(c_S(j) \otimes r_{S-1}(j-1)),\\
H_{31} &= \mu F_S, H_{32} = \mu(c_{S-s}(1) \otimes r_S(s+1)),\\
H_{33} &= \Sigma_{j=2}^{S-s}(s+j)\omega(c_{S-s}(j) \otimes r_{S-s}(j-1)) - (\mu+\delta_1)I_{S-s}\\
&\quad - \Sigma_{j=1}^{S-s}(s+j)\omega(c_{S-s}(j) \otimes r_{S-s}(j)),\\
H_{34} &= (s+1)\omega(c_{S-s}(1) \otimes r_{S-1}(s)),\\
H_{35} &= \delta_1 I_{S-s}, H_{42} = \Sigma_{j=1}^{S-1}\mu(c_{S-1}(j) \otimes r_S(j)),\\
H_{43} &= \beta(c_{S-1}(S-1) \otimes r_{S-s}(S-s)),\\
H_{44} &= \Sigma_{j=2}^{S-1} j\omega(c_{S-1}(j) \otimes r_{S-1}(j-1)) - (\beta+\mu+\delta_1)I_{S-s}\\
&\quad - \Sigma_{j=2}^{s-1} j\omega(c_{S-1}(j) \otimes r_{S-1}(j)) + \beta E_{S-1},\\
H_{46} &= \delta_1 I_{S-1}, H_{53} = \delta_2 I_{S-s},\\
H_{55} &= \Sigma_{j=2}^{S-s}(s+j)\omega(c_{S-s}(j) \otimes r_{S-s}(j-1)) - \delta_2 I_{S-s},\\
H_{56} &= (s+1)\omega(c_{S-s}(1) \otimes r_{S-1}(s)), H_{64} = \delta_2 I_{S-1},\\
H_{65} &= \beta c_{S-1}(S-1) \otimes r_{S-s}(S-s),\\
H_{66} &= \beta E_{S-1} - (\beta+\delta_2)I_{S-1} + \Sigma_{j=2}^{S-1} j\omega(c_{S-1}(j) \otimes r_{S-1}(j-1))\\
&\quad - \Sigma_{j=2}^{s-1} j\omega(c_{S-1}(j) \otimes r_{S-1}(j)).
\end{aligned}$$

Let, $\pi_N = (\pi_{00}^N, \pi_{01}^N, \pi_{10}^N, \pi_{11}^N, \pi_{20}^N, \pi_{21}^N)$, where

$$\begin{aligned}
\pi_{00}^N &= (\pi_{N,0,0,s+1}, \pi_{N,0,0,s+2} \cdots, \pi_{N,0,0,S}), \pi_{01}^N = (\pi_{N,0,1,0}, \pi_{N,0,1,1}, \cdots, \pi_{N,0,1,S-1}),\\
\pi_{10}^N &= (\pi_{N,1,0,s+1}, \pi_{N,1,0,s+2}, \cdots, \pi_{N,1,0,S}), \pi_{11}^N = (\pi_{N,1,1,1}, \pi_{N,1,1,2}, \cdots, \pi_{N,1,1,S-1}),\\
\pi_{20}^N &= (\pi_{N,2,0,s+1}, \pi_{N,2,0,s+2}, \cdots, \pi_{N,2,0,S}), \pi_{21}^N = (\pi_{N,2,1,1}, \pi_{N,2,1,2}, \cdots, \pi_{N,2,1,S-1})
\end{aligned}$$

be the steady state vector of the generator matrix A_N.

Then the relation $\pi_N A_N = 0$ and the normalizing condition $\pi_N e = 1$ gives rise to the following equations:

$$\begin{aligned}
&\pi_{00}^N H_{11} + \pi_{01}^N H_{21} + \pi_{10}^N H_{31} = 0, && \Longrightarrow \pi_{00}^N = -(\pi_{01}^N H_{21} + \pi_{10}^N H_{31})H_{11}^{-1},\\
&\pi_{00}^N H_{12} + \pi_{01}^N H_{22} + \pi_{10}^N H_{32} + \pi_{11}^N H_{42} = 0, && \Longrightarrow \pi_{01}^N = -(\pi_{00}^N H_{12} + \pi_{10}^N H_{32} + \pi_{11}^N H_{42})H_{22}^{-1},\\
&\pi_{00}^N H_{13} + \pi_{10}^N H_{33} + \pi_{11}^N H_{43} + \pi_{20}^N H_{53} = 0, && \Longrightarrow \pi_{10}^N = -(\pi_{00}^N H_{13} + \pi_{11}^N H_{43} + \pi_{20}^N H_{53})H_{33}^{-1},\\
&\pi_{01}^N H_{24} + \pi_{10}^N H_{34} + \pi_{11}^N H_{44} + \pi_{21}^N H_{64} = 0, && \Longrightarrow \pi_{11}^N = -(\pi_{01}^N H_{24} + \pi_{10}^N H_{34} + \pi_{21}^N H_{64})H_{44}^{-1},\\
&\pi_{10}^N H_{35} + \pi_{20}^N H_{55} + \pi_{21}^N H_{65} = 0, && \Longrightarrow \pi_{20}^N = -(\pi_{10}^N H_{35} + \pi_{21}^N H_{65})H_{55}^{-1},\\
&\pi_{11}^N H_{46} + \pi_{20}^N H_{56} + \pi_{21}^N H_{66} = 0 && \Longrightarrow \pi_{21}^N = -(\pi_{11}^N H_{46} + \pi_{20}^N H_{56})H_{66}^{-1},
\end{aligned}$$

The invertibility of the matrices $H_{ii}; i = 1, 2, .., 6$ follows from the fact that they are strictly diagonally dominant. Thus by Block Gauss-Seidel iteration, we can find the vector π_N.

Now, we know that the truncated system is stable if and only if $\pi_N A_0 e < \pi_N A_{2N} e$. By rearranging and using the limiting technique used by Krishnamoorthy et.al. [4], as $N \to \infty$, we get $\lim_{N\to\infty}(\frac{\pi_N A_0 e}{\pi_N A_{2N} e}) < 1$, where

$$\begin{aligned}\pi_N A_0 e &= \pi_{01}^N(\lambda\gamma c_S(1) \otimes r_S(1))e + ((\pi_{10}^N + \pi_{11}^N + \pi_{20}^N + \pi_{21}^N)\lambda\gamma I_{S-s})e,\\ \pi_N A_{2N} e &= (\pi_{00}^N N\theta I_{S-s})e + \pi_{01}^N(N\theta(1-\delta)c_S(1) \otimes r_S(1))e\\ &\quad + ((\pi_{10}^N + \pi_{20}^N)N\theta(1-\delta)I_{S-s})e + ((\pi_{11}^N + \pi_{21}^N)N\theta(1-\delta)I_{S-1})e,\end{aligned}$$

as the stability condition which was verified numerically.

4 Steady State Distribution

Since $\mathbf{X}$ is a level dependent quasi-birth-death process, we use the method described by Neuts-Rao [5] to calculate the steady state probability vector. Consider the steady state probability vector $\mathbf{x} = (x_0, x_1, x_2, \dots)$ of Q, where,

$$\begin{aligned}x_i = (&z_{i,0,0,s+1}, z_{i,0,0,s+2} \cdots, z_{i,0,0,S}, z_{i,0,1,0}, z_{i,0,1,1}, \dots, z_{i,0,1,S-1},\\ &z_{i,1,0,s+1}, z_{i,1,0,s+2}, \dots, z_{i,1,0,S}, z_{i,1,1,1}, z_{i,1,1,2}, \dots, z_{i,1,1,S-1},\\ &z_{i,2,0,s+1}, z_{i,2,0,s+2}, \dots, z_{i,2,0,S}, z_{i,2,1,1}, z_{i,2,1,2}, \dots, z_{i,2,1,S-1})(i \geq 0).\end{aligned}$$

Here, x_i satisfies the relation

$$x_{N+k-1} = x_{N-1}R^k, \ k \geq 1,$$

where the matrix R is the unique non-negative solution of the matrix quadratic equation,

$$R^2 A_2 + RA_1 + A_0 = \mathbf{0},$$

with $A_1 = A_{1N}, A_2 = A_{2N}$ and $R = \lim_{n\to\infty} R_n$, where $\{R_n\}$ is defined such that $R_{n+1} = -A_0 A_1^{-1} - R_n A_2 A_1^{-1}; n \geq 0$ and $R_0 = \mathbf{0}$. The vectors $x_0, x_1, \dots, x_{N-1}$ are obtained by solving the equations

$$\begin{aligned}&x_0 A_{10} + x_1 A_{21} = 0\\ &x_{i-1}A_0 + x_i A_{1i} + x_{i+1}A_{2(i+1)} = 0; (1 \leq i \leq N-2)\\ &x_{N-2}A_0 + x_{N-1}(A_{1(N-1)} + RA_2) = 0\end{aligned}$$

subject to normalizing condition

$$[\Sigma_{i=0}^{N-2} x_i + x_{N-1}(1-R)^{-1}]e = 1.$$

5 Performance Measures

Using the above probability vectors, we calculated some important performance measures which are given below,

1. Expected inventory level in the system,

$$E_{inv} = \Sigma_{i=0}^{\infty}\Sigma_{k=0}^{2}\Sigma_{j=s+1}^{S} j z_{i,k,0,j} + \Sigma_{i=0}^{\infty}\Sigma_{k=0}^{2}\Sigma_{j=1}^{S-1} j z_{i,k,1,j}.$$

2. Mean number of customers in the orbit,

$$E_{orbit} = \Sigma_i \Sigma_k \Sigma_l \Sigma_j i z_{i,k,l,j} = \left(\Sigma_{i=1}^{\infty} i x_i\right) \mathbf{e}.$$

3. Expected rate at which production is switched ON,

$$E_{ON} = \mu \Sigma_{i=0}^{\infty} z_{i,1,0,s+1} \\ + (s+1)\omega \left(\Sigma_{i=0}^{\infty} z_{i,0,0,s+1} + \Sigma_{i=0}^{\infty} z_{i,1,0,s+1} + \Sigma_{i=0}^{\infty} z_{i,2,0,s+1}\right).$$

4. Expected perishability rate,

$$E_p = \omega \left(z_{i,0,1,1} + \Sigma_{i=0}^{\infty} \Sigma_{k=0}^{2} \Sigma_{j=s+2}^{S} j z_{i,k,0,j} + \Sigma_{i=0}^{\infty} \Sigma_{k=0}^{2} \Sigma_{j=2}^{S-1} j z_{i,k,1,j}\right).$$

5. Average number of departures after service completion,

$$E_{ds} = \mu \Sigma_{i=0}^{\infty} \left(\Sigma_{j=s+2}^{S} z_{i,1,0,j} + \Sigma_{j=1}^{S-1} z_{i,1,1,j}\right).$$

6. Average number of customers lost prior to entering the orbit,

$$E_{la} = \lambda(1-\gamma) \Sigma_{i=0}^{\infty} \left(z_{i,0,1,0} + \Sigma_{j=s+1}^{S} z_{i,1,0,j} + \Sigma_{j=1}^{S-1} z_{i,1,1,j}\right) \\ + \lambda(1-\gamma) \Sigma_{i=0}^{\infty} \left(\Sigma_{j=s+1}^{S} z_{i,2,0,j} + \Sigma_{j=1}^{S-1} z_{i,2,1,j}\right).$$

7. Average number of customers lost during retrials,

$$E_{lr} = \theta(1-\delta) \Sigma_{i=0}^{\infty} i \left(z_{i,0,1,0} + \Sigma_{j=s+1}^{S} z_{i,1,0,j} + \Sigma_{j=1}^{S-1} z_{i,1,1,j}\right) \\ + \theta(1-\delta) \Sigma_{i=0}^{\infty} i \left(\Sigma_{j=s+1}^{S} z_{i,2,0,j} + \Sigma_{j=1}^{S-1} z_{i,2,1,j}\right).$$

8. Average rate of breakdown

$$A_{br} = \delta_1 \Sigma_{i=0}^{\infty} \left(\Sigma_{j=s+1}^{S} z_{i,1,0,j} + \Sigma_{j=1}^{S-1} z_{i,1,1,j}\right).$$

9. Average rate of repair

$$A_{rr} = \delta_2 \Sigma_{i=0}^{\infty} \left(\Sigma_{j=s+1}^{S} z_{i,2,0,j} + \Sigma_{j=1}^{S-1} z_{i,2,1,j}\right).$$

10. Overall rate of retrials,

$$\theta_1^* = \theta \left(\Sigma_{i=1}^{\infty} i x_i\right) \mathbf{e}.$$

11. Successful rate of retrials,

$$\theta_2^* = \theta \Sigma_{i=0}^{\infty} i \left(\Sigma_{j=s+1}^{S} z_{i,0,0,j} + \Sigma_{j=1}^{S-1} z_{i,0,1,j}\right).$$

12. Fraction of time production is ON

$$F_{ON} = \Sigma_{i=0}^{\infty} \Sigma_{j=0}^{S-1} z_{i,0,1,j} + \Sigma_{i=0}^{\infty} \Sigma_{j=1}^{S-1} z_{i,1,1,j}.$$

13. Ratio of successful rate of retrials,

$$R_{sr} = \frac{\theta_2^*}{\theta_1^*}.$$

6 Cost Analysis

To determine the optimum cost for the model under consideration, we construct the expected total cost (ETC), per unit time in terms of performance measures as follows:

$$ETC = k_1 E_{ON} + k_2 Einv + k_3 Eorbit + k_4(E_{la} + E_{lr}) + k_5 E_p + k_6 F_{ON} + k_7 A_{br} + k_8 A_{rr},$$

where k_1 = Production switch on cost per unit time, k_2 = Holding cost of inventory per unit per unit time, k_3 = Holding cost of customers per unit per unit time, k_4 = Cost due to loss of customers per unit per unit time, k_5 = Cost due to the decay of items per unit per unit time, k_6 = Cost of running production process per unit time, k_7 = Penalty due to breakdown of server per unit per unit time and k_8 = Cost of server repair per unit per unit time.

7 Numerical Experiments

In this section, we provide results of numerical illustration that has been carried out for studying the effects of variation of different parameters on various performance measures. Numerical experiments are conducted by considering some artificial data. Assume that the production switch on level, $s = 6$ and the maximum permissible inventory level, $S = 18$. To study the variation of each parameter on system performances, we consider the following cases 7.1 to 7.9 with table representations.

7.1 Effect of the Arrival Rate λ

As the arrival rate λ increases, the number of customers in the orbit E_{orbit} also increases which in turn leads to the loss of arriving customers as well as retrying customers. The increase in E_{orbit} results in the increase of E_{ds}, θ_1^* and θ_2^* (see Table 1). The decrease in expected inventory level can be seen due to a decrease in expected production switching rate.

Table 1. Effect of arrival rate λ on various performance measures

λ	E_{inv}	E_{orbit}	E_{ON}	E_p	E_{ds}	E_{la}	E_{lr}	A_{br}	A_{rr}	θ_1^*	θ_2^*
2.2	1.4045	1.6734	2.9503e−11	1.6854	1.1589	0.49056	0.55055	0.15452	0.022975	3.3468	0.59409
2.3	1.392	1.8047	2.8287e−11	1.6704	1.1804	0.52135	0.59821	0.15739	0.020866	3.6093	0.61826
2.4	1.3804	1.9388	2.7248e−11	1.6565	1.2005	0.55227	0.64725	0.16006	0.018881	3.8776	0.64138
2.5	1.3698	2.0757	2.6356e−11	1.6437	1.2191	0.5833	0.69759	0.16255	0.017027	4.1514	0.66344
2.6	1.3599	2.2151	2.5591e−11	1.6319	1.2364	0.61441	0.74914	0.16486	0.015307	4.4302	0.68449
2.7	1.3508	2.3568	2.493e−11	1.621	1.2526	0.64558	0.80182	0.16701	0.013719	4.7137	0.70454
2.8	1.3424	2.5007	2.4359e−11	1.6109	1.2676	0.6768	0.85557	0.16902	0.012261	5.0015	0.72363
2.9	1.3346	2.6467	2.3862e−11	1.6016	1.2817	0.70804	0.9103	0.17089	0.01093	5.2933	0.74178

$S = 18; s = 6; \mu = 3; \omega = 1.2; \beta = 2.6; \theta = 2; \gamma = 0.7; \delta = 0.8; \delta_1 = 0.4; \delta_2 = 1.6$

7.2 Effect of the Service Rate μ

Intuitively, an increase in service rate leads to a greater number of service completions. Therefore, E_{ds} also increases and the number of customers in the orbit E_{orbit} decreases. The overall and successful rate of retrials decreases because E_{orbit} is decreasing. Expected inventory level E_{inv} gets decreased when more and more customers get served, leading to a decrease in E_p. So the production process need not have to switch ON frequently. The breakdown rate is decreasing and the repair rate is increasing. The number of unsatisfied customers decreases, that is E_{la} and E_{lr} in Table 2 support the intuition.

Table 2. Effect of service rate μ on various performance measures

μ	E_{inv}	E_{orbit}	E_{ON}	E_p	E_{ds}	E_{la}	E_{lr}	A_{br}	A_{rr}	θ_1^*	θ_2^*
2.6	1.4414	2.2047	3.302e−11	1.7297	1.1483	0.59858	0.75316	0.17666	0.016614	4.4093	0.64352
2.7	1.4223	2.1698	3.1106e−11	1.7067	1.1674	0.59456	0.73807	0.17294	0.016744	4.3396	0.64924
2.8	1.404	2.1368	2.937e−11	1.6848	1.1855	0.59068	0.72382	0.16936	0.016855	4.2735	0.65443
2.9	1.3865	2.1054	2.7793e−11	1.6638	1.2027	0.58693	0.71034	0.16589	0.016949	4.2109	0.65915
3	1.3698	2.0757	2.6356e−11	1.6437	1.2191	0.5833	0.69759	0.16255	0.017027	4.1514	0.66344
3.1	1.3538	2.0475	2.5046e−11	1.6245	1.2347	0.5798	0.68552	0.15931	0.01709	4.0949	0.66734
3.2	1.3384	2.0206	2.3848e−11	1.6061	1.2495	0.57641	0.67408	0.15619	0.01714	4.0413	0.67088
3.3	1.3238	1.9951	2.275e−11	1.5885	1.2636	0.57314	0.66323	0.15317	0.017177	3.9902	0.67409
3.4	1.3097	1.9708	2.1742e−11	1.5716	1.2771	0.56998	0.65294	0.15025	0.017202	3.9417	0.67701

$S = 18; s = 6; \lambda = 2.5; \omega = 1.2; \beta = 2.6; \theta = 2; \gamma = 0.7; \delta = 0.8; \delta_1 = 0.4; \delta_2 = 1.6$

7.3 Effect of the Perishable Rate ω

When decay rate increases, obviously E_p increases, which leads to decrease in expected inventory level E_{inv} as well as in expected departure from service E_{ds}. The production switch on rate is decreasing but it is very negligible. As E_{inv} decreases, more customers joins the orbit i.e. E_{orbit} increases. When E_{orbit} increases, we expect increase in measures like E_{la}, E_{lr} and θ_1^*. The breakdown rate is decreasing along with the repair rate. Table 3 supports these intuitions.

Table 3. Effect of perishable rate ω on various performance measures

ω	E_{inv}	E_{orbit}	E_{ON}	E_p	E_{ds}	E_{la}	E_{lr}	A_{br}	A_{rr}	θ_1^*	θ_2^*
0.8	1.7962	1.9275	4.1765e−09	1.437	1.3022	0.56171	0.63607	0.17363	0.020244	3.8549	0.67457
0.9	1.6558	1.9696	1.0079e−09	1.4902	1.2785	0.56798	0.65347	0.17047	0.019302	3.9392	0.67182
1	1.5424	2.008	2.7263e−10	1.5424	1.257	0.57361	0.66939	0.1676	0.018463	4.016	0.66904
1.1	1.4487	2.0432	8.1281e−11	1.5935	1.2373	0.57869	0.68404	0.16497	0.017709	4.0864	0.66624
1.2	1.3698	2.0757	2.6356e−11	1.6437	1.2191	0.5833	0.69759	0.16255	0.017027	4.1514	0.66344
1.3	1.3024	2.1058	9.195e−12	1.6931	1.2023	0.58752	0.71019	0.16031	0.016407	4.2116	0.66067
1.4	1.244	2.1339	3.4221e−12	1.7416	1.1866	0.59139	0.72197	0.15822	0.015841	4.2678	0.65793
1.5	1.193	2.1602	1.3486e−12	1.7894	1.172	0.59496	0.73302	0.15627	0.01532	4.3203	0.65523
1.6	1.1478	2.1848	5.6019e−13	1.8366	1.1583	0.59827	0.74342	0.15444	0.01484	4.3696	0.65255

$S = 18; s = 6; \lambda = 2.5\mu = 3; \beta = 2.6; \theta = 2; \gamma = 0.7; \delta = 0.8; \delta_1 = 0.4; \delta_2 = 1.6$

7.4 Effect of the Replenishment Rate β

As the replenishment rate β increases, the expected inventory E_{inv} increases and hence the expected perishable rate E_p increases. The production switch on rate also increases with an increase in β. When the inventory available to customers increases the service completion becomes faster, so E_{ds}. The breakdown rate is increasing along with the repair rate as the replenishment rate increases. Accordingly, the expected number of customers in the orbit E_{orbit} decreases, due to this, the measures E_{la}, E_{lr} and θ_1^* decreases (see Table 4).

Table 4. Effect of replenishment rate β on various performance measures

β	E_{inv}	E_{orbit}	E_{ON}	E_p	E_{ds}	E_{la}	E_{lr}	A_{br}	A_{rr}	θ_1^*	θ_2^*
2.3	1.1888	2.1977	3.4125e−12	1.4266	1.1518	0.59831	0.74984	0.15358	0.014924	4.3954	0.64623
2.4	1.2484	2.1546	6.9567e−12	1.4981	1.1756	0.59314	0.7313	0.15674	0.015641	4.3092	0.65269
2.5	1.3087	2.114	1.374e−11	1.5705	1.198	0.58813	0.71391	0.15973	0.016343	4.2279	0.65841
2.6	1.3698	2.0757	2.6356e−11	1.6437	1.2191	0.5833	0.69759	0.16255	0.017027	4.1514	0.66344
2.7	1.4315	2.0396	4.9211e−11	1.7178	1.2391	0.57864	0.68229	0.16521	0.017694	4.0793	0.66787
2.8	1.494	2.0057	8.9604e−11	1.7928	1.2579	0.57415	0.66793	0.16772	0.018341	4.0114	0.67175
2.9	1.5571	1.9737	1.5938e−10	1.8685	1.2757	0.56983	0.65446	0.1701	0.01897	3.9474	0.67514
3	1.6209	1.9436	2.7736e−10	1.9451	1.2925	0.56567	0.64183	0.17233	0.019578	3.8872	0.67808
3.1	1.6854	1.9152	4.7292e−10	2.0225	1.3083	0.56168	0.62997	0.17445	0.020167	3.8305	0.68062

$S = 18; s = 6; \lambda = 2.5\mu = 3; \omega = 1.2; \theta = 2; \gamma = 0.7; \delta = 0.8; \delta_1 = 0.4; \delta_2 = 1.6$

7.5 Effect of the Retrial Rate θ

As the retrial rate θ increases, one would expect a decrease in expected number of customers in the orbit E_{orbit}. Which is the reason for decrease in E_{ds} and θ_2^*. As the production switch on rate increases, expected inventory level E_{inv} and E_p increases. The breakdown rate increases and the repair rate decreases along with theta. The decrease in E_{la} is very negligible because E_{inv} is increasing. From Table 5, as θ increases most of the retrying customers fail to access a free server so E_{lr} increases.

Table 5. Effect of retrial rate θ on various performance measures

θ	E_{inv}	E_{orbit}	E_{ON}	E_p	E_{ds}	E_{la}	E_{lr}	A_{br}	A_{rr}	θ_1^*	θ_2^*
1.6	1.3647	2.5704	2.5777e−11	1.6376	1.2273	0.58675	0.6859	0.16365	0.01142	4.1127	0.68318
1.7	1.366	2.4251	2.5925e−11	1.6392	1.2252	0.58587	0.68889	0.16337	0.012832	4.1226	0.67814
1.8	1.3673	2.2957	2.6071e−11	1.6407	1.2232	0.585	0.69183	0.16309	0.014241	4.1323	0.67317
1.9	1.3685	2.18	2.6215e−11	1.6422	1.2211	0.58414	0.69473	0.16282	0.015641	4.1419	0.66827
2	1.3698	2.0757	2.6356e−11	1.6437	1.2191	0.5833	0.69759	0.16255	0.017027	4.1514	0.66344
2.1	1.371	1.9813	2.6495e−11	1.6452	1.2171	0.58247	0.7004	0.16228	0.018396	4.1607	0.65869
2.2	1.3722	1.8954	2.6632e−11	1.6467	1.2152	0.58165	0.70317	0.16202	0.019744	4.1699	0.65401
2.3	1.3734	1.8169	2.6765e−11	1.6481	1.2133	0.58085	0.7059	0.16177	0.02107	4.1789	0.6494
2.4	1.3746	1.7449	2.6896e−11	1.6495	1.2114	0.58005	0.70859	0.16151	0.022372	4.1878	0.64486

$S = 18; s = 6; \lambda = 2.5\mu = 3; \omega = 1.2; \beta = 2.6; \gamma = 0.7; \delta = 0.8; \delta_1 = 0.4; \delta_2 = 1.6$

7.6 Effect of the Probability γ

When the probability γ increases, unsatisfied customers move to orbit, hence E_{orbit} increases. This in turn leads to the reduced loss of customers upon arrival, E_{la} decreases. As E_{orbit} increases retrials become unsuccessful that force to increase in E_{lr}. As E_{orbit} increases, we expect an increase in E_{ds}, θ_1^* and θ_2^*. Here breakdown rate increases while the repair rate decreases. Table 6 supports these intuitions. As expected production switch on rate decreases, inventory level also decreases which leads to a decrease in E_p.

Table 6. Effect of probability γ on various performance measures

γ	E_{inv}	E_{orbit}	E_{ON}	E_p	E_{ds}	E_{la}	E_{lr}	A_{br}	A_{rr}	θ_1^*	θ_2^*
0.1	1.5035	0.20387	3.805e−11	1.8042	0.98716	1.4512	0.061623	0.13162	0.098574	0.40774	0.099623
0.2	1.4792	0.43658	3.5559e−11	1.7751	1.0284	1.3369	0.13474	0.13712	0.076136	0.87316	0.19948
0.3	1.4554	0.7	3.3275e−11	1.7465	1.0692	1.2105	0.22031	0.14256	0.058056	1.4	0.29846
0.4	1.4323	0.99543	3.1211e−11	1.7187	1.1091	1.0718	0.31909	0.14789	0.043668	1.9909	0.39542
0.5	1.4102	1.3234	2.9371e−11	1.6922	1.1477	0.92077	0.43153	0.15303	0.032379	2.6469	0.48924
0.6	1.3893	1.6839	2.7754e−11	1.6671	1.1845	0.75776	0.55777	0.15793	0.023657	3.3677	0.57887
0.7	1.3698	2.0757	2.6356e−11	1.6437	1.2191	0.5833	0.69759	0.16255	0.017027	4.1514	0.66344
0.8	1.3518	2.4973	2.5161e−11	1.6222	1.2513	0.39819	0.85047	0.16685	0.012073	4.9946	0.7423
0.9	1.3355	2.9465	2.4155e−11	1.6026	1.281	0.2034	1.0156	0.1708	0.0084355	5.893	0.815

$S = 18; s = 6; \lambda = 2.5, \mu = 3; \lambda = 2; \omega = 1.2; \beta = 2.6; \theta = 2; \delta = 0.8; \delta_1 = 0.4; \delta_2 = 1.6$

7.7 Effect of the Probability δ

As δ increases, the unsuccessful retrying customers return to the orbit faster, so E_{orbit} increases. This leads to a decrease in the expected loss of retrying customers. Since the number of orbiting customers increases it makes the server busy so the expected loss upon arrival E_{la} increases. The increase in E_{orbit} leads to an increase in E_{ds}, θ_1^* and θ_2^*. From Table 7, the production switch on rate increases with an increase in δ which results the increase in E_{inv}.

Table 7. Effect of probability δ on various performance measures

δ	E_{inv}	E_{orbit}	E_{ON}	E_p	E_{ds}	E_{la}	E_{lr}	A_{br}	A_{rr}	θ_1^*	θ_2^*
0.1	1.4575	0.65421	3.3376e−11	1.749	1.0653	0.51705	0.91763	0.14204	0.062076	1.3084	0.28882
0.2	1.4518	0.7211	3.2843e−11	1.7421	1.0752	0.52128	0.90352	0.14336	0.05773	1.4422	0.3128
0.3	1.4449	0.80406	3.2226e−11	1.7339	1.087	0.52633	0.8867	0.14493	0.052832	1.6081	0.3414
0.4	1.4367	0.90996	3.1497e−11	1.724	1.1013	0.53247	0.86625	0.14684	0.047293	1.8199	0.37617
0.5	1.4264	1.0505	3.062e−11	1.7117	1.1191	0.54013	0.84075	0.14922	0.041017	2.101	0.41954
0.6	1.4132	1.2476	2.9542e−11	1.6959	1.1421	0.55003	0.80784	0.15228	0.033908	2.4952	0.47556
0.7	1.3955	1.5478	2.8172e−11	1.6746	1.1733	0.56348	0.7632	0.15644	0.025901	3.0956	0.55159
0.8	1.3698	2.0757	2.6356e−11	1.6437	1.2191	0.5833	0.69759	0.16255	0.017027	4.1514	0.66344
0.9	1.3265	3.351	2.3749e−11	1.5918	1.2978	0.61761	0.58455	0.17305	0.0076108	6.7021	0.85654

$S = 18; s = 6; \lambda = 2.5, \mu = 3; \lambda = 2; \omega = 1.2; \beta = 2.6; \theta = 2; \gamma = 0.7; \delta_1 = 0.4; \delta_2 = 1.6$

7.8 Effect of the Breakdown Rate δ_1

As the breakdown rate δ_1 increases, one would expect an increase in expected number of customers in the orbit E_{orbit}. It can be seen that even though θ_1^* increases, θ_2^* is decreasing. As the production switch on rate increases, the expected inventory level E_{inv} and E_p increases. When θ increases, also, intuitively, it is clear that E_{ds} decreases. The increase in E_{inv} causes an increase in E_{la}. From Table 8, as δ_1 increases most of the retrying customers fail to access a free server so E_{lr} increases.

Table 8. Effect of retrial rate δ_1 on various performance measures

δ_1	E_{inv}	E_{orbit}	E_{ON}	E_p	E_{ds}	E_{la}	E_{lr}	A_{br}	A_{rr}	θ_1^*	θ_2^*
0.2	1.3174	1.9764	1.924e−11	1.5808	1.2735	0.57208	0.65446	0.084897	0.0094144	3.9527	0.6804
0.3	1.3441	2.0269	2.2801e−11	1.6129	1.2458	0.57784	0.67637	0.12458	0.013422	4.0538	0.67192
0.4	1.3698	2.0757	2.6356e−11	1.6437	1.2191	0.5833	0.69759	0.16255	0.017027	4.1514	0.66344
0.5	1.3945	2.1228	2.9901e−11	1.6733	1.1934	0.58848	0.71813	0.1989	0.020269	4.2457	0.65499
0.6	1.4182	2.1684	3.343e−11	1.7018	1.1686	0.5934	0.73802	0.23371	0.023185	4.3367	0.64658
0.7	1.441	2.2123	3.6941e−11	1.7292	1.1446	0.59808	0.75729	0.26708	0.025808	4.4247	0.63823
0.8	1.463	2.2548	4.0429e−11	1.7556	1.1215	0.60253	0.77594	0.29907	0.028166	4.5097	0.62995
0.9	1.4841	2.2959	4.3891e−11	1.781	1.0992	0.60676	0.79402	0.32977	0.030285	4.5918	0.62175
1	1.5045	2.3356	4.7324e−11	1.8054	1.0777	0.61079	0.81153	0.35923	0.032189	4.6713	0.61365

$S = 18; s = 6; \lambda = 2.5; \mu = 2; \omega = 1.2; \beta = 2.6; \theta = 2; \gamma = 0.7; \delta = 0.8; \delta_2 = 1.6$

7.9 Effect of the Breakdown Rate δ_2

As repair rate δ_2 increases, one would expect an increase in service completion and subsequently increase in E_{ds}. Expected number of customers in the orbit E_{orbit} decreases. θ_1^* decreases while θ_2^* increases. The production switch on rate increases, while the expected inventory level E_{inv} and E_p decreases. When δ_2 increases, the decrease in E_{la} is lesser. From Table 9, as δ_2 increases, retrying customers who try to access a free server reduces, so E_{lr} decreases.

Table 9. Effect of retrial rate δ_2 on various performance measures

δ_2	E_{inv}	E_{orbit}	E_{ON}	E_p	E_{ds}	E_{la}	E_{lr}	A_{br}	A_{rr}	θ_1^*	θ_2^*
1.4	1.3843	2.1048	2.955e−11	1.6612	1.2034	0.58619	0.71043	0.16045	0.015879	4.2095	0.65735
1.5	1.3766	2.0893	2.7825e−11	1.6519	1.2117	0.58466	0.7036	0.16157	0.016472	4.1786	0.66062
1.6	1.3698	2.0757	2.6356e−11	1.6437	1.2191	0.5833	0.69759	0.16255	0.017027	4.1514	0.66344
1.7	1.3637	2.0636	2.5096e−11	1.6364	1.2257	0.58208	0.69227	0.16342	0.017547	4.1272	0.66591
1.8	1.3582	2.0528	2.4006e−11	1.6299	1.2315	0.58097	0.68752	0.1642	0.018036	4.1057	0.66808
1.9	1.3533	2.0432	2.3058e−11	1.624	1.2368	0.57997	0.68327	0.1649	0.018496	4.0863	0.67
2	1.3489	2.0344	2.2227e−11	1.6187	1.2415	0.57906	0.67943	0.16553	0.018929	4.0689	0.6717
2.1	1.3448	2.0265	2.1496e−11	1.6138	1.2458	0.57822	0.67596	0.16611	0.019339	4.053	0.67323
2.2	1.3411	2.0193	2.0849e−11	1.6094	1.2497	0.57745	0.6728	0.16663	0.019726	4.0386	0.6746

$S = 18; s = 6; \lambda = 2.5; \mu = 2; \omega = 1.2; \beta = 2.6; \theta = 2; \gamma = 0.7; \delta = 0.8; \delta_1 = 0.4$

8 Conclusion

The paper studied the impact of an unreliable server on a perishable inventory system with production and retrials. Exponential distribution is considered for inter-arrival time as well as the service time. The production is switched ON based on an (s, S) policy. The customer would be allowed to join the orbit if the inventory level is zero or the server is busy or breakdown occurs. The Matrix Geometric Method is used to find the stationary probability vector, which make it easier to obtain key performance measures. A suitable cost function is constructed on the basis of system characteristics. Numerical verification of the convexity of the cost function is conducted. This work can be extended further by considering the arrivals to follow a Markovian Arrival Process (MAP) instead of assuming Poisson arrivals.

References

1. Artalejo, J.R., Krishnamoorthy, A., Lopez-Herrero, M.J.: Numerical analysis of (s, S) inventory systems with repeated attempts. Ann. Oper. Res. **141**(1), 67–83 (2006)
2. Jose, K.P.,Reshmi, P.S.: A production inventory model with deteriorating items and retrial demands. Opsearch **58**(1), 71–82 (2020)
3. Krishnamoorthy, A., Islam, M.E.: Production inventory with random life time and retrial of customers. In: Second National Conference on Mathematical and Computational Models, pp. 89–110. NCMCM (2003)
4. Krishnamoorthy, A., Sathian, M. K., C Viswanath, N.: Reliability of a k-out-of-n system with a single server extending non-preemptive service to external customers-Part II. Reliabil. Theory Appl. **11**(3), 76–88 (2016)
5. Neuts, M.F., Rao, B.: Numerical investigation of a multiserver retrial model. Queue. Syst. **7**(2), 169–190 (1990)
6. Periyasamy, C.: A finite population perishable inventory system with customers search from the orbit. Int. J. Comput. Appl. Math. **12**(1) (2017)
7. Reshmi, P.S., Jose, K.P.: A queueing inventory system with perishable items and retrial of customers. Malaya J. Matematik **7**(2), 165–170 (2019)
8. Reshmi, P.S., Jose, K.P.: A MAP/PH/1 perishable inventory system with dependent retrial loss. Int. J. Appl. Comput. Math. **6**(153) (2020)
9. Shivakumar, B.: A perishable inventory system with retrial demands and a finite population. J. Comput. Appl. Math. **224**(1), 29–38 (2009)
10. Ushakumari, P.: A retrial inventory system with an unreliable server. Int. J. Math. Oper. Res. **10**(2), 190–210 (2017)

Leaf Node Polling Model Analysis in an Integrated Access and Backhaul Network

Dmitry Nikolaev[1(✉)] and Yuliya Gaidamaka[1,2]

[1] Department of Probability Theory and Cybersecurity, Peoples' Friendship University of Russia named after Patrice Lumumba, 6 Miklukho-Maklaya Street, Moscow 117198, Russian Federation
nikolaev-di@rudn.ru, gaydamaka-yuv@rudn.ru

[2] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS), 44-2 Vavilov Street, Moscow 119333, Russian Federation

Abstract. In this paper, a polling service model is used for performance analysis of a leaf node in an Integrated Access and Backhaul (IAB) network. is constructed in order to analyse the performance measures of the system, which will enable the estimation of packet transmission delays in an IAB network. The Markovian polling queueing system with two queues and nonzero switching time is constructed to analyse the performance measures of the system, which will enable the estimation of packet transmission delays in an IAB network. One of the queues is designed to store packets transmitted in downlink from a parent node in the IAB network, and the second queue is designed to store packets transmitted in uplink from child nodes and user equipment associated with an analyzed node. Cyclic polling of queues allows to take into account the main feature of IAB technology — half–duplex mode of packet transmission. Using the apparatus of Markov processes, formulas for calculating the waiting time and the sojourn time of the request in the system, affecting the delay of packet transmission through the node, are obtained. The results of the numerical analysis illustrate an upper bound estimate for the server switching time at which the 5G NR delay constraints are fulfilled for networks with a single relay node.

Keywords: polling · queueing system · integrated access and backhaul · 5G · leaf node

1 Introduction

Integrated Access and Backhaul (IAB) via gNodeB base stations (gNB) for backhaul connections in 5G networks was one of the approved objectives of the 17th Release of the 3GPP (3rd Generation Partnership Project) [1]. When using IAB

The reported study was funded by RSF, project № 23-79-10084, https://rscf.ru/project/23-79-10084/..

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 106–117, 2024.
https://doi.org/10.1007/978-3-031-65385-8_8

technology, only a small fraction of node base stations are connected to the traditional fibre-optic infrastructure of the fixed network, while the remaining nodes carry backhaul traffic over wireless channels [2].

Although existing base station specifications in 3GPP Long Term Evolution Advanced (LTE-Advanced) standards allow backhaul through a single hop link, IAB is a more advanced solution capable of supporting multi-hop links, dynamic resource multiplexing and plug-and-play design, which reduces the complexity of network deployment [3]. Taking into account the above mentioned advantages of IAB technology, designing an efficient and high-performance 5G/6G network using this technology remains a relevant research task [4].

When designing a network, it is necessary to consider not only the radio channel conditions, available radio band, topology of the designed network and energy efficiency constraints, but also the requirements for Quality of Service (QoS), including the delay in service provisioning as one of the most important metrics. Although IAB technology supports the ability to transmit data in bidirectional full duplex mode, most implementations will be restricted to unidirectional half duplex mode, which causes increased latency. Half duplex mode obliges us to be particularly careful about the value of end-to-end data transmission delay.

In this paper, to analyse a single node performance in an IAB network with half duplex mode of data transmission, a Markovian queueing system in the form of a polling service model with two queues, non-zero switching time between them and packet arrivals during the server switching periods is proposed. Among the obtained performance measures of the system, the emphasis is on the delays and access blockages, for which numerical analyses have been carried out.

2 System Model

Figure 1 schematically depicts an example of IAB network topology in the form of a Spanning Tree (SP) with the root in the IAB-donor, the only node having a wired connection to the fibre optic network, with all other IAB-nodes connected only to one parent node carrying traffic over wireless channels. The object of study is packet flow through the boundary IAB-node corresponding to the leaf node of the tree, and the subject of study are performance measures of this flow.

In the case of half-duplex transmission mode, the packets arrive at the IAB-node alternately

- from the parent node through the downlink channel;
- from child nodes and from User Equipment (UE) associated with the considered node through the uplink channels.

Half-duplex mode forbids simultaneously enabling

- an access link and a backhaul link on the same node;
- bidirectional data transmission on the access link or on the backhaul link.

In the following sections, a half-duplex service model with two queues, cyclic queue traversal, non-zero switching time between queues, and requests arriving

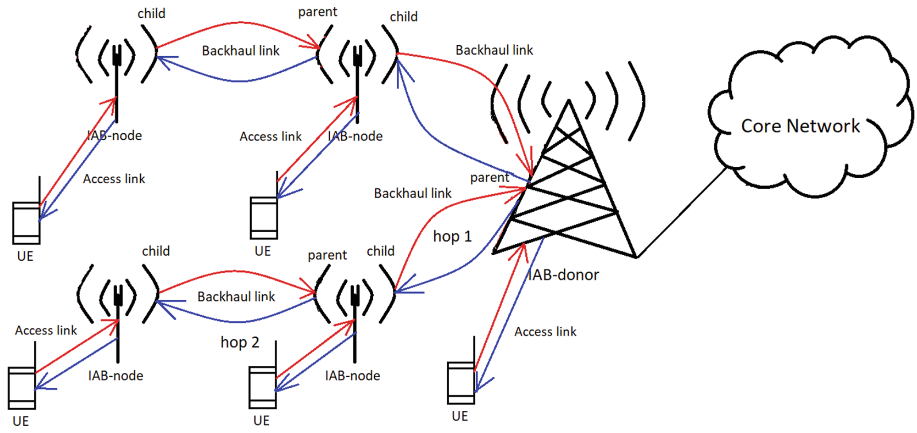


Fig. 1. IAB network fragment in the form of a spanning tree.

exclusively during the switching periods due to the above mentioned half-duplex constraints is proposed to describe the data packet processing at the IAB leaf nodes. Thus, one of the queues will be associated with the downlink channel and the other with the uplink channel, and the packets transmitted through these channels will correspond to the requests in the queues. The aim of the study is to analyse the dependence of the request waiting time on the duration of the switching interval between queues, which corresponds to the duration of the packet arrival phase in the IAB-node from the parent node and the UE in half-duplex mode.

It is worth noting that the apparatus of queueing theory [5,6] known in the analysis of telecommunication systems has so far been applied to IAB networks only in a few cases [7,8]. In the first case, by means of building a discrete model with probabilistic service, the packet delay was estimated, and in the second case, the limiting distribution of the number of users at the IAB-donor and IAB-node was found.

3 Polling Service Model

The ordered polling [9–12] means that the server switches from queue to queue according to some given switching rule and serves the requests in each queue according to a given service rule. Polling systems with two queues and exponential distributed inter-arrival intervals are investigated in [13–15]. Thus, in [15] analytical expressions for average waiting times of requests in the queue for gateway and exhaustive service disciplines are found.

To analyse the process of data packet transmission by the IAB leaf node, we propose a model of polling service with two finite queues $M_2|M_2|1|(r_1,r_2)$, corresponding to the conditions described in the previous section, the scheme of which is shown in Fig. 2. The queue Q_1 receives requests from the parent node through the downlink and the queue Q_2 receives requests from the UEs through the uplink. Due to the peculiarities of the proposed system, it is not possible to unambiguously define its service discipline, as it is exhaustive due to the service of all the requests received in the queue and, at the same time, global-gated, as from the instant of the service cycle's start new requests are not allowed to join to the"gated" requests which were already in queues at the beginning of the current polling cycle. As an assumption, let us assume that switching from the first queue to the second queue is instantaneous, and switching from the second queue to the first queue takes non-zero time, that models simultaneous requests arrival to both queues.

As a result, the system will operate according to the following cycle: requests arrival (i.e., filling of the queues while server switches from queue Q_2 to Q_1) — queue Q_1 service (if empty, then immediately switching to Q_2) — queue Q_2 service.

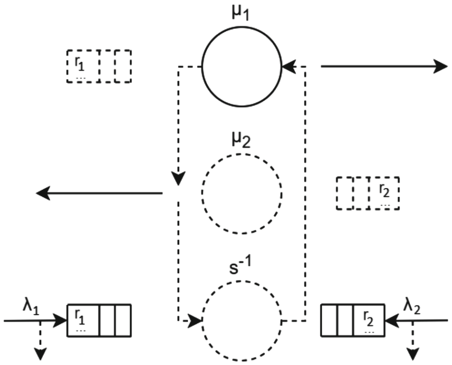


Fig. 2. Polling service model with two queues.

Since the model is assumed to be Markovian, the durations of intervals between neighbouring arrivals of requests to the system, the durations of requests servicing and the switching time between queues have exponential distribution, which corresponds to 3GPP recommendations on IAB network modelling.

Let us introduce the system parameters: λ_i — intensities of requests arrival in Q_i; μ_i — intensities of requests servicing; r_i — storage capacity; s^{-1} — intensity of server switching from Q_2 to Q_1 (i.e., average switching time is equal to s), where $i = 1, 2$.

The functioning of the system is described by a random process $\mathbf{X}(t)$:

$$\mathbf{X}(t) = \{(q(t), n_1(t), n_2(t)), t \geq 0\}, \tag{1}$$

where $q(t) \in \{0, 1, 2\}$ — the state of the server ($q = 0$ — requests arrival, $q = 1, 2$ — service of the first and second queues, respectively), $n_i(t) \in \{0, 1, \ldots, r_i\}$ — the number of requests in Q_i at time t.

The main feature of the system reflecting the half-duplex mode of packet transmission in the IAB network can be formulated as follows:

$$\lambda_i(q) = \begin{cases} \lambda_i, & q = 0 \text{ (server switching)}, \\ 0, & q = 1,2 \text{ (service of the queue } Q_1 \text{ or } Q_2\text{)}, \end{cases} \quad i = 1,2. \tag{2}$$

The state space of this system has the following form:

$$\begin{aligned} \mathbb{X} = \{&(0, n_1, n_2) : n_1 \in \{0,1,\dots,r_1\}, n_2 \in \{0,1,\dots,r_2\}; \\ &(1, n_1, n_2) : n_1 \in \{1,\dots,r_1\}, n_2 \in \{0,1,\dots,r_2\}; \\ &(2, 0, n_2) : n_2 \in \{1,\dots,r_2\}\} \end{aligned} \tag{3}$$

which cardinality is equal to

$$|\mathbb{X}| = (r_1+1)(r_2+1) + r_1(r_2+1) + r_2. \tag{4}$$

Note that it can be represented as a union of three non-intersecting subspaces, i.e.

$$\mathbb{X} = \mathbb{X}_0 \cup \mathbb{X}_1 \cup \mathbb{X}_2, \tag{5}$$

where $\mathbb{X}_0 = \{(0, n_1, n_2) : n_1 \in \{0,1,\dots,r_1\}, n_2 \in \{0,1,\dots,r_2\}\}$ — state space of arriving requests into the system (switching of the server), $\mathbb{X}_1 = \{(1, n_1, n_2) : n_1 \in \{1,\dots,r_1\}, n_2 \in \{0,1,\dots,r_2\}\}$ — queue Q_1 service state space, $\mathbb{X}_2 = \{(2, 0, n_2) : n_2 \in \{1,\dots,r_2\}\}$ — queue Q_2 service state space.

Let us now introduce access blocking state spaces. In our system the requests are blocked in case of overflow of the queues only during the server switching phase (because requests arrive to the system only during this phase). Then the blocking state space of Q_i is expressed as:

$$\mathfrak{B}_i = \{(q, n_1, n_2) : q = 0, n_i = r_i\}, \quad i = 1,2. \tag{6}$$

So the blocking state space of partial blocking (i.e., at least one of the queues) is represented as:

$$\mathfrak{B} = \mathfrak{B}_1 \cup \mathfrak{B}_2 = \mathfrak{B}_1 + \mathfrak{B}_2 - \mathfrak{B}_1 \cap \mathfrak{B}_2, \tag{7}$$

where the intersection $\mathfrak{B}_1 \cap \mathfrak{B}_2$ is the full blocking state space represented by the only state:

$$\mathfrak{B}_1 \cap \mathfrak{B}_2 = (0, r_1, r_2). \tag{8}$$

The main task is to find the stationary distribution of the process $\mathbf{X}(t)$

$$p(q, n_1, n_2) = \lim_{t\to\infty} \mathsf{P}\{\mathbf{X}(t) = (q(t), n_1(t), n_2(t))\}, (q, n_1, n_2) \in \mathbb{X}. \tag{9}$$

To find it, it is necessary to solve the system of equilibrium equations (or Equilibrium System, ES), which can be compiled using the graph of transition intensities.

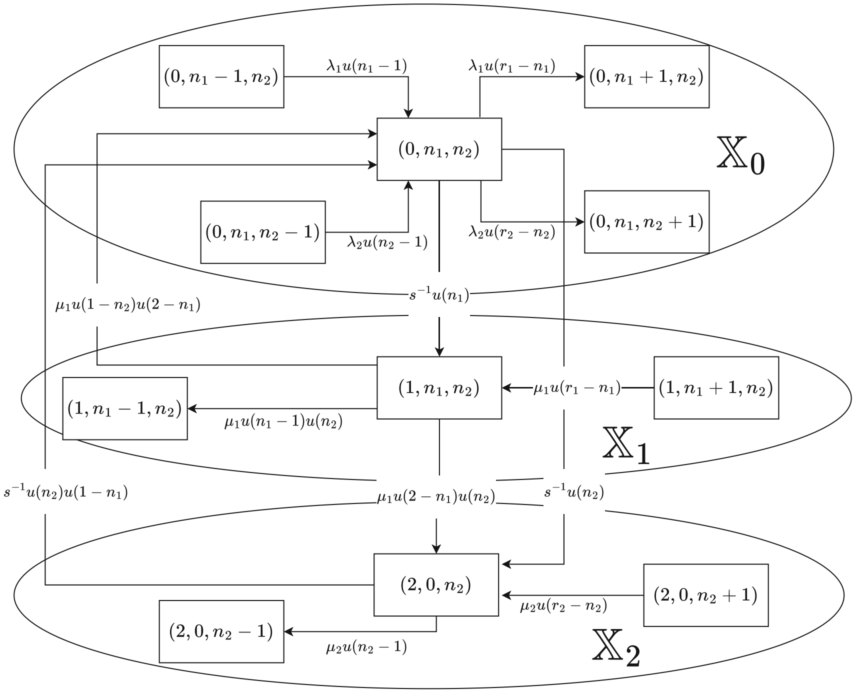


Fig. 3. Transition intensity graph with three subsets of states.

Figure 3 shows the fragment of the transition intensity graph of this system. It consists of three groups of states belonging to each of the three non-intersecting state spaces, respectively.

Now, using the graph shown in Fig. 3 and the principle of global balance, let's construct an ES:

$$\begin{cases} \mu_1 p(1,1,0) + \mu_2 p(2,0,1) = (\lambda_1 + \lambda_2) p(0,0,0), \\ \lambda_1 p(0, i-1, 0) = (u(r_1 - i)\lambda_1 + \lambda_2 + s^{-1}) p(0, i, 0), \quad i = 1, ..., r_1, \\ \lambda_2 p(0, 0, i-1) = (\lambda_1 + u(r_2 - i)\lambda_2 + s^{-1}) p(0, 0, i), \quad i = 1, ..., r_2, \\ \lambda_1 p(0, i-1, j) + \lambda_2 p(0, i, j-1) = (u(r_1 - i)\lambda_1 + u(r_2 - j)\lambda_2 + s^{-1}) p(0, i, j), \\ i = 1, ..., r_1, j = 1, ..., r_2, \\ s^{-1} p(0, i, j) + u(r_1 - i)\mu_1 p(1, i+1, j) = \mu_1 p(1, i, j), \quad i = 1, ..., r_1, j = 0, ..., r_2, \\ s^{-1} p(0, 0, j) + \mu_1 p(1, 1, j) + u(r_2 - j)\mu_2 p(2, 0, j+1) = \mu_2 p(2, 0, j), \quad j = 1, ..., r_2, \end{cases} \tag{10}$$

where $u(x)$ is the Heaviside function

$$u(x) = \begin{cases} 0, & x \leq 0, \\ 1, & x > 0. \end{cases} \tag{11}$$

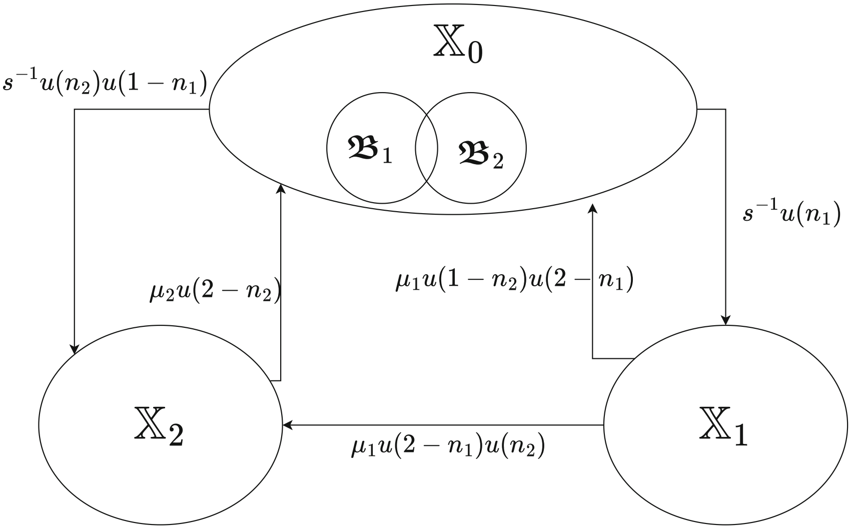


Fig. 4. Enlarged graph of intensities of transitions between non-intersecting spaces.

For a better understanding of the transitions between non-intersecting spaces, a reduced graph of the system transitions is presented in Fig. 4. Thus, the left part of the first ES equation is formed by transitions from $\mathbb{X}_1$ and $\mathbb{X}_2$ to $\mathbb{X}_0$; the summands of the second, third and fourth equations containing the multiplier s^{-1}, are responsible for transitions to states of spaces $\mathbb{X}_1$ or $\mathbb{X}_2$ from $\mathbb{X}_0$; while the remaining summands represent transitions within spaces.

We can write the ES more compactly, where each equation is related to one of the spaces:

$$\begin{cases} u(1-i-j)(\mu_1 p(1,1,0)+\mu_2 p(2,0,1))+u(i)\lambda_1 p(0,i-1,j)+u(j)\lambda_2 p(0,i,j-1)= \\ =(u(r_1-i)\lambda_1+u(r_2-i)\lambda_2+u(i+j)s^{-1})p(0,i,j), \quad i=0,\dots,r_1, j=0,\dots,r_2, \\ s^{-1}p(0,i,j)+u(r_1-i)\mu_1 p(1,i+1,j)=\mu_1 p(1,i,j), \quad i=1,...,r_1, j=0,...,r_2, \\ s^{-1}p(0,0,j)+\mu_1 p(1,1,j)+u(r_2-j)\mu_2 p(2,0,j+1)=\mu_2 p(2,0,j), \quad j=1,...,r_2. \end{cases} \tag{12}$$

Solving the ES together with the normalisation condition

$$\sum_{(q,n_1,n_2)\in\mathbb{X}} p(q,n_1,n_2)=1, \tag{13}$$

we find the stationary distribution. Thus allows us to obtain the analytical expressions for desired performance measures of the system. The following metrics are essential in further numerical analysis.

Probability for the system to be in a state of queues filling

$$P_0 = \mathsf{P}\{(q, n_1, n_2) \in \mathbb{X}_0\} = \sum_{n_1=0}^{r_1} \sum_{n_2=0}^{r_2} p(0, n_1, n_2). \tag{14}$$

Probability for the system to be in a state of queue Q_1 service

$$P_1 = \mathsf{P}\{(q, n_1, n_2) \in \mathbb{X}_1\} = \sum_{n_1=1}^{r_1} \sum_{n_2=0}^{r_2} p(1, n_1, n_2). \tag{15}$$

Probability for the system to be in a state of queue Q_2 service

$$P_2 = \mathsf{P}\{(q, n_1, n_2) \in \mathbb{X}_2\} = \sum_{n_2=1}^{r_2} p(2, 0, n_2). \tag{16}$$

Blocking probability of a request for queue Q_1

$$B_1 = \mathsf{P}\{(q, n_1, n_2) \in \mathfrak{B}_1\} = \sum_{n_2=0}^{r_2} p(0, r_1, n_2). \tag{17}$$

Blocking probability of a request for queue Q_2

$$B_2 = \mathsf{P}\{(q, n_1, n_2) \in \mathfrak{B}_2\} = \sum_{n_1=0}^{r_1} p(0, n_1, r_2). \tag{18}$$

Probability of partial blocking (at least one of the queues)

$$B = \mathsf{P}\{(q, n_1, n_2) \in \mathfrak{B}\} = B_1 + B_2 - B_{12} = \sum_{n_2=0}^{r_2} p(0, r_1, n_2) + \sum_{n_1=0}^{r_1-1} p(0, n_1, r_2), \tag{19}$$

where $B_{12} = \mathsf{P}\{(q, n_1, n_2) \in \mathfrak{B}_1 \cap \mathfrak{B}_2\} = p(0, r_1, r_2)$.

Average number of requests in queue Q_i, where if $i = 2$, then $i + 1 = 1$

$$\overline{N_i} = \sum_{n_i=1}^{r_i} n_i \sum_{q=0}^{2} \sum_{n_{i+1}}^{r_{i+1}} p(q, n_1, n_2), \quad i = 1, 2. \tag{20}$$

Average waiting time in queue Q_i

$$\omega_i = \frac{\overline{N_i}}{\lambda_i(1 - B_i)}, \quad i = 1, 2, \tag{21}$$

which formulas are obtained using Little's law and considering blocking probabilities of corresponding queues.

The next section provides a numerical analysis of the average waiting time of a request in a queue, the access blocking of at least one of the queues and the probability of non-receipt of requests (i.e., the probability of servicing one of the queues) as a function of the switching intensity between queues.

4 Numerical Analysis

The initial data for the numerical experiment are taken from [17], which considers scenarios for FR2 band, 200 MHz of bandwidth and subcarrier spacing of 120 kHz corresponding to the NR numerology 3. The uplink and downlink overheads are defined by following 3GPP TS 38.306 [16]. To determine the service intensity parameters we will use the formula of throughput capacity C from 3GPP standard [18]

$$C = 10^{-6}\nu \mathcal{Q} f R \frac{12N}{T}(1 - H), \tag{22}$$

where the number of multiplexed layers ν put equal to 1, scaling factor $f = 0.75$, modulation index $\mathcal{Q}$ put equal to 6 (i.e., assume the use of quadrature amplitude modulation QAM64), coding error R take equal to $\frac{438}{1024}$. Symbol length $T = 8.92{\cdot}10^{-6}$ and the number of resource blocks $N = 132$ unambiguously determined using the third numerology and the bandwidth of 200 MHz. The overlap factor $H = 0.18$ is given by the standard.

Applying abovementioned formula, we obtain a downlink access channel capacity of 280 Mbit/s. Considering the request size equal to the maximum TCP packet size, i.e. 1500 bytes, and subframe duration in NR, i.e. 1 ms [18], we convert the value of the throughput capacity from $C_0 = 280$ Mbit/s to C_1 measured in number of packets per subframe using the following formula

$$C_1 = 10^{-3}\frac{C_0 \cdot 10^6}{8 \cdot 1500} = \frac{C_0}{12} \approx 23. \tag{23}$$

As a result, we get the average number of transmitted packets per subframe (for 1 ms) equal to 23. This number will be considered as the first queue requests service intensity.

Since backhaul channels are distinguished by the possibility of using large antenna arrays, the formation of a directional beam and a stable state of line of sight, we will consider the average number of packets transmitted through the backhaul channel (also having size equal to 1500 bytes) for 1 ms in one and a half times more than the obtained value for the access channel. Consequently, we will take the second queue requests service intensity as 34.

It is worth noting that similar approach to the determining the initial parameters for the numerical analysis of the IAB network in the discrete case was used in [17] for other numerologies as well.

Figure 5 plots the mean waiting time for requests in queues Q_1, Q_2 with following initial parameters: $\mu_1 = 23$, $\mu_2 = 34$, $\lambda_1 = 12$, $\lambda_2 = 16$, $r_1 = r_2 = 100$, where $\mu_1, \mu_2, \lambda_1, \lambda_2$ have values of packets/ms (or requests/ms). As shown in Fig. 5, in order to satisfy the 1 ms limit to an end-to-end delay imposed by the 5G network standards, the switching duration of a single-hop IAB network must satisfy the upper estimate of $s < 0.8$ ms (since $s^{-1} > 1.25$). For a multi-hop IAB network, such an estimate will not fit, and for a more accurate estimate, the number and parameters of links and relay nodes must be taken into account. In particular, it is necessary to take into account the transmission rates in the radio links connecting the transit nodes, the duration of the phases of the half-duplex transmission mode and the processor performance of the transit nodes.

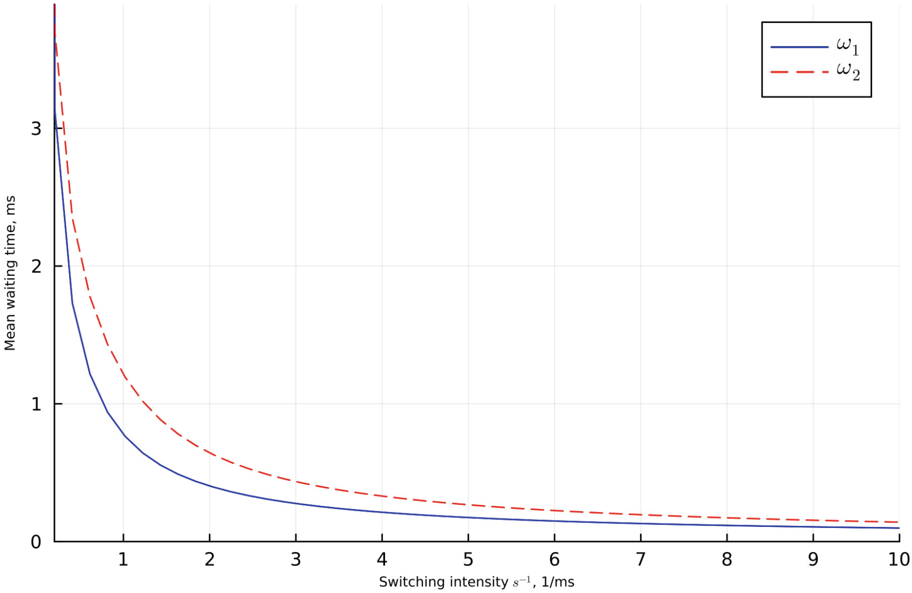


Fig. 5. Average waiting times of requests in queues Q_1 and Q_2.

Figure 6 shows the graphs of probabilities of queue service by the server, blocking access to at least one of the queues and their sum depending on the intensity of switching between queues. The graph in Fig. 6 shows that starting from approximately $s^{-1} > 1.5$ the probabilities of arrival and non-arrival (probabilities of queue service together with access blocking) of requests coincide. This is explained by the fact that the total proposed load $\rho = \rho_1 + \rho_2 \approx 1$, where $\rho_i = \frac{\lambda_i}{\mu_i}$ $(i = 1, 2)$, that is, in total, the time of arrival of requests (with almost no blocking) will coincide with the time of their service, which gives us equal probabilities with the value of ≈ 0.5. Thus, as the total load decrease, the probabilities described above will also decrease.

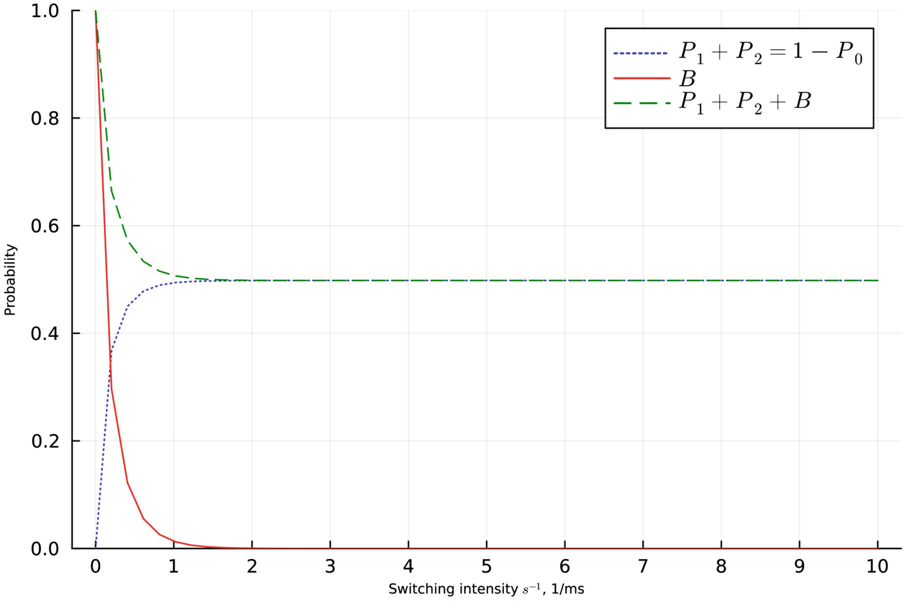


Fig. 6. Probabilities of blocking and queue service.

5 Conclusion

Integrated access and backhaul technology is considered one of the key technologies in the transition to next-generation networks, as it enables the deployment of dense networks without additional connections to the core network. However, due to the use of multi-hop retransmission in half-duplex mode, additional verification of fullfilment of delay requirements is necessary.

To study the packet delay in a half-duplex IAB network, this paper provides a model of the IAB-node in the form of a polling queueing system. The study of the system by methods of mathematical teletraffic theory and Markov processes, allows to estimate the performance measures of the system at a range of initial parameters values and to make a choice of the most appropriate values necessary for the construction of the physical implementation of the technology. Considering packet delay and blocking probabilities as key metrics, we have performed a numerical analysis that allows us to give an upper bound estimate for the switching duration corresponding to the 5G NR requirements. Note that some other QoS paramrters can also be estimated based on the constructed model.

One of the objectives of further research is the formulation and solution of the optimisation problem for the construction of the most efficient multi-hop IAB network, in which the load on the network nodes will be balanced, and the quality of service for users will remain high.

References

1. 3GPP Technical Specification Group Services and System Aspects (2020) Technical Report TR 21.917, Release 17 description; Summary of Rel-17 Work Items (Release 17)
2. ETSI Technical Specification TS 138 474 V16.0.0 (2020-07) 5G NG-RAN. F1 data transport (3GPP TS 38.874 version 16.0.0 Release 16)
3. Polese, M., et al.: Integrated Access and Backhaul in 5G mmWave networks: potential and challenges. IEEE Commun. Mag. **58**(3), 62–68 (2020). https://doi.org/10.1109/MCOM.001.1900346
4. Molchanov D.A., Begishev V.O., Samuylov K.E., Koucheryavy D.A.: 5G/6G networks: architecture, technologies, analysis, and calculation methods. M.: RUDN Publishing House, 516 p (2022)
5. Moiseev, A.N., Nazarov, A.A.: Infinitely Linear Systems and Queueing Networks, p. 240. STL Publishing House, Tomsk (2015)
6. Nazarov, A.A., Terpugov, A.F.: Queueing theory: textbook, p. 228. STL Publishing House, Tomsk (2010)
7. Khayrov E., Koucheryavy Y.: Packet level performance of 5G NR system under blockage and micromobility impairments. IEEE Access **11**, 90383–90395 (2023). https://doi.org/10.1109/ACCESS.2023.3307021
8. Salimzyanov, R., Moiseev, A.: Local balance equation for the probability distribution of the number of customers in the IAB nerwork, pp. 284–289. Omsk, SUITMM (2023)
9. Vishnevskiy V.M., Semenova, O.V.: Polling systems: Theory and application in the broadband wireless networks. M.: Technosphere, 312 p (2007)
10. Rykov V.V.: To the analysis of polling systems. Autom. Remote Control **70** (6), 997–1018 (2009)
11. Takagi H.: Analysis of polling systems, p. 175. MIT Press (1986)
12. Takagi H., Kleinrock L.: A tutorial on the analysis of polling systems. Computer Science Department, University of California, Los Angeles. Report No. CSD-850005, p. 172 (1985)
13. Takagi H.: Mean message waiting times in symmetric multiqueue systems with cyclic service. Performance Evaluation, 271–277 (1985)
14. Gaidamaka Yu, V., Zaripova, E.R.: Model of SIP-server with gateway and exhaustive queuing service disciplines. Bulletin of PFUR. Ser. Math. Inform. Phys. 1, 52–57 (2013)
15. Orlov, Y., Gaidamaka, Y., Zaripova, E.: Approach to estimation of performance measures for sip server model with batch arrivals. In: Vishnevsky, V., Kozyrev, D. (eds.) DCCN 2015. CCIS, vol. 601, pp. 141–150. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30843-2_15
16. 3GPP. NR; User Equipment (UE) radio access capabilities (Release 17). Technical Specification (TS) 38.306, 3rd Generation Partnership Project (3GPP). Version V17.6.0 (2023-09)
17. Yarkina, N., Moltchanov, D., Koucheryavy, Y.: Counter waves link activation policy for latency control in In-Band IAB systems. IEEE Commun. Lett. **27**(11), 3108–3112 (2023). https://doi.org/10.1109/LCOMM.2023.3313233
18. 3GPP. NR; Physical channels and modulation (Release 18). Technical Specification (TS) 38.211, 3rd Generation Partnership Project (3GPP). Version V18.0.0 (2023-09)

A Semi-Markovian Analysis of an Inventory Model with Inventory-Level Dependent Arrival and Service Processes

K. R. Ranjith[1,2](✉), B. Gopakumar[3], and Sajeev S. Nair[4]

[1] Government Engineering College, Thrissur, India
kr.ranjith3@gmail.com
[2] APJ Abdul Kalam Technological University, CET Campus, Thiruvananthapuram 695016, Kerala, India
[3] Maharaja's Technological Institute, Chembukkavu, Thrissur 680020, Kerala, India
[4] Government Polytechnic College Chelakkara Thonoorkara(PO), Chelakkara, Thrissur 680586, Kerala, India

Abstract. Two Semi Markov processes are defined to describe the service and inter-arrival times of an $s - S$ inventory model with zero lead time, in which both inter-arrival time and service time depend upon the inventory level. It is assumed that both the service time and the inter-arrival time follow Phase-Type distributions, which are determined by the current inventory level. The marginal distributions of both the service time and the inter-arrival time are obtained. A continuous parameter Markov chain is used to model the queue size. Condition for stability and the steady state characteristics of the system are derived. The impact of interdependence between service and arrival processes, along with inventory level on the system is examined. Furthermore, a numerical analysis is also done to explain the consequences of this dependency on steady-state system characteristics.

Keywords: Interdependent processes · Semi-Markov Process · Matrix analytic method · $s - S$ inventory model

1 Introduction

Inventory models had received significant attention in academic research. While numerous classical models assume negligible or no time for the inventory to be served, real-world scenarios often demand considerations of the time to serve the inventory. Pioneering the exploration of Inventory models with positive service time were Berman et al. [12] and Sigman et al. [14]. Comprehensive insight into studies by various authors in this direction can be found in the survey articles by Krishnamoorthy et al. [8,9].

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 118–133, 2024.
https://doi.org/10.1007/978-3-031-65385-8_9

Unexpectedly, there have been relatively few prior studies exploring models where the service time and/or inter-arrival time are contingent upon the inventory level. This stands in contrast to the significant emphasis placed on queueing models with state-dependent arrival and service processes in the existing literature. S. K. Gupta [4] analyzed such a model with a finite queue size. In his investigation, both arrival and service rates were treated as arbitrary functions of the number of customers in the system. Other pioneering works in a similar direction were done by Hiller et al. [6] as well as Conway et al. [3]. Later, Bekker et al. [2] provided detailed descriptions of both M/G/1 and G/G/1 models incorporating workload-dependent arrival and service rates.

However, there are classical models (with zero service time) that have analyzed inventory models with stock dependent demands. These models are derived from the observation that maintaining abundant inventory has a favourable impact on demand.

Customers are often enticed to make purchases by prominently displaying a considerable amount of inventory in stores. Larson et al. [10] coined the term "psychic stock" for this displayed inventory. Hadley et al. [5], Wolfe [16], T L Urban [15] and Johnson [7] are among the researchers who have investigated the stimulating impact of inventory level on demand. The influence of inventory level on demand is evident in cases involving distinct inventory items, such as ornaments or clothing materials. A diverse assortment of these items offers customers a wider choice, consequently elevating demand. Another instance is when there's an abundance of inventory or when dealing with perishable items. In these situations, sellers may introduce special offers to entice buyers and clear out excess/old stock, resulting in a notable increase in demand.

Queueing models with interdependent arrival and service processes are introduced by Ranjith et al. [13]. Much before that Guy Latouche [11] derived interdependent phase type processes by a semi Markovian point process. He constructed this by considering a finite state irreducible Markov chain. In this paper we follows a similar approach. But with the difference that our focus is not on the dependence between these processes, but on the interdependence between the state of the embedded Markov chain and the phase type distribution constructed.

Using this method, this paper analyses an $s - S$ inventory model with no lead time and positive service time. In this model both the service and arrival processes depend on the inventory level. The structure of the paper is in the following manner. Section 2 presents the description of the underlying Markov chains for the service and arrival processes. In Sect. 3, the marginal distribution of the service time and inter arrival time are found. In Sect. 4, a continuous-time Markov chain is employed to model a queueing-inventory model with inventory dependent arrival and service processes. It also covers the condition for stability of the system. Section 5 focuses on the steady state analysis and evaluation of the key system performance measures. Furthermore, we conduct a comprehensive numerical investigation of the system in Sect. 6. Finally, in Sect. 7 we conclude the discussion.

2 Semi-Markovian Service and Arrival Processes Depending on Inventory Level

To construct Semi-Markovian point processes suitable for modeling arrival and service processes of a queueing-inventory model in which these two processes depend on the level of inventory, we proceed as follows:-

Consider an irreducible Markov chain $\mathcal{X} = \{X_i | i = 0, 1, 2, 3, ...\}$ with state space $\{1, 2, 3, ..., r\}$. Let $P^X = \left[p_{ij}^X\right]$ where

$$p_{ij}^X = \begin{cases} 1 \text{ if } j = i - 1, 2 \leq i \leq r \\ 1 \text{ if } i = 1, j = r \\ 0 \text{ otherwise.} \end{cases}$$

be the transition probability matrix of the chain $\mathcal{X}$.

Assume that the transitions of the chain $\mathcal{X}$ occur at random epochs $\gamma_i, i = 1, 2, 3, \ldots$. Let τ_i be the interval of time between the successive transitions.

$$\tau_i = \begin{cases} \gamma_i - \gamma_{i-1}, \text{ if } i = 2, 3, ... \\ \gamma_1 \text{ if } i = 1 \end{cases}$$

For each i, if $X_{i-1} = j$, assume that τ_i follows a Phase type distribution $F_j(.)$ with representation (α_j, D_j), where D_j is an $n_j \times n_j$ matrix. Thus we have a semi-Markov Process $\{Z_X(t) | t \geq 0\}$ defined by

$$Z_X(t) = X_i, \;\; \gamma_i \leq t < \gamma_{i+1}, \;\; i = 0, 1, 2, ...$$

In the present study, states of the chain $\mathcal{X}$ are the inventory levels and each state transition corresponds to a service completion. $\gamma_i, i = 1, 2, 3, ...$ are the epochs of completion of i^{th} service, $Z_X(t)$ represents the inventory level at time t and the distribution of duration of the service happening at time t is $F_{Z_X(t)}(.)$.

For the arrival process, we proceed as follows:- Consider the phase type distributions $G_i(.)$ with representations $(\beta_i, T_i), i = 1, 2, .., r$ where T_i is a square matrix of order m_i. Define a Markov chain $\mathcal{Y} = \{Y_i | i = 0, 1, 2, 3, ...\}$ with state space {0,1,2,...}, $Y_0 = 0$ and having the transition probability matrix $P^Y = \left[p_{ij}^Y\right]$ where

$$p_{ij}^Y = \begin{cases} 1 \text{ if } j = i + 1, i \geq 0 \\ 0 \text{ otherwise.} \end{cases}$$

Starting from time $t = 0$, let ν_i be the epoch at which the chain $\mathcal{Y}$ makes the i^{th} transition, $i = 1, 2, 3, ...$ Let φ_i be the inter occurrence time $\nu_i - \nu_{i-1}$ between the $i - 1^{\text{th}}$ and i^{th} transitions of the chain $\mathcal{Y}$. Assume that φ_i follows the distribution $G_j(.)$ where $j = Z_X(\nu_{i-1})$. Hence we have a semi Markov Process

$$Z_Y(t) = Y_i, \nu_i \leq t < \nu_{i+1}, i = 0, 1, 2, ...$$

We may take the states of the chain $\mathcal{Y}$ to be the number of arrivals. The distribution of the inter arrival times are then determined by the inventory level at the epochs of the preceding arrivals.

3 Marginal Distributions of Service Times and Inter-arrival Times

Consider the continuous parameter Markov chain

$$\mathcal{N}_1 = \{(N_1(t), I_X(t), J_\tau(t)) | t \geq 0\}$$

where $N_1(t)$ is the number of transitions occurred during the time interval $(0, t]$ of the chain $\mathcal{X}$, $I_X(t)$ is the state of the chain $\mathcal{X}$ and $J_\tau(t)$ is the phase of the distribution of the ongoing service process at time t. The state space of this process is

$$\bigcup_{i=1}^{r} \{(n, i, j) | n = 0, 1, 2, ..., j = 1, 2, 3, ..., n_i\}$$

Since in the steady state all the states of the chain $\mathcal{X}$ are equally likely, the initial probability distribution of the chain $\mathcal{N}_1$ is given by $\frac{1}{r}\tilde{\alpha}$, where $\tilde{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_r)$ The infinitesimal generator of the chain $\mathcal{N}_1$ is

$$Q_s = \begin{bmatrix} U & U^0 & 0 & 0 & ... \\ 0 & U & U^0 & 0 & ... \\ 0 & 0 & U & U^0 & ... \\ . & . & . & . & ... \\ . & . & . & . & ... \end{bmatrix}$$

where

$$U = diag(D_1, D_2, ..., D_r)$$

and

$$U^0 = \begin{bmatrix} 0 & 0 & 0 ... 0 & 0 & D_1^0\alpha_1 \\ D_2^0\alpha_2 & 0 & 0 ... 0 & 0 & 0 \\ 0 & D_3^0\alpha_3 & 0 ... 0 & 0 & 0 \\ . & . & . ... . & . & . \\ . & . & . ... . & . & . \\ 0 & 0 & 0 ... 0 & D_r^0\alpha_r & 0 \end{bmatrix}$$

Here $D_i^0 = -D_i e, i = 1, 2, ..., r.$

The service time τ_i of the i^{th}customer is the time taken by the chain $\mathcal{N}_1$ for the transition from level i to level $i+1$. From the infinitesimal generator of $\mathcal{N}_1$ it follows that τ_i's are identically distributed and that their common marginal distribution $F(t)$ is of phase type with representation $\left(\frac{1}{r}\tilde{\alpha}, U\right)$.

Therefore,

$$\begin{aligned} F(t) &= 1 - \frac{1}{r}\tilde{\alpha} exp(Ut)e \\ &= 1 - \frac{1}{r}\sum_{i=1}^{r} \alpha_i exp(D_i t)e \end{aligned}$$

where e is a column vector of 1's of appropriate order and the density function is

$$f(t) = \frac{1}{r}\sum_{i=1}^{r} \alpha_i exp(D_i t)D_i^0 = \frac{1}{r}\sum_{i=1}^{r} F_i'(t)$$

Thus in the steady state, marginal density of the service time is the mixture of the densities $F_1'(.), F_2'(.), ..., F_r'(.)$.

To determine the marginal distribution of the inter arrival time, we consider the Markov chain

$$\mathcal{N}_2 = \{(N_2(t), I_X(\nu), I_X(t), J_\varphi(t))\,|t \geq 0\}$$

where $N_2(t)$ is the number of transitions of the chain $\mathcal{Y}$ occurred in the interval $(0,t]$, $I_X(\nu)$ and $I_X(t)$ are the states of the chain $\mathcal{X}$ at time ν and t respectively where $\nu = max\{\nu_j \in (o,t]\}$ and $J_\varphi(t)$ is the phase of the ongoing arrival process at time t. Note that $N_2(t)$ is the total number of arrivals in $(0,t]$ and $Z_X(\nu)$ is the inventory level at the epoch of previous arrival. This chain has the state space

$$\bigcup_{i=1}^{r} \{(n,i,k,j)|n = 0,1,2,...,k = 1,2,...,r, j = 1,2,3,...,m_i\}$$

and the infinitesimal generator

$$Q_V = \begin{bmatrix} V & V^0 & 0 & 0 & ... \\ 0 & V & V^0 & 0 & ... \\ 0 & 0 & V & V^0 & ... \\ . & . & . & . & ... \\ . & . & . & . & ... \end{bmatrix}$$

where

$$V = diag(V_1, V_2, ..., V_r)$$

and

$$V^0 = \begin{bmatrix} V_1^0 \\ V_2^0 \\ \vdots \\ V_r^0 \end{bmatrix}$$

in which

$$V_i = \begin{bmatrix} T_i - \mu_1 I_{m_1} & 0 & 0 & ... & 0 & \mu_1 I_{m_1} \\ \mu_2 I_{m_2} & T_i - \mu_2 I_{m_2} & 0 & ... & 0 & 0 \\ . & . & . & ... & . & . \\ . & . & . & ... & . & . \\ 0 & 0 & 0 & ... & \mu_r I_{m_r} & T_i - \mu_r I_{m_r} \end{bmatrix},$$

$\mu_i = -\alpha_i D_i^{-1} e$ and V_0^i is a block matrix of order $r \times r^2$ with $T_i^0 \beta_i$ at positions $(j, (j-1)r + j)$ and zero matrices at every other positions.

Hence in the steady state, the distribution $G(.)$ of the inter arrival time is a Phase type distribution with representation $(\frac{1}{r}\tilde{\beta} \otimes \zeta, V)$ where $\tilde{\beta} = (\beta_1, \beta_2, \ldots, \beta_r)$ and $\zeta = \left(\frac{1}{r}, \frac{1}{r}, \ldots \frac{1}{r}\right)$ is the stationary probability vector of the chain $\mathcal{X}$. Therefore, we have

$$\begin{aligned} G(t) &= 1 - \left(\frac{1}{r}\tilde{\beta} \otimes \zeta\right) exp(Vt)e \\ &= 1 - \frac{1}{r}\sum_{i=1}^{r}(\beta_i \otimes \zeta) exp(V_i t)e \end{aligned}$$

and the density function is

$$g(t) = \frac{1}{r}\sum_{i=1}^{r}(\beta_i \otimes \zeta) exp(V_i t)\left(T_i^0 \otimes e_r\right)$$

where e_r is a column vector of 1's of dimension $r \times 1$. Thus in the steady state, marginal distribution of the service time is the mixture of the phase type distributions with representations $((\beta_i \otimes \zeta), V_i), i = 1, 2, ..., r.$

A simplified version of the model we discussed so far may be obtained by assuming that D_i's and D_j's are linearly dependent and so are T_i's and T_j's. This is made by taking $D_i = \epsilon_i D$, $\alpha_i = \alpha$ and $T_i = \delta_i T$, $\beta_i = \beta$, where (α, D) and (β, D) represents two phase type distributions. This assumption gives us the freedom to switch to the process (β_i, T_i) from (β_{i+1}, T_{i+1}), even before the absorption of the latter, whenever there is a state transition occurs in the chain $\mathcal{X}$. Such a model is introduced in the next section.

4 A Queueing-Inventory Model with Inventory Dependent Arrival and Service Processes

Consider a single server inventory model. The inventory is instantaneously replenished according to $(s - S)$ policy. At any time t, the arrival of customers is according to the inventory level at that time. When the inventory level is $s + i$, the distribution of the inter-arrival time is phase-type with representation $(\beta, \delta_i T)$ where T is a square matrix of order n and δ_i is a real number, $i = 1, 2, ..., r = S - s$. The service time distributions too are determined by the inventory level. While the inventory level is $s + i$, the service time distribution is phase type with representation $(\alpha, \epsilon_i D)$. Here D is of order $m \times m$, ϵ_i are real numbers $i = 1, 2, .., r$ and α is the initial distribution.

Let $N(t)$ be the number of customers in the system, $I(t)$ be the inventory level, $J_1(t)$ and $J_2(t)$ be the states of arrival and service processes respectively at time t. Then the system under discussion can be modelled by the continuous time Markov chain

$$\mathcal{N} = \{(N(t), I(t), J_1(t), J_2(t)) / t \geq 0\}$$

with state space

$$\{(0,i,j_1)/1 \le i \le r, 1 \le j_1 \le n\} \cup \{(k,i,j_1,j_2)/k = 1,2,3,...1 \le i \le r, 1 \le j_1 \le n, 1 \le j_2 \le m\}$$

The infinitesimal generator of the chain $\mathcal{N}$ is given by

$$Q = \begin{bmatrix} A_{00} & A_{01} & 0 & 0 & 0 & ... \\ A_{10} & A_1 & A_0 & 0 & 0 & ... \\ 0 & A_2 & A_1 & A_0 & 0 & ... \\ 0 & 0 & A_2 & A_1 & A_0 & ... \\ . & . & . & . & . & ... \\ . & . & . & . & . & ... \end{bmatrix}$$

where

$$\begin{aligned} A_{00} &= \Delta \otimes T \\ A_{01} &= \Delta \otimes T^0\beta \otimes \alpha \\ A_{10} &= E^{\perp} \otimes I_n \otimes D^0 \\ A_0 &= \Delta \otimes T^0\beta \otimes I_m \\ A_1 &= \Delta \otimes T \otimes I_m + E \otimes I_n \otimes D \\ A_2 &= E^{\perp} \otimes I_n \otimes D^0\alpha \end{aligned}$$

Here

$$\begin{aligned} \Delta &= \text{diag}(\delta_1, \delta_2, ..., \delta_r) \\ E &= \text{diag}(\epsilon_1, \epsilon_2, ..., \epsilon_r) \\ E^{\perp} &= \begin{bmatrix} 0 & 0 & 0 \, ... & 0 & \epsilon_1 \\ \epsilon_2 & 0 & 0 \, ... & 0 & 0 \\ 0 & \epsilon_2 & 0 \, ... & 0 & 0 \\ . & . & . \, ... & . & . \\ . & . & . \, ... & . & . \\ 0 & 0 & 0 \, ... & \epsilon_r & 0 \end{bmatrix} \end{aligned}$$

Now let π and θ be the stationary probability vectors of $T + T^0\beta$ and $D + D^0\alpha$ respectively and

$$A = A_0 + A_1 + A_2 = \Delta \otimes (T + T^0\beta) \otimes I_m + E \otimes I_n \otimes D + E^{\perp} \otimes I_n \otimes D^0\alpha.$$

For any row vector φ of length r such that $\varphi E^{\perp} = \varphi E$,

$$\begin{aligned} (\varphi \otimes \pi \otimes \theta) A &= \varphi E \otimes \pi \otimes \theta D + \varphi E^{\perp} \otimes \pi \otimes \theta D^0\alpha \\ &= \varphi E \otimes \pi \otimes \theta \left(D + D^0\alpha\right) \\ &= \mathbf{0}. \end{aligned}$$

Thus we have the following lemma.

Lemma 1. *For any row vector φ of length r such that $\varphi E^{\perp} = \varphi E$, $(\varphi \otimes \pi \otimes \theta)$ is a null vector of A.*

Choose $\phi = \left(\frac{1}{\epsilon_1}, \frac{1}{\epsilon_2}, ..., \frac{1}{\epsilon_r}\right)$. For this ϕ, $\phi E^{\perp} = \phi E$. Hence by lemma 1, $(\phi \otimes \pi \otimes \theta)$ is a null vector of A. Therefore $\Pi = \frac{1}{\phi . e}(\phi \otimes \pi \otimes \theta)$ is the stationary probability vector of A.

Now

$$\Pi A_2 e = \frac{r}{\phi e}\theta D^0 = \frac{r}{\phi e}\mu$$

and

$$\Pi A_0 e = \frac{1}{\phi e}\left(\sum_{i=1}^{r}\frac{\delta_i}{\epsilon_i}\right)\pi T^0 = \frac{1}{\phi e}\left(\sum_{i=1}^{r}\frac{\delta_i}{\epsilon_i}\right)\lambda$$

Hence we have the following theorem.

Theorem 1. *The continuous parameter irreducible Markov chain $\mathcal{N}$ is positive recurrent if and only if*

$$\left(\sum_{i=1}^{S-s}\frac{\delta_i}{\epsilon_i}\right)\lambda < (S-s)\mu$$

Note that when $\delta_i = \epsilon_i \ \forall i$, the condition for stability reduces to $\lambda < \mu$. In particular if $\delta_i = \epsilon_i = 1 \ \forall i$, inventory level and the arrival and service processes are independent.

5 Stationary Distribution of the Markov Chain $\mathcal{N}$

The stationary probability vector $\mathbf{z} = (\mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2 \ldots)$ is given by

$$\mathbf{z}_i = \mathbf{z}_1 R^{i-1}, i = 2, 3, 4... \tag{1}$$

$$\mathbf{z}_0 (\Delta \otimes T) + \mathbf{z}_1 \left(E^{\perp} \otimes I_n \otimes D^0\right) = 0 \tag{2}$$

$$\mathbf{z}_0 \left(\Delta \otimes T^0\beta \otimes \alpha\right) + \mathbf{z}_1 (\Delta \otimes T \otimes I_m + E \otimes I_n \otimes D) + \mathbf{z}_2 \left(E^{\perp} \otimes I_n \otimes D^0\alpha\right) = 0 \tag{3}$$

where the matrix R is the minimal solution of the matrix quadratic equation

$$R^2 \left(E^{\perp} \otimes I_n \otimes D^0\alpha\right) + R(\Delta \otimes T \otimes I_m + E \otimes I_n \otimes D) + \left(\Delta \otimes T^0\beta \otimes I_m\right) = 0$$

From Eq. (2),

$$\mathbf{z}_0 \left(\Delta \otimes T^0\right) = \mathbf{z}_1 \left(E^{\perp} \otimes e_n \otimes D^0\right)$$

So that

$$\begin{aligned}\mathbf{z}_0 \left(\Delta \otimes T^0\beta \otimes \alpha\right) &= \mathbf{z}_0 \left(\Delta \otimes T^0\right)(I_r \otimes \beta \otimes \alpha) \\ &= \mathbf{z}_1 \left(E^{\perp} \otimes e_n \otimes D^0\right)(I_r \otimes \beta \otimes \alpha) \\ &= \mathbf{z}_1 \left(E^{\perp} \otimes e_n\beta \otimes D^0\alpha\right)\end{aligned} \tag{4}$$

Using Eqs. (3) and (4), we get

$$\mathbf{z}_1\left[E^{\perp}\otimes e_n\beta\otimes D^0\alpha+\varDelta\otimes T\otimes I_m+E\otimes I_n\otimes D+R\left(E^{\perp}\otimes I_n\otimes D^0\alpha\right)\right]=0 \tag{5}$$

Therefore the vector z_1 can be uniquely determined up to a multiplicative constant. This constant can be found by normalizing the total probability to one.

We partitioned each steady state vector z_i as $z_i = (z_{i1}, z_{i2}, \dots, z_{ir})$ where $z_{0j} = (z_{0j1}, z_{0j2}, \dots, z_{0jn})$, $z_{ij} = (z_{ij11}, z_{ij12}, \dots, z_{ij1m}, \dots, z_{ijn1}, z_{ijn2} \dots, z_{ijnm})$, $j = 1, 2, \dots, r, i = 1, 2, 3, \dots$ in which z_{0jk} and $z_{ijkl}, k = 1, 2, \dots, n, l = 1, 2, \dots, m$ are scalars.

Some of the important system characteristics in the steady state are as follows.

1. Probability that the server is idle $= z_0 e_{rn}$.
2. For $k > 0$, Probability that there are k customers in the system,
$$P\left(N=k\right)=z_k e_{rnm}.$$
3. Expected number of customers in the system, $E\left(N\right)=\sum\limits_{i=1}^{n} i z_i e_{rnm}$.
4. Probability that the inventory level is j, $P\left(I=j\right)=z_{0j}e_n+\sum\limits_{i=1}^{\infty} z_{ij}e_{nm}$.
5. Expected inventory level, $E\left(I\right)=\sum\limits_{j=1}^{r} jP\left(I=j\right)$.

5.1 Expected Waiting Time

Consider a customer who joins the queue as the k^{th} customer. The waiting time W_k of this customer in the queue is the time until absorption of the Markov chain
$$W(t)=\{(r(t),I(t),J_s(t))/t\geq 0\}$$
where $r(t)$ is the position of the particular customer in the queue, $I(t)$ is the inventory level and $J_s(t)$ is the state of the ongoing service process at time t. The infinitesimal generator of this chain is

$$\tilde{Q}=\begin{bmatrix} Q_w & -Q_w e \\ 0 & 0 \end{bmatrix}$$

where

$$Q_w=\begin{bmatrix} E\otimes D & E^{\perp}\otimes D^0\alpha & 0 & 0 & 0 & \dots & 0 \\ 0 & E\otimes D & E^{\perp}\otimes D^0\alpha & 0 & 0 & \dots & 0 \\ 0 & 0 & E\otimes D & E^{\perp}\otimes D^0\alpha & 0 & \dots & 0 \\ 0 & 0 & 0 & E\otimes D & E^{\perp}\otimes D^0\alpha & \dots & 0 \\ . & . & . & . & . & & \dots \\ 0 & 0 & . & . & . & & E\otimes D \end{bmatrix}$$

Hence W_k follows a Phase type distribution with representation (ψ_w, Q_w) where $\psi_w = (\psi_k, \mathbf{0}, \mathbf{0}, ..\mathbf{0})$ in which ψ_k is a vector of length rm. The ij^{th} entry of ψ_k is the conditional probability that the chain is in a state with inventory level i and service phase j given that it is in level k in the steady state.

Hence, when the system is in the steady state, the expected waiting time of this customer is

$$\begin{aligned} W_k &= -\psi_w Q_w^{-1} e \\ &= \psi_k \left[\sum_{i=0}^{k} (-1)^i \left[(E \otimes D)^{-1} (E^{\perp} \otimes D^0 \alpha) \right]^i \right] \left[E \otimes D \right]^{-1} e \\ &= \psi_k \left[I + \left(\sum_{i=1}^{k} (I^{\perp})^k \right) \otimes e\alpha \right] \left[E \otimes D \right]^{-1} e. \\ &\text{where } I^{\perp} = E^{-1} E^{\perp}. \end{aligned}$$

Hence the Expected waiting time of an arbitrary customer in the steady state is given by

$$E(W) = \sum_{k=1}^{\infty} P(N = k) W_k$$

6 A Sample Problem on Cost Optimization and Numerical Analysis of the Chain $\mathcal{N}$

In this section we present an example of a cost optimization problem that arises in queueing-inventory situations with interdependent arrival and service processes. In this example we assumed that the demand increases with the inventory level according to the relation $\rho i^{1-\kappa}$, where ρ and $0 < \kappa < 1$ are constants and i is the inventory level. Baker et al. [1] used such a functional to model a situation with inventory level dependent demand which is for a relatively short season. Also we take the service rate to be proportional to the inventory level with σ as the proportionality constant.

6.1 A Cost Optimization Problem

We assumed that the multipliers ϵ_i and $\delta_i, i = 1, 2, \ldots r$ are related to the inventory level i by the relation $\delta_i = \rho i^{1-\kappa}$ and $\epsilon_i = \sigma i$ where ρ and σ are two positive parameters.

This illustration encompasses four distinct cost categories. The initial type involves the cost associated with providing the service at inventory level i, represented as $c(i)$. The second, denoted as c_0, refers to the cost of maintaining the server in a state of readiness during idle periods. Additionally, there are holding costs, denoted as c_h, incurred to ensure customer comfort within the system, along with the cost c_s associated with preserving the integrity of the inventory. All these costs are calculated per unit time.

Taking all these costs into account, we construct the cost function, $Cost = c_h \times E(N) + c_s \times E(I) + \sum_1^n c(i)P(I=i) + c_0 \times P\,(N=0)$ where $E(N)$ is the expected number of customers in the system, $E(I)$ is the expected inventory level, $P(I=i)$ is the probability that the inventory level is i and $P\,(N=0)$ is the probability that the system is idle.

For illustration, we choose $\rho = 0.7$, $\kappa = 0.4$, $c_h = 2$, $c_s = 0.5$, $c(i) = E(i,i)$, $c_0 = 6$. $T = \begin{bmatrix} -12 & 4 & 6 \\ 3 & -10 & 5 \\ 4 & 3 & -9 \end{bmatrix}$, $D = \begin{bmatrix} -7 & 1 & 2 & 1 \\ 3 & -11 & 2 & 3 \\ 2 & 2 & -10 & 3 \\ 5 & 3 & 4 & -15 \end{bmatrix}$, $\beta = (0.4, 0.35, 0.25)$ and $\alpha = (0.2, 0.3, 0.4, 0.1)$.

Our objective is to determine the rate at which the service rate should increase with the inventory level in order to minimize the incurred cost. That is we would like to find the value of the proportionality constant σ that optimizes the cost function. Figure 1 depicts the cost function plotted against σ. The convex nature of the curve indicates the presence of a minimum cost. Specifically, the cost reaches its minimum value when σ equals 0.54, with the minimal cost being 15.5280. Consequently, by exerting control over the service process, we ensure system stability even in the presence of heightened arrival rates, and we achieve this at the lowest possible cost.

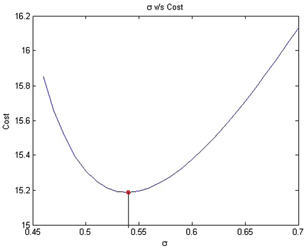


Fig. 1. Variation of cost wrt κ

Our numerical analysis demonstrates that it is possible to adjust the service rate based on the inventory level, thereby enabling control over the system's characteristics.

6.2 Numerical Analysis of the Chain $\mathcal{N}$

For the specified parameter values, we computed the expected number of customers $E(N)$, system idle probability $P(N=0)$, expected inventory level $E(I)$ expected waiting time $E(W)$ for various values of σ and the results are displayed in Figs. 2, 3 4 and 5. The calculated values are given in Table 2 in Appendix.

As σ increases, the service rates corresponding to each inventory level also rise. The escalation in service rate provides support for the growth in the arrival rate, stemming from increased inventory levels. This leads to a decrease in the expected number of customers in the system. With higher σ values, the traffic intensity decreases. Consequently, the system idle probability increases. For a small value of σ, service is delivered at a reduced rate when the inventory level is low. Consequently, it takes more time to complete a service and subsequently replenish the low inventory. As a result, the system experiences prolonged periods with low inventory levels. In contrast, higher values of σ increase the probability that the system is in a state with a higher inventory level.

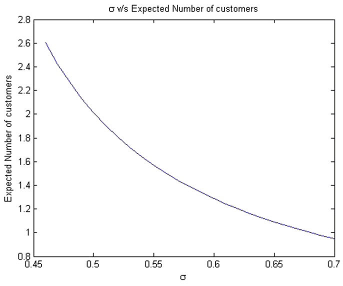


Fig. 2. Expected number of customers

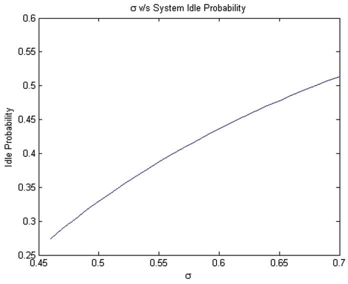


Fig. 3. Idle probability

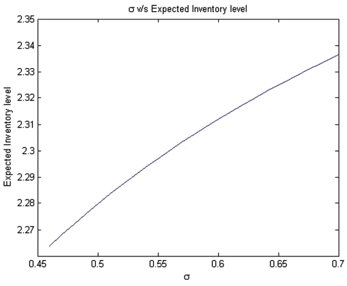


Fig. 4. Inventory level

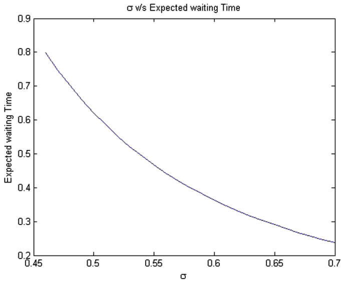


Fig. 5. Expected Waiting time

6.3 A Comparison Between Proposed Model and One with Stock Dependent Arrival Process and Independent Service Process

The model under discussion (Model 1) is compared with a similar one where the service rate remains unaffected by the inventory level (Model 2) giving the following values to the parameters. $\varDelta = diag(0.7000, 1.0610, 1.3532, 1.6082, 1.8386)$, and

$$D = \begin{bmatrix} -12 & 1 & 2 & 2 \\ 3 & -16 & 2 & 4 \\ 2 & 3 & -15 & 3 \\ 6 & 3 & 4 & -20 \end{bmatrix}, T = \begin{bmatrix} -10-d & 4 & 2 \\ 3 & -8-d & 1 \\ 1 & 3 & -8-d \end{bmatrix}$$ where d varies from 0 to 2 in increments of 0.1, $r = 5, n = 3, m = 4, \alpha = (.2, .3, .4, .1), \beta = (0.4, 0.35, 0.25), E = diag(0.4567, 0.9133, 1.3700, 1.8267, 2.2833)$

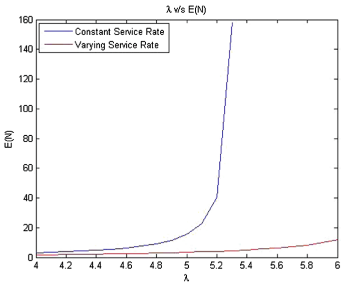


Fig. 6. Copmarison of two models

In Model 2, the service rate remains the same, while the arrival rate escalates with the inventory level. A comparison of the expected number of customer in both systems is presented in Fig. 6. The numerical results are tabulated in Table 1 in the Appendix. Notably, Model 2 experiences a higher inflow of customers compared to Model 1. In both systems, the arrival rate increases with the inventory level. When the inventory level is high, the arrival rate is also elevated. In the case, where the service rate is constant, the inventory level has nearly a uniform change, resulting in approximately equal probabilities for all inventory levels. This, in turn, leads to a high average arrival rate, causing the system to quickly burst out.

Conversely, in Model 1, the service rate diminishes as the inventory level decreases. Consequently, when the inventory level is low, service occurs at a slow pace. Since an increase in inventory level through replenishment occurs only after these long service periods, the system spends a considerable proportion of time in a state of low inventory. Consequently, at these times the expected arrival rate will be low. Therefore, even with higher demand v, the effective arrival rate will be moderate, ensuring the stability of the system.

In scenarios where there is high demand for the inventory, opting for stock-dependent service processes becomes advantageous in regulating the inflow of customers and maintaining system stability. This proves particularly useful in contexts such as ration distribution systems or the distribution of essential commodities to a large population. In such situations, the arrivals increase with the available stock. Consequently, we can exert control over the arrival rate by managing the inventory level. Our numerical study indicates that an effective method to achieve this control is by employing stock-dependent service processes.

7 Conclusion

In our investigation, we delved into a queueing inventory model incorporating interdependent arrival and service processes, along with the inventory level, utilizing two semi-Markov processes. The marginal distributions of both service time and inter-arrival times were identified as mixtures of phase-type distributions. Further analysis focused on a specific instance of this model, with the derivation of conditions for system stability. The stationary distribution was determined numerically. We also explored the distribution of waiting time, its expected value, and other critical system performance measures. This model has the potential for extension to more complex scenarios involving positive lead time. Examining the impact of replenishment time on system performance would be particularly intriguing in such cases.

Appendix

Results of numerical analysis mentioned in Sects. 6.1, 6.2 and 6.3 are tabulated in the following tables.

Table 1. Comparison between the models in terms of expected number of customers

Arrival Rate $-\alpha T^{-1}e$	Expected Number of Customers	
	Model 1	Model 2
4	1.6968	3.0114
4.1	1.8088	3.3473
4.2	1.9302	3.7428
4.3	2.0624	4.2155
4.4	2.2069	4.7899
4.5	2.3656	5.5027
4.6	2.5405	6.4105
4.7	2.7345	7.6053
4.8	2.9508	9.2483
4.9	3.1935	11.6489
5	3.4679	15.4862
5.1	3.7807	22.598
5.2	4.1406	40.2836
5.3	4.5592	157.9844
5.4	5.0521	
5.5	5.6413	
5.6	6.358	
5.7	7.249	
5.8	8.3869	
5.9	9.8911	
6	11.9727	

Table 2. Variation of important performance measures with σ

σ	Inventory Level	$E(N)$	$P(I)$	Cost	Expected Service rate	E(W)
0.46	2.2639	2.6019	0.2744	15.8498	3.0219	0.798
0.47	2.2681	2.4235	0.289	15.6555	3.0876	0.7485
0.48	2.2721	2.2683	0.3031	15.5059	3.1533	0.7026
0.49	2.276	2.1319	0.3165	15.3925	3.219	0.6602
0.5	2.2798	2.0111	0.3295	15.309	3.2847	0.6211
0.51	2.2835	1.9034	0.342	15.2502	3.3504	0.585
0.52	2.287	1.8068	0.354	15.2123	3.4161	0.5518
0.53	2.2905	1.7195	0.3656	15.192	3.4818	0.5211
0.54	2.2938	1.6404	0.3768	15.1868	3.5474	0.4929
0.55	2.2971	1.5683	0.3876	15.1945	3.6131	0.4668
0.56	2.3002	1.5024	0.398	15.2135	3.6788	0.4427
0.57	2.3033	1.4418	0.4081	15.2423	3.7445	0.4204
0.58	2.3062	1.3859	0.4178	15.2796	3.8102	0.3997
0.59	2.3091	1.3342	0.4272	15.3244	3.8759	0.3805
0.6	2.3119	1.2863	0.4363	15.3759	3.9416	0.3627
0.61	2.3147	1.2417	0.4451	15.4333	4.0073	0.3462
0.62	2.3173	1.2001	0.4537	15.496	4.073	0.3307
0.63	2.3199	1.1613	0.462	15.5634	4.1387	0.3163
0.64	2.3225	1.1248	0.47	15.635	4.2044	0.3028
0.65	2.3249	1.0907	0.4778	15.7105	4.2701	0.2902
0.66	2.3273	1.0585	0.4854	15.7893	4.3358	0.2784
0.67	2.3297	1.0282	0.4928	15.8713	4.4015	0.2673
0.68	2.332	0.9996	0.4999	15.9562	4.4672	0.2569
0.69	2.3342	0.9725	0.5069	16.0436	4.5328	0.2471
0.7	2.3364	0.9469	0.5136	16.1333	1.0873	0.2378

References

1. Baker, R.C., Urban, T.L.: Single-period inventory dependent demand models. Omega **16**(6), 605–607 (1988)
2. Bekker, R., Borst, S., Boxma, O.J., Kella, O.: Queues with workload-dependent arrival and service rates: In honor of vladimir kalashnikov (guest editors: Evsey morozov and richard serfozo). Queue. Syst. **46**, 03 (2004)
3. Conway, R.W., Maxwell, W.L.: A queuing model with state dependent service rates. J. Ind. Eng. **12**(2), 132–136 (1962)
4. Gupta, S.K.: Queues with hyper-poisson input and exponential service time distribution with state dependent arrival and service rates. Oper. Res. **15**(5), 847–856 (1967)

5. Hadley, G., Whitin, T.M.: Analysis of inventory systems. International Series in Management. Prentice-Hall (1963)
6. Hillier, F.S., Conway, R.W., Maxwell, W.L.: A multiple server queueing model with state dependent service rate. J. Ind. Eng. **15**(3), 153–157 (1964)
7. Johnson, E.L.: On (s, S) policies. Manage. Sci. **15**(1), 80–101 (1968)
8. Krishnamoorthy, A., Lakshmy, B., Rangaswamy, M.: A survey on inventory models with positive service time. OPSEARCH **48**, 153–169 (2011)
9. Krishnamoorthy, A., Shajin, D., Narayanan, V.: Inventory with positive service time: a survey. Adv. Trends Queue. Theory **2**, 201–238 (2021)
10. Larson, P.D., DeMarais, R.A.: Psychic stock: retail inventory for stimulating demand. In: Dunlap, B.J. (ed.) Proceedings of the 1990 Academy of Marketing Science (AMS) Annual Conference. DMSPAMS, pp. 447–450. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-13254-9_89
11. Latouche, G.: A phase-type semi-Markov point process. SIAM J. Algeb. Discrete Meth. **3**(1), 77–90 (1982)
12. Kaplan, E.H., Berman, O., Shevishak, D.G.: Deterministic approximations for inventory management at service facilities. IIE Trans. **25**(5), 98–104 (1993)
13. Ranjith, K.R., Krishnamoorthy, A., Gopakumar, B., Nair, S.: Analysis of a single server queue with interdependence of arrival and service processes - a semi-Markov approach, pp. 417–429 (2021)
14. Sigman, K., Simchi-Levi, D.: Light traffic heuristic for an $M/G/1$ queue with limited inventory. Ann. Oper. Res. **40**, 371–380 (1993)
15. Urban, T.L.: Inventory models with inventory-level-dependent demand: a comprehensive review and unifying theory. Eur. J. Oper. Res. **162**(3), 792–804 (2005)
16. Wolfe, H.B.: A model for control of style merchandise. Ind. Manag. Rev. **9**, 69–82 (1968)

Infinite-Server Queueing System with Two States of Service and Abandonments

Radmir Salimzyanov$^{(\boxtimes)}$ and Alexander Moiseev

Tomsk State University, Tomsk, Russian Federation
rsalimzyanov@yahoo.com, moiseev.tsu@gmail.com

Abstract. Infinite-server queueing system with Poisson arrival process, two states of service and abandonments is considered in the paper. Such system can be used as a simple mathematical model of a subscriber communication network based on IAB (Integrated Access and Backhaul) technology with two mobile nodes. Joint probability distribution of the number of customers in the states of service is obtained under asymptotic condition of high intensity of the arrival process. Numerical experiments are performed to estimate precision and applicability area of the approximation built on the results of the asymptotic analysis.

Keywords: IAB · asymptotic analysis · infinite-server queue · service abandonments

1 Introduction

Integrated Access and Backhaul (IAB) is a technology that provides fast and cost-effective deployment on millimeter waves (mmWave) due to self-connection in the same spectrum [1]. Wireless autonomous reverse transmission uses the same wireless channel to cover and connect to other base stations (BS), which leads to increased productivity, more efficient use of spectrum resources, and lower equipment costs, as well as to reducing dependence on the availability of wired reverse transmission at each location of the access node [2].

Mathematical modeling of an IAB-based network using queueing theory is a promising research direction. In addition to the mentioned standard [3], there have been studies conducted by various authors on the coverage of BS [4], signal transmission speeds under different conditions, and the utilization of fifth-generation networks with IAB on the Internet of Things [5]. However, the question of modeling of such systems still remains open.

In this paper, we propose a mathematical model of IAB-based network with two mobile nodes in the form of infinite-server queueing system with two states of service and abandonments. This model takes into account the roaming of a user from one node to another during the entire service time and the possibility of early leaving the system.

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 134–145, 2024.
https://doi.org/10.1007/978-3-031-65385-8_10

In paper [6], the authors considered a similar model in which there are two phases of service, but due to the specifics of the real problem, the phases of service are considered as sequential and each of them has its own service parameter (there is no a separate total time of successful service). Some results and literature reviews on models with abandonments of service (customers impatient in service) can be found in [7,8].

The rest part of the paper is organized as follows. In Sect. 2, the problem is formulated and a mathematical model in the form of a queueing system is proposed. In Sect. 3, the system of Kolmogorov equations is formulated and its exact solution obtained under the condition of equivalence of the local and global balance equations is provided. In Sect. 4, the asymptotic analysis method is applied for solution of the problem for a wider class of systems than the exact solution may be used for. As a result, an approximation of the joint probability distribution of the number of customers in the states of service is obtained. For estimating precision of the approximation and its applicability area, series of numerical experiments have been conducted. Their results are presented in Sect. 5. Conclusions are formulated in Sect. 6.

2 Problem and Mathematical Model

Making necessary assumptions and generalizations, we can depict the behavior of the entire system as follows. Let us consider an IAB system consisting of one donor and two mobile network customer service nodes (Fig. 1). Users move between two communication nodes and the following options are possible in the system:

- abandonment of service – user goes beyond the range of his or her communication node and does not connect to any other node (Fig. 1: a, b);
- internal migration – user goes beyond the range of his or her communication node, but immediately after that, he (or she) enters into the range of another communication node and can continue servicing (Fig. 1: a, b, and c);
- successful service completion – user completes his or her work and logs out of the system.

For this model, we are interested in how much the system is loaded, e.g. how many users are connected to each node, taking into account their possible migrations.

For modeling the system described above, we propose a mathematical model in the form of an infinte-server queue with two states of servicing (Fig. 2). The input flow is a Poisson arrival process with intensity λ. An incoming customer occupies any available server and starts its service in state 1 or 2 with probabilities v_1 or v_2, respectively. Duration of the service is an exponentially distributed random variable with parameter μ. While the customer is servicing, during time period of length Δt, it can move from state i to state k $(i, k \in \{1, 2\})$ with probability $\alpha_{ik}\Delta t$ (internal migration) or leave the system without completing

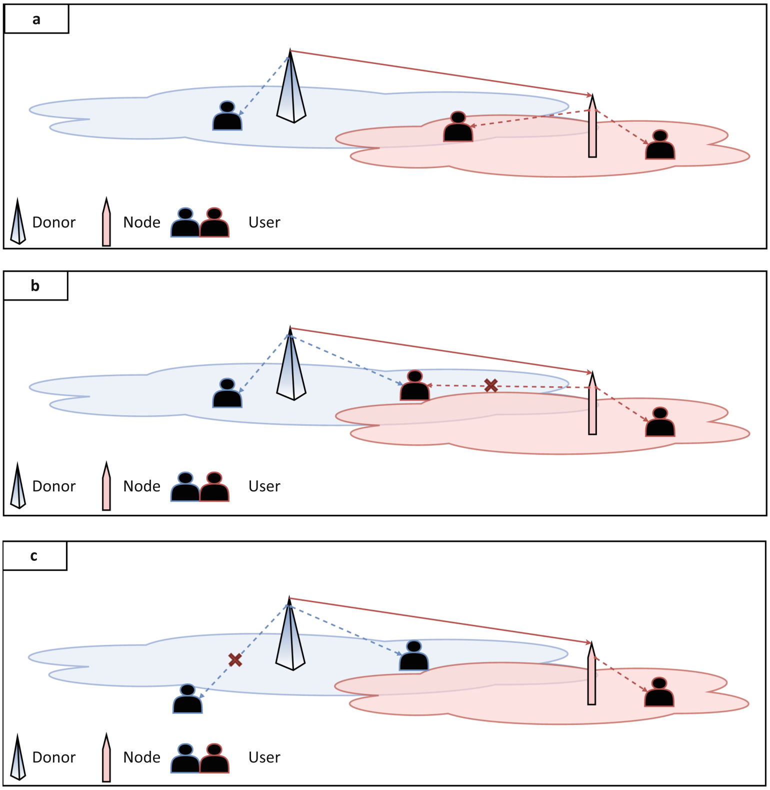


Fig. 1. System model

its service with probability $\alpha_{i0}\Delta t$ (abandonment of service). At the end of the service (successful service completion), the customer also leaves the system.

Let us denote the number of customers serviced in state i at instant t by $n_i(t)$ $(i = 1, 2)$. The problem is to find joint probability distribution of the number of customers in the states

$$P(n_1, n_2) = \Pr\{n_1(t) = n_1, n_2(t) = n_2\}$$

which we consider in a steady-state regime.

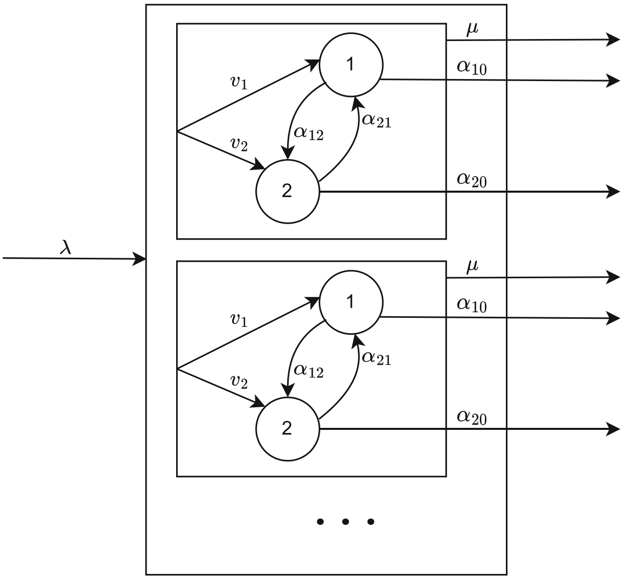


Fig. 2. Mathematical model

3 Kolmogorov Equations and Exact Solution

Described problem was considered in our recent paper [9]. The system of Kolmogorov equations for distribution $P(n_1, n_2)$ may be written as follows:

$$\begin{gathered}
P(n_1,n_2)[\lambda + n_1\mu + n_2\mu + n_1\alpha_{10} + n_2\alpha_{20} + n_1\alpha_{12} + n_2\alpha_{21}] = \\
P(n_1+1,n_2)(n_1+1)[\mu+\alpha_{10}] + P(n_1,n_2+1)(n_2+1)[\mu+\alpha_{20}] \\
+P(n_1-1,n_2)v_1\lambda + P(n_1,n_2-1)v_2\lambda \\
+P(n_1-1,n_2+1)(n_2+1)\alpha_{21} + P(n_1+1,n_2-1)(n_1+1)\alpha_{12}.
\end{gathered} \tag{1}$$

In [9], the following exact solution of the system was obtained:

$$\begin{gathered}
P(n_1,n_2) = \frac{1}{n_1!n_2!}\left[\frac{v_1\lambda}{\mu+\alpha_{10}}\right]^{n_1}\left[\frac{v_2\lambda}{\mu+\alpha_{20}}\right]^{n_2} P(0,0), \\
P(0,0) = \left(\sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}\frac{1}{n_1!n_2!}\left[\frac{v_1\lambda}{\mu+\alpha_{10}}\right]^{n_1}\left[\frac{v_2\lambda}{\mu+\alpha_{20}}\right]^{n_2}\right)^{-1}
\end{gathered} \tag{2}$$

under condition of the equivalence of the local and global balance equations which has the following form for the considered model:

$$\alpha_{21}[\mu+\alpha_{10}]v_2 = \alpha_{12}[\mu+\alpha_{20}]v_1. \tag{3}$$

Solution (2) can be applied only when condition (3) is satisfied and can not be used in other cases [10].

4 Asymptotic Analysis

Condition (3) imposes severe constraints, and solution (2) is almost inapplicable in practice. So, it is necessary to find a solution of system (1) for a wider range of model parameters. Because direct solution of the problem seems unreachable, we propose to use the asymptotic analysis method [11,12] for obtaining the solution.

Let us introduce the characteristic function

$$H(u_1, u_2) = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} e^{ju_1 n_1} e^{ju_2 n_2} P(n_1, n_2) \tag{4}$$

(here $j = \sqrt{-1}$) and make corresponding transformations in (1). We obtain

$$\begin{gathered} H(u_1, u_2)(v_1 \lambda (e^{ju_1} - 1) + v_2 \lambda (e^{ju_2} - 1)) + \\ + j \frac{\partial H(u_1, u_2)}{\partial u_1} (\mu + \alpha_{10} + \alpha_{12} - e^{-ju_1}(\mu + \alpha_{10} - e^{ju_2} \alpha_{12})) + \\ + j \frac{\partial H(u_1, u_2)}{\partial u_2} (\mu + \alpha_{20} + \alpha_{21} - e^{-ju_2}(\mu + \alpha_{20} - e^{ju_1} \alpha_{21})) = 0. \end{gathered} \tag{5}$$

We look for the solution of (5) under the condition of increasing intensity of the incoming flow: $\lambda \to \infty$.

4.1 First-Order Asymptotic

As the solution will be sought under the condition of increasing intensity of the incoming flow, we introduce the following notation:

$$\varepsilon = \frac{1}{\lambda},$$

where $\varepsilon \to 0$ while $\lambda \to \infty$. Also, we introduce the following notations:

$$u_1 = \varepsilon w_1, \qquad u_2 = \varepsilon w_2, \qquad H(u_1, u_2) = F_1(w_1, w_2, \varepsilon).$$

Let us make these substitutions in Eq. (5):

$$\begin{gathered} F_1(w_1, w_2, \varepsilon) \frac{1}{\varepsilon} \{ v_1 (e^{j\varepsilon w_1} - 1) + v_2 (e^{j\varepsilon w_2} - 1) \} + \\ + j \frac{\partial F_1(w_1, w_2, \varepsilon)}{\partial w_1} \frac{1}{\varepsilon} \{ \mu + \alpha_{10} + \alpha_{12} - e^{-jw_1 \varepsilon} (\mu + \alpha_{10} + e^{j\varepsilon w_2} \alpha_{12}) \} + \\ + j \frac{\partial F_1(w_1, w_2, \varepsilon)}{\partial w_2} \frac{1}{\varepsilon} \{ \mu + \alpha_{20} + \alpha_{21} - e^{-jw_2 \varepsilon} (\mu + \alpha_{20} + e^{j\varepsilon w_1} \alpha_{21}) \} = 0. \end{gathered}$$

Using the expansion

$$e^{j\varepsilon w_k} = 1 + j\varepsilon w_k + +O(\varepsilon^2),$$

after completing limit transition $\varepsilon \to 0$, we obtain

$$\begin{gathered}
F_1(w_1,w_2)(jw_1v_1+jw_2v_2)+\\
+j^2\frac{\partial F_1(w_1,w_2)}{\partial w_1}(w_1\mu+w_1\alpha_{10}-w_2\alpha_{12}+w_1\alpha_{12})+\\
j^2\frac{\partial F_1(w_1,w_2)}{\partial w_2}(w_2\mu+w_2\alpha_{20}-w_1\alpha_{21}+w_2\alpha_{21})=0.
\end{gathered} \tag{6}$$

Let us rewrite Eq. (6) in the following form:

$$\begin{gathered}
F_1(w_1,w_2)jw_1v_1+j^2\frac{\partial F_1(w_1,w_2)}{\partial w_1}(w_1\mu+w_1\alpha_{10}+w_1\alpha_{12})+\\
j^2\frac{\partial F_1(w_1,w_2)}{\partial w_2}(-w_1\alpha_{21})=0,\\
F_1(w_1,w_2)jw_2v_2+j^2\frac{\partial F_1(w_1,w_2)}{\partial w_1}(-w_2\alpha_{12})+\\
j^2\frac{\partial F_1(w_1,w_2)}{\partial w_2}(w_2\mu+w_2\alpha_{20}+w_2\alpha_{21})=0.
\end{gathered}$$

Dividing the first equation by $F(w_1,w_2)w_1$ and the second one by $F(w_1,w_2)w_2$, we obtain

$$\begin{gathered}
jv_1+j^2\frac{\partial F_1(w_1,w_2)}{\partial w_1}\frac{1}{F_1(w_1,w_2)}(\mu+\alpha_{10}+\alpha_{12})+\\
j^2\frac{\partial F_1(w_1,w_2)}{\partial w_2}\frac{1}{F_1(w_1,w_2)}(-\alpha_{21})=0,\\
jv_2+j^2\frac{\partial F_1(w_1,w_2)}{\partial w_1}\frac{1}{F_1(w_1,w_2)}(-\alpha_{12})+\\
j^2\frac{\partial F_1(w_1,w_2)}{\partial w_2}\frac{1}{F_1(w_1,w_2)}(\mu+\alpha_{20}+\alpha_{21})=0.
\end{gathered} \tag{7}$$

We will look a solution of this system in the following form:

$$F_1(w_1,w_2)=\exp\{jw_1a_1+jw_2a_2\},$$

where a_1 and a_2 are some constants. Making corresponding substitutions in (7), we derive

$$\begin{gathered}
jw_1(v_1-a_1\mu-a_1\alpha_{10}-\alpha_{12}a_1+a_2\alpha_{21})+\\
jw_2(v_2-a_2\mu-a_1\alpha_{10}-\alpha_{12}a_1+a_2\alpha_{21})=0,
\end{gathered}$$

which we write in the form

$$\begin{cases}
v_1-a_1(\mu+\alpha_{10}+\alpha_{12})+a_2\alpha_{21}=0,\\
v_2-a_2(\mu+\alpha_{20}+\alpha_{21})+a_1\alpha_{12}=0.
\end{cases}$$

Solving this system, we obtain

$$a_1 = \frac{v_1 + v_2\alpha_{21}}{(\mu+\alpha_{20}+\alpha_{21})(\mu+\alpha_{10}+\alpha_{12})+\alpha_{12}\alpha_{21}},$$
$$a_2 = \frac{\alpha_{12}v_1 + v_2(\mu+\alpha_{10}+\alpha_{12})}{(\mu+\alpha_{10}+\alpha_{12})(\mu+\alpha_{20}+\alpha_{21})+\alpha_{12}\alpha_{21}}.$$

4.2 Second-Order Asymptotic

Let us perform the following substitution in Eq. (5):

$$H(u_1,u_2) = H^{(2)}(u_1,u_2)\exp\{ju_1\lambda a_1 + ju_2\lambda a_2\}, \tag{8}$$

where $H^{(2)}(u_1,u_2)$ is the characteristic function of two-dimensional centered random process $\{n_1(t)-a_1\lambda, n_2(t)-a_2\lambda\}$. We obtain

$$\begin{gathered}H(u_1,u_2)\{-\lambda + j^2(\mu+\alpha_{10}+\alpha_{12})\lambda a_1 + j^2(\mu+\alpha_{20}+\alpha_{21})\\ -j^2e^{-ju_1}(\mu+\alpha_{10})\lambda a_1 - j^2e^{-ju_1}(\mu+\alpha_{20})\lambda a_2\\ +v_1\lambda e^{ju_1} + v_2\lambda e^{ju_2} - j^2e^{-ju_1}\alpha_{12}e^{ju_2}\lambda a_1 - j^2e^{-ju_2}\alpha_{21}e^{ju_1}\lambda a_2\}\\ +\frac{\partial H^{(2)}(u_1,u_2)}{\partial u_1}\left\{j(\mu+\alpha_{10}+\alpha_{12}) - je^{-ju_1}(\mu+\alpha_{10}) - je^{-ju_1}\alpha_{12}e^{ju_2}\right\}\\ +\frac{\partial H^{(2)}(u_1,u_2)}{\partial u_2}\left\{j(\mu+\alpha_{20}+\alpha_{21}) - je^{-ju_2}(\mu+\alpha_{20}) - je^{-ju_2}\alpha_{21}e^{ju_1}\right\} = 0.\end{gathered}$$

By making the following substitutions:

$$\lambda = \frac{1}{\varepsilon^2}, \quad u_1 = \varepsilon w_1, \quad u_2 = \varepsilon w_2,$$
$$H^{(2)}(u_1,u_2) = F_2(w_1,w_2,\varepsilon),$$

we derive

$$\begin{gathered}F_2(w_1,w_2,\varepsilon)\frac{1}{\varepsilon^2}\{-1 + j^2(\mu+\alpha_{10}+\alpha_{12})a_1 + j^2(\mu+\alpha_{20}+\alpha_{21})\\ -j^2e^{-j\varepsilon w_1}(\mu+\alpha_{10})a_1 - j^2e^{-j\varepsilon w_1}(\mu+\alpha_{20})a_2\\ +v_1e^{j\varepsilon w_1} + v_2e^{j\varepsilon w_2} - j^2e^{-j\varepsilon w_1}\alpha_{12}e^{j\varepsilon w_2}a_1 - j^2e^{-j\varepsilon w_2}\alpha_{21}e^{j\varepsilon w_1}a_2\}\\ +\frac{\partial F_2(w_1,w_2,\varepsilon)}{\partial w_1}\frac{1}{\varepsilon}\left\{j(\mu+\alpha_{10}+\alpha_{12}) - je^{-j\varepsilon w_1}(\mu+\alpha_{10}) - je^{-j\varepsilon w_1}\alpha_{12}e^{j\varepsilon w_2}\right\}\\ +\frac{\partial F_2(w_1,w_2,\varepsilon)}{\partial w_2}\frac{1}{\varepsilon}\left\{j(\mu+\alpha_{20}+\alpha_{21}) - je^{-j\varepsilon w_2}(\mu+\alpha_{20}) - je^{-j\varepsilon w_2}\alpha_{21}e^{j\varepsilon w_1}\right\} = 0.\end{gathered}$$

Using expansions

$$e^{j\varepsilon w_k} = 1 + j\varepsilon w_k + \frac{(j\varepsilon w_k)^2}{2} + O(\varepsilon^2)$$

and

$$e^{-j\varepsilon w_k} = 1 - j\varepsilon w_k + \frac{(j\varepsilon w_k)^2}{2} + O(\varepsilon^2),$$

we obtain

$$\begin{gathered}
-\frac{1}{2}F_2(w_1,w_2)\Big\{w_1^2(v_1 + a_1(\mu + \alpha_{10} + \alpha_{12}) + \alpha_{21}a_2) \\
+w_2^2(v_1 + a_2(\mu + \alpha_{20} + \alpha_{21}) + \alpha_{12}a_1) - 2w_1w_2(\alpha_{12}a_1 + \alpha_{21}a_2)\Big\} \\
-\frac{\partial F_2(w_1,w_2)}{\partial w_1}jw_1\Big\{\alpha_{12} - \alpha_{10} - \mu\Big\} + \frac{\partial F_2(w_1,w_2)}{\partial w_1}jw_2\alpha_{12} \\
-\frac{\partial F_2(w_1,w_2)}{\partial w_2}jw_1\Big\{\alpha_{21} - \alpha_{20} - \mu\Big\} + \frac{\partial F_2(w_1,w_2)}{\partial w_2}jw_2\alpha_{21} = 0.
\end{gathered} \tag{9}$$

We will look for a solution of this equation in the form

$$F(w_1,w_2) = \exp\{-\frac{1}{2}w_1^2K_{11} - \frac{1}{2}w_2^2K_{22} - w_1w_2K_{12}\}, \tag{10}$$

where K_{11}, K_{22}, and K_{12} are some constants.

Substituting (10) into (9), we obtain:

$$\begin{gathered}
w_1^2\Big\{-\frac{1}{2}(v_1 + a_1(\mu + \alpha_{10} + \alpha_{12}) + \alpha_{21}a_2) - 2K_{11}(\alpha_{12} - \alpha_{10} - \mu) + K_{12}\alpha_{21}\Big\}+ \\
+w_2^2\Big\{-\frac{1}{2}(v_1 + a_2(\mu + \alpha_{20} + \alpha_{21}) + \alpha_{12}a_1) + K_{12}\alpha_{12} + 2K_{22}\alpha_{21}\Big\}+ \\
+w_1w_2\Big\{\alpha_{12}a_1 + \alpha_{21}a_2 - K_{12}(\alpha_{12} - \alpha_{10} - \mu) + 2K_{11}\alpha_{12}- \\
-2K_{22}(\alpha_{21} - \alpha_{20} - \mu) + K_{12}\alpha_{21}\Big\} = 0.
\end{gathered}$$

After some derivations, we obtain the following expressions for evaluation of constants K_{11}, K_{22}, K_{12}:

$$K_{11} = 2\frac{v_1 + a_2\alpha_{21} + K_{12}\alpha_{21}}{\alpha_{12} + \alpha_{10} + \mu},$$

$$K_{22} = 2\frac{v_2 + a_1\alpha_{12} + K_{12}\alpha_{12}}{\alpha_{21} + \alpha_{20} + \mu},$$

$$K_{12} = \frac{\alpha_{12}\dfrac{v_1 + \alpha_{21}a_2}{\alpha_{12} + \alpha_{10} + \mu} + \alpha_{21}\dfrac{v_2 + \alpha_{12}a_1}{\alpha_{21} + \alpha_{20} + \mu} - (a_1\alpha_{12} + a_2\alpha_{21})}{\alpha_{12} + \alpha_{10} + \mu + \alpha_{21} + \alpha_{21} + \mu - \dfrac{\alpha_{12}\alpha_{21}}{\alpha_{12} + \alpha_{10} + \mu} - \dfrac{\alpha_{12}\alpha_{21}}{\alpha_{21} + \alpha_{20} + \mu}}.$$

4.3 Approximation of Joint Probability Distribution of the Number of Customers in States of Service

Taking into account derived expressions for constants $a_1, a_2, K_{11}, K_{22}, K_{12}$ and using expression (8), we obtain the following approximation for characteristic function of the number of customers in states of service in the steady-state regime:

$$H(u_1, u_2) \approx \exp\left\{ju_1\lambda a_1 + ju_2\lambda a_2 - u_1u_2\lambda K_{12} - \frac{u_1^2\lambda K_{11}}{2} - \frac{u_2^2\lambda K_{22}}{2}\right\}, \quad (11)$$

which can be applied for enough big values of the arrival process intensity λ. So, the probability distribution of the number of customers in the states of service in the steady-state regime $P(n_1, n_2)$ is a two-dimensional Gaussian distribution with vector of mathematical expectations

$$\mathbf{m} = \lambda \left[a_1 \quad a_2\right] \quad (12)$$

and covariance matrix

$$\mathbf{K} = \lambda \begin{bmatrix} K_{11} & K_{12} \\ K_{12} & K_{22} \end{bmatrix}. \quad (13)$$

Because (11) represents characteristic function of continuous random variable with possible negative values, we need in constructing of probability distribution for integer non-negative values which can be applied as an approximation for the probability distribution of the number of customers. To do this, we propose to use the following cumulative distribution function (c.d.f.):

$$F(i, k) = \frac{G(i + 0.5, k + 0.5) - G(i - 0.5, k - 0.5)}{1 - G(-0.5, -0.5)}, \quad (14)$$

where $i, k \in \{0, 1, \dots\}$ mean the number of customers in service states 1 and 2 respectively, $G(i, k)$ is a c.d.f. of two-dimensional Gaussian distribution with vector of mathematical expectations (12) and covariance matrix (13).

5 Numerical Example

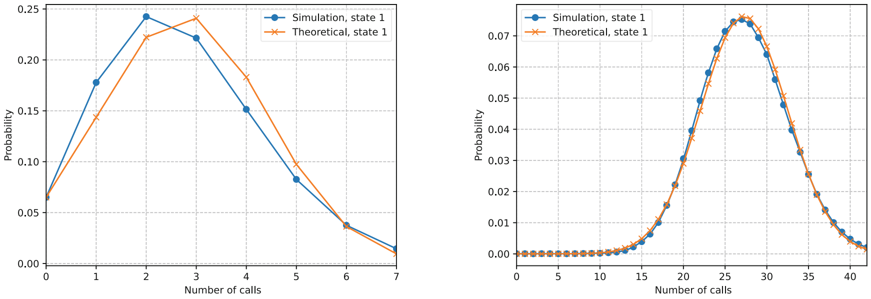
To evaluate the accuracy of approximation (14), we conduct the following experiment: for different values of parameter λ, using simulations, we obtain an empirical probability distribution function and compare it with approximation (14). For the comparison, we will take into account only marginal distributions for the corresponding states of service. For accuracy estimation, we use the Kolmogorov distance

$$\Delta = \max_i |F(i) - F_{\text{sim}}(i)|,$$

where $F_{\text{sim}}(i)$ is an empirical c.d.f. built on the base of results of simulations, and $F(i)$ is a marginal one-dimensional Gaussian c.d.f. built on the base of expression (14). Choosen values of system parameters are given in Table 1. Also, we preformed similar numerical comparison of the approximation with the exact solution obtained under the condition of equivalence of the local and global balance equations (3). The experiments were conducted with the same values of infinite (Table 1), for which condition (3) is satisfied.

Table 1. Values of parameters for numerical experiments

Parameter	Value
μ	0.1
v_1	0.3
v_2	0.7
α_{12}	12.83
α_{21}	1
α_{10}	1
α_{20}	0.1



a) $\lambda = 10$ b) $\lambda = 100$

Fig. 3. Probability distribution of the number of customers in state 1 for $\lambda = 10$ and $\lambda = 100$

Figure 3 shows a comparison of the stationary probability distributions of the number of customers serviced in state 1 for different intensities of the arrival process. Table 2 shows corresponding values of the Kolmogorov distance. We can see that accuracy of the theoretical approximation increases with increasing of λ. The same results for state 2 can be found in Table 3. We consider the results can be an acceptable if Kolmogorov distance $\Delta \leq 0.05$ (highlighted in boldface in the tables). So, as we see from the tables, we reach the acceptable results for the obtained approximation for values $\lambda \geq 10$.

Table 2. Kolmogorov distance between probability distributions of the number of customers in state 1 for various values of λ: Δ_{sim} – approximation against simulation; Δ_{ex} – approximation against exact solution

λ	1	5	10	15	20
Δ_{sim}	0,1388	0.0712	**0,0354**	**0.0227**	**0.0165**
Δ_{ex}	0,1387	0.0711	**0,0348**	**0.0229**	**0.0170**

Table 3. Kolmogorov distance between probability distributions of the number of customers in state 2 for various values of λ: Δ_{sim} – approximation against simulation; Δ_{ex} – approximation against exact solution

λ	1	5	10	15	20
Δ_{sim}	**0,0288**	**0.0058**	**0,0031**	**0.0016**	**0.0016**
Δ_{ex}	**0,0284**	**0.0056**	**0,0027**	**0,0018**	**0.0013**

6 Conclusion

Mathematical model for subscriber communication network using IAB technology with two mobile nodes is proposed in the paper. The model is formulated in the form of an infinite-server queueing system with two states of service and abandonments. The method of asymptotic analysis is applied to find the joint two-dimensional probability distribution of the number of customers in the first and second states of service. Obtained result in the form of an approximation can be applied in the case when the condition of equivalence of the local and global balance equations is not met but it is limited by enough big intensity of the arrival process. Conducted numerical experiments approve applicability of the obtained approximation. We think that the approach may be applied for models with an arbitrary number of service states and for models with non-Poisson arrivals and non-exponential service times.

References

1. Sadovaya, Ye., et al.: Integrated access and backhaul in millimeter-wave cellular: benefits and challenges. IEEE Commun. Mag. **60**(9), 81–86 (2022)
2. Polese, M., Giordani, M., Roy, A., Goyal, S., Castor, D., Zorzi, M.: End-to-end simulation of integrated access and backhaul at mmwaves. In: 2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), pp. 1–7. IEEE (2018)
3. 3GPP, TS 38.174, NR; Integrated Access and Backhaul Radio Transmission and Reception, V.16.1.0 (2020). https://www.3gpp.org/ftp/Specs/html-info/38174.htm
4. Łukowa, A., Venkatasubramanian, V., Visotsky, E., Cudak, M.: On the coverage extension of 5g millimeter wave deployments using integrated access and backhaul. In: 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 1–7. IEEE (2020)
5. Monteiro, V.F., et al.: Paving the way toward mobile IAB: problems, solutions and challenges. IEEE Open J. Commun. Society **3**, 2347–2379 (2022)
6. Fedorova, E., Lapatin, I., Lizyura, O., Moiseev, A., Nazarov, A., Paul, S.: Queueing system with two phases of service and service rate degradation. Axioms **12**(2), 104 (2023)
7. Garnett, O., Mandelbaum, A., Reiman, M.: Designing a call center with impatient customers. Manufact. Serv. Oper. Manage. **4**(3), 208–227 (2002)

8. Dai, J.G., He, S.: Many-server queues with customer abandonment: numerical analysis of their diffusion model. Stoch. Syst. **3**(1), 96–146 (2013)
9. Salimzyanov R., Moiseev A.: Local balance equation for the probability distribution of the number of customers in IAB network. In: Systemy Upravleniya, informatsionnye technologii i matematicheskoe modelirovanie (SUITMM), V Russian conference, Omsk 2023, 284–289. Omsk State University, Russia, Omsk (2023) (in Russian)
10. Nazarov, A., Yuzhakov, A.: Equivalence criterion of the equations of global and detailed balance for Markov chains. Autom. Remote. Control. **56**(12), 1718–1724 (1995)
11. Moiseev, A., Nazarov, A.: Asymptotic analysis of the infinite-server queueing system with high-rate semi-Markov arrivals. In: 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, ICUMT 2014, 6-8 Oct., St. Petersburg, pp. 507–513. IEEE (2015)
12. Nazarov, A., Moiseev, A.: Analysis of the $GI/PH/\infty$ system with high-rate arrivals. Autom. Control. Comput. Sci. **49**(6), 328–339 (2015)

Cost Analysis in a Production Inventory System: Managing Varied Production Rates, Service Interruption and Customer Retrial

Susmi Skaria[1], Salini S. Nair[2], and K. P. Jose[1,2](✉)

[1] Department of Mathematics, Union Christian College, Aluva 683102, Kerala, India
kpjspc@gmail.com
[2] PG & Research Department of Mathematics, St. Peter's College, Kolenchery 682311, Kerala, India

Abstract. This paper introduces the notion of service interruption within a production inventory system. Customer arrivals follow a Poisson process and service times are exponentially distributed. The system allows retrials for service, and production rates vary suitably to accommodate increased demand. If an item is out of stock, the server is occupied or an interruption occurs, a primary customer enters the orbit of infinite capacity; otherwise, the customer lost permanently. Retrial attempts for service can be made from the orbit. The system's stability is assessed and key performance measures are defined. A relevant cost function is formulated and subjected to numerical and graphical analysis.

Keywords: Retrial inventory · Different production rates · Service interruption · Cost Analysis

1 Introduction

Service interruption models include different types of service unavailability. This may be due to server breakdown, server interruptions, server taking vacations, arrival of a priority customer,unreliable server, etc. There are numerous studies on inventory systems where interruption occurs due to unreliable supplier. Our study is due to the unreliable sever.

The first study on an inventory system was by Berman et al. [1]. He studied a deterministic model in which a processing time is required for serving the inventory. Rashid et al. [9] analyzed a production inventory system by considering demand and production time as stochastic parameters. They calculated the

The authors acknowledge the financial support provided by the FIST Program, Dept. of Science and Technology, Govt. of India, to the PG & Research Dept. of Mathematics through SR/FST/College-2018- XA 276(C).
The authors would like to express their sincere gratitude to the referees for their insightful comments and valuable suggestions that greatly improved the quality of this paper.

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 146–160, 2024.
https://doi.org/10.1007/978-3-031-65385-8_11

transition probabilities in steady state and the long-run inventory costs. Krishnamoorthy et al. [5] studied queues with service interruption and repair. This paper presents an infinite-capacity queueing system with a single server where the service rule is FIFO. Long-run system distribution is also obtained under a stable regime. Krishnamoorthy and Jose [4] analysed and compared three production inventory systems with positive service time and retrial of customers and they found that the model with a buffer size equal to the inventoried items is the best profitable model.

Server interruptions and retrials in an inventory model were studied by Krishnamoorthy et al. [6]. They calculated the waiting time of a customer in the orbit and their dependence on different system parameters. A production inventory system with different rates of production and retrials was studied by Jose and Salini [3]. They employed the Matrix analytic method to find an algorithmic solution. They also compared two production inventory systems with the retrial of customers and varying production rates by introducing a buffer with different capacities [2] and found that the model with varying buffer-size is more efficient for practical applications.

A solution for an Inventory model with server interruption and retrials was investigated by E. Sandhya et al. [13]. They found an explicit expression for the steady state distribution and several performance measures are evaluated explicitly and numerically. Salini and Jose [11] studied the production inventory system with retrial and varying service rates by assuming the arrival of customers as MAP and service time following Phase Type distribution.

Rejitha K.R. and K.P.Jose [10] studied a queueing inventory system with MAP, retrials, and different replenishment rates. Here the arrival of customers follows a Markovian arrival process and service time follows Phase-Type distribution. They used the Matrix analytic method to analyze the model. In the paper, A PH distributed production inventory model with different modes of service and MAP arrivals, Salini and Jose [12] studied a production inventory model with the retrial of customers under (s, S) policy. Here also arrival pattern follows MAP. The production process follows Phase-Type distribution. They analyzed the effect of correlation between two successive inter-arrival times. In the present study, the concept of interruption is introduced to a production inventory system with varying production rates.

2 Mathematical Modelling and Analysis of the Problem

We consider a production inventory system under (s, S) policy. Retrial of customers and service interruptions are also allowed. The item in the inventory is served through a single server counter. The production process starts whenever the inventory falls to s and continues the production till the inventory reaches level S. The service of a customer can be interrupted at any time. Interruptions occur during the service and there is no restriction on the number of possible interruptions. An arriving customer who finds that the server is busy, server is interrupted or inventory level is zero can proceed to an orbit where they can retry for the service. The following assumptions are made for modeling this problem.

- The arrival of customers is according to the Poisson process with rate λ and the service pattern follows an exponential distribution with rate μ.
- The production rate is $\alpha\beta$ where $\alpha \in [1, c]$ and c is a finite number greater than 1, when production starts and the rate falls to β when the inventory level crosses above s.
- The inter occurrence-time of interruption is exponentially distributed with parameter δ_1 and an exponentially distributed amount of time with parameter δ_2 is required to resume service from where it is stopped.
- The inter retrial-time is exponentially distributed with linear rate $i\theta$, when there are i customers in the orbit.
- An arriving customer who finds the server busy, on interruption, or inventory level zero can proceed to an orbit with probability γ and lost forever with probability $1 - \gamma$.
- A retrial customer who finds the server busy, on interruption or inventory level zero returns to the orbit with probability δ and lost forever with probability $1 - \delta$.
- Inventory as well as customers are not lost due to server interruptions.

The following are the notations used in this model.
N(t): Number of customers in the orbit at time t.
I(t): Inventory level at time t.

J(t):The production status $J(t) : \begin{cases} 0, \text{ if the production is on OFF mode,} \\ 1, \text{ if the production is on ON mode.} \end{cases}$

S(t) :The server status $S(t) : \begin{cases} 0, \text{ if the server is idle,} \\ 1, \text{ if the server is busy,} \\ 2, \text{ if the server is on interruption.} \end{cases}$

Now $X(t) = \{(N(t), S(t), J(t), I(t)) : t \geq 0.\}$ is a level dependent quasi birth death process on the state space $\{(i, j, 0, k) : i \geq 0; j = 0, 1, 2; k = s + 1, ..., S\}$ $\bigcup\{(i, 0, 1, k) : i \geq 0; k = 0, ..., S - 1\}$ $\bigcup\{(i, j, 1, k) : i \geq 0; j = 1, 2; k = 1, ..., S - 1\}$.

The infinitesimal generator Q of the process is a block tri-diagonal matrix and has the form:

$$Q = \begin{bmatrix} A_{1,0} & A_0 & & & \\ A_{2,1} & A_{1,1} & A_0 & & \\ & A_{2,2} & A_{1,2} & A_0 & \\ & & A_{2,3} & A_{1,3} & A_0 \\ & & & \ddots & \ddots & \ddots \end{bmatrix}$$

where the block matrices $A_0, A_{1,i}; (i \geq 0)$ and $A_{2.i}; (i \geq 1.)$ are square matrices of order $6S - 3s - 2$ and are given by

$$A_0 = \begin{array}{c} \underline{0,0} \\ \underline{0,1} \\ \underline{1,0} \\ \underline{1,1} \\ \underline{2,0} \\ \underline{2,1} \end{array} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & D_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & D_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & D_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & D_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & D_3 \end{bmatrix}$$

$$A_{2,i} = \begin{array}{c} \underline{0,0} \\ \underline{0,1} \\ \underline{1,0} \\ \underline{1,1} \\ \underline{2,0} \\ \underline{2,1} \end{array} \begin{bmatrix} 0 & 0 & B_1 & 0 & 0 & 0 \\ 0 & B_2 & 0 & B_3 & 0 & 0 \\ 0 & 0 & B_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & B_5 & 0 & 0 \\ 0 & 0 & 0 & 0 & B_4 & 0 \\ 0 & 0 & 0 & 0 & 0 & B_5 \end{bmatrix} , \qquad A_{1,i} = \begin{array}{c} \underline{0,0} \\ \underline{0,1} \\ \underline{1,0} \\ \underline{1,1} \\ \underline{2,0} \\ \underline{2,1} \end{array} \begin{bmatrix} G_1 & 0 & G_2 & 0 & 0 & 0 \\ C_3 & C_4 & 0 & G_3 & 0 & 0 \\ C_5 & C_6 & G_4 & 0 & G_5 & 0 \\ 0 & C_7 & C_8 & C_9 & 0 & G_6 \\ 0 & 0 & G_7 & 0 & G_8 & 0 \\ 0 & 0 & 0 & G_9 & 0 & G_{10} \end{bmatrix}$$

where $\underline{a,b}$ denotes the entry corresponding to the variations of the inventory level j for fixed i, the number of customers in the orbit and a and b stands for the server status and production status respectively. The $(m,n)^{th}$ element of the matrices contained in $A_0, A_{2,i}$, and $A_{1,i}$ are given below.

$$D_1 = (\lambda\gamma)C_1, \qquad D_2 = (\lambda\gamma)I_{S-s},$$

$$D_3 = (\lambda\gamma)I_{S-1}, \qquad (C_1)_{mn} = \begin{cases} 1; \text{m=n=1} \\ 0; \text{otherwise} \end{cases} ,$$

$$B_1(i) = (i\theta)I_{S-s}, \qquad B_2(i) = (i\theta)(1-\delta)C_1, \qquad B_3(i) = (i\theta)C_2,$$
$$B_4(i) = (i\theta)(1-\delta)I_{S-s}, \qquad B_5(i) = (i\theta)(1-\delta)I_{S-1},$$

$$(C_2)_{mn} = \begin{cases} 1;\ \text{m=2,...S, n=m-1} \\ 0;\ \text{otherwise} \end{cases} , \quad (C_3)_{mn} = \begin{cases} \beta;\ \text{m=S, n=S-s} \\ 0;\ \text{otherwise} \end{cases} ,$$

$$(C_4)_{mn}(i) = \begin{cases} -(\lambda\gamma + \alpha\beta + i\theta(1-\delta));\ \text{m=n=1} \\ -(\lambda + \alpha\beta + i\theta);\ \text{m=2,...s, n=m} \\ -(\lambda + \beta + i\theta);\ \text{m=s+1,...S, n=m} \\ \alpha\beta;\ \text{m=1,...s, n=m+1} \\ \beta;\ \text{m=s+1,...S-1, n=m+1} \\ 0;\ \text{otherwise} \end{cases} ,$$

$$(C_5)_{mn} = \begin{cases} \mu;\ \text{m=2,...S-s, n=m-1} \\ 0;\ \text{otherwise} \end{cases} , \qquad (C_6)_{mn} = \begin{cases} \mu;\ \text{m=1, n=s+1} \\ 0;\ \text{otherwise} \end{cases} ,$$

$$(C_7)_{mn} = \begin{cases} \mu;\ \text{m=1,...S-1, n=m} \\ 0; \text{otherwise} \end{cases} \qquad (C_8)_{mn} = \begin{cases} \beta;\ \text{m=S-1, n=S-s} \\ 0;\ \text{otherwise} \end{cases} ,$$

$$(C_9)_{mn}(i) = \begin{cases} -(\lambda\gamma + \alpha\beta + \mu + i\theta(1-\delta) + \delta_1); \text{ m=1,...s-1, n=m} \\ -(\lambda\gamma + \beta + \mu + i\theta(1-\delta) + \delta_1); \text{ m=s,... S-1, n=m} \\ \alpha\beta; \text{ m=1,...s-1, n=m+1} \\ \beta; \text{ m=s,...S-2,n=m+1} \\ 0; \text{ otherwise} \end{cases},$$

$$\begin{aligned} &G_1(i) = -(\lambda + i\theta)I_{S-s}, \qquad G_2 = \lambda I_{S-s}, \\ &G_3 = \lambda C_2, \qquad G_4(i) = -(\lambda\gamma + \mu + i\theta(1-\delta) + \delta_1)I_{S-s}, \\ &G_5 = \delta_1 I_{S-s}, \qquad G_6 = \delta_1 I_{S-1}, \\ &G_7 = \delta_2 I_{S-s}, \qquad G_8(i) = -(\lambda\gamma + i\theta(1-\delta) + \delta_2)I_{S-s}, \\ &G_9 = \delta_2 I_{S-1}, \qquad G_{10}(i) = -(\lambda\gamma + i\theta(1-\delta) + \delta_2)I_{S-1}. \end{aligned}$$

Using the Neuts-Rao [9] truncation method we can find an N such that $A_{1,i} = A_1$ and $A_{2,i} = A_2$ when $i \geq N$.Then the infinitesimal generator Q of the process will take the following form,

$$Q = \begin{bmatrix} A_{1,0} & A_0 & & & & & & & \\ A_{2,1} & A_{1,1} & A_0 & & & & & & \\ & A_{2,2} & A_{1,2} & A_0 & & & & & \\ & & A_{2,3} & A_{1,3} & A_0 & & & & \\ & & & \ddots & \ddots & \ddots & & & \\ & & & & A_{2,N-1} & A_{1,N-1} & A_0 & & \\ & & & & & A_2 & A_1 & A_0 & \\ & & & & & & A_2 & A_1 & A_0 \\ & & & & & & & \ddots & \ddots & \ddots \end{bmatrix}.$$

3 Steady State Analysis

3.1 System Stability

Let $A = A_0 + A_1 + A_2$. Then

$$A = \begin{bmatrix} M_1 & 0 & M_2 & 0 & 0 & 0 \\ M_3 & M_4 & 0 & M_5 & 0 & 0 \\ M_6 & M_7 & M_8 & 0 & M_9 & 0 \\ 0 & M_{10} & M_{11} & M_{12} & 0 & M_{13} \\ 0 & 0 & M_{14} & 0 & M_{15} & 0 \\ 0 & 0 & 0 & M_{16} & 0 & M_{17} \end{bmatrix}$$

where $M_1 = G_1, M_2 = B_1 + G_2, M_3 = C_3, M_4 = D_1 + B_2 + C_4, M_5 = B_3 + G_3, M_6 = C_5, M_7 = C_6, M_8 = D_2 + B_4 + G_4, M_9 = G_5, M_{10} = C_7, M_{11} =$

$C_8, M_{12} = B_5 + C_9 + D_3, M_{13} = G_6, M_{14} = G_7, M_{15} = D_2 + B_4 + G_8, M_{16} = G_9, M_{17} = D_3 + B_5 + G_{10}$.
Le the steady state probability vector of A be $\boldsymbol{\pi} = (\pi(0), \pi_1(0), \pi(1), \pi_1(1), \pi(2), \pi_1(2))$. Then the equation $\boldsymbol{\pi}A$=0 gives the following equations.

$$\begin{array}{l}
\pi(0)M_1 + \pi_1(0)M_3 + \pi(1)M_6 = 0.\\
\pi(0)M_2 + \pi(1)M_8 + \pi_1(1)M_{11} + \pi(2)M_{14} = 0.\\
\pi_1(0)M_5 + \pi_1(1)M_{12} + \pi_1(2)M_{16} = 0.\\
\pi(1)M_9 + \pi(2)M_{15} = 0.\\
\pi_1(1)M_{13} + \pi_1(2)M_{17} = 0.
\end{array}$$

From these equations it follows that

$$\begin{array}{l}
\pi(0) = (\pi_1(0)M_3 + \pi(1)M_6)(-M_1)^{-1}\\
\pi_1(0) = (\pi(1)M_7 + \pi_1(1)M_{10})(-M_4)^{-1}\\
\pi(1) = (\pi(0)M_2 + \pi_1(1)M_{11} + \pi(2)M_{14})(-M_8)^{-1}\\
\pi_1(1) = (\pi_1(0)M_5 + \pi_1(2)M_{16})(-M_{12})^{-1}\\
\pi(2) = (\pi(1)M_9(-M_{15})^{-1}\\
\pi_1(2) = (\pi_1(1)M_113)(-M_{17})^{-1}
\end{array}$$

where $M_1 = -(\lambda + N\theta)I_{S-s}$, $M_8 = -(\mu + \delta_1)I_{S-s}$, $M_{15} = (-\delta_2)I_{S-s}$,$M_{17} = (-\delta_2)I_{S-1}$. M_4 and M_{12} are upper triangular matrices given by $M_4 = (\lambda\gamma + N\theta(1-\delta))C_1 + C_4$, $M_{12} = (\lambda\gamma + N\theta(1-\delta))I_{S-1} + C_9$. All these matrices are invertible. So the above equations can be solved using Block Gauss Seidel iteration procedure to find the vector $\boldsymbol{\pi}$. The stability condition can be stated as $lim_{N\to\infty}\frac{\pi A_0 e}{\pi A_2 e} < 1$, which we checked numerically where
$\pi A_0 e$=$\pi_1(0) + (\pi(1) + \pi(2))(\lambda\gamma)I_{S-s} + (\pi_1(1) + \pi_1(2))(\lambda\gamma)I_{S-1}$ and
$\pi A_2 e$=$\pi(0)(N\theta)I_{S-s} + \pi_1(0)(N\theta)(1-\delta))C_1 + \pi_1(0)(N\theta)C_2 + (\pi(1) + \pi(2))(N\theta)(1-\delta)I_{S-s} + (\pi_1(1) + \pi_1(2))(N\theta)(1-\delta))I_{S-1}$.

3.2 Steady State Probability Vector

Let $\mathbf{x} = (x_0, x_1, ..., x_{N-1}, x_N ...)$ be the steady state probability vector of Q, where each x_i is given by

$$x_i = (y_{i,0,0,s+1} \cdots, y_{i,0,0,S}, y_{i,0,1,0} \cdots y_{i,0,1,S-1}, y_{i,1,0,s+1}, \cdots y_{i,1,0,S},$$
$$y_{i,1,1,1}, \cdots y_{i,1,1,S-1}, y_{i,2,0,s+1}, \cdots y_{i,2,0,S}, y_{i,2,1,1}, \cdots y_{i,2,1,S-1}).$$

Under the stability condition, x_i's are given by $x_{N+r-1} = x_{N-1}R^r (r \geq 1)$ where R is the unique non negative solution of the equation $R^2A_2 + RA_1 + A_0 = 0$ for which the spectral radius is less than one and the vectors $x_0, x_1, ...x_{N-1}$ are obtained by solving

$$\begin{array}{l}
x_0A_{1,0} + x_1A_{2,1} = 0\\
x_{i-1}A_0 + x_iA_{1,i} + x_{i+1}A_{2,i+1} = 0 (1 \leq i \leq N-2)\\
x_{N-2}A_0 + x_{N-1}(A_{1,N-1} + RA_2) = 0
\end{array}$$

subject to the normlizing condition

$$[\textstyle\sum_{i=0}^{N-2} x_i + x_{N-1}(1-R)^{-1}]e = 1.$$

3.3 Rate Matrix R and Truncation Level N

The rate matrix R is evaluated using an iterative method. We denote the sequence of R by $\{R(N)\}$and is defined by $R_0(N) = 0$ and
$R_{n+1}(N){=}{-}(R^2(N)A_2(N) - A_0(N))A_1^{-1}(N)$.
Elsner's algorithm [7] is used to find the spectral radius $\eta(N)$ in such a way that $\|\eta(N) - \eta(N + l)\| < \epsilon$ where ϵ is an arbitrarily small value and $\eta(N)$ is the spectral radius of $R(N)$.

3.4 System Performance Measures

Some important performance measures are

(i) Expected inventory Level in the system

$$EIL = \Sigma_{i=0}^{\infty}\Sigma_{j=0}^{2}\Sigma_{k=s+1}^{S} k y_{i,j,0,k} + \Sigma_{i=0}^{\infty}\Sigma_{j=0}^{2}\Sigma_{k=1}^{S-1} k y_{i,j,1,k}.$$

.

(ii) Expected Number of customers in the orbit

$$ENC = \left(\Sigma_{i=1}^{\infty} i x_i\right) e = \left(\Sigma_{i=1}^{N-1} i x_i + x_N(N(I - R)^{-1} + R(I - R)^{-2})\right) e.$$

.

(iii) Expected interruption rate

$$EIR = \delta_1 \Sigma_{i=0}^{\infty}\left(\Sigma_{k=s+1}^{S} y_{i,1,0,k} + \Sigma_{k=1}^{S-1} y_{i,1,1,k}\right).$$

.

(iv) Expected repair rate of the server

$$ERR = \delta_2 \Sigma_{i=0}^{\infty}\left(\Sigma_{k=s+1}^{S} y_{i,2,0,k} + \Sigma_{k=1}^{S-1} y_{i,2,1,k}\right).$$

.

(v) Expected number of external customers lost, before entering the orbit is

$$ECL_P = (1 - \gamma)\lambda\left(\Sigma_{i=0}^{\infty} y_{i,0,1,0} + \Sigma_{i=0}^{\infty}\Sigma_{j=1}^{2}\Sigma_{k=s+1}^{S} y_{i,j,0,k} + \Sigma_{i=0}^{\infty}\Sigma_{j=1}^{2}\Sigma_{k=1}^{S-1} y_{i,j,1,k}\right).$$

(vi) Expected number of departures after completing service is

$$END = \mu\left(\Sigma_{i=0}^{\infty}\Sigma_{k=s+1}^{S} y_{i,1,0,k} + \Sigma_{i=0}^{\infty}\Sigma_{k=1}^{S-1} y_{i,1,1,k}\right).$$

.

(vii) Expected number of customers lost due to retrials

$$ECL_R = \theta(1 - \delta)\left(\Sigma_{i=1}^{\infty} i\left(y_{i,0,1,0} + \Sigma_{j=1}^{2}\Sigma_{k=s+1}^{S} y_{i,j,0,k} + \Sigma_{j=1}^{2}\Sigma_{k=1}^{S-1} y_{i,j,1,k}\right)\right).$$

.

(viii) Overall rate of retrials

$$ORR = \theta\left(\Sigma_{i=1}^{\infty} i x_i\right) e.$$

.

(ix) Successful rate of retrials

$$SRR = \theta \Sigma_{i=0}^{\infty} \left(\Sigma_{k=s+1}^{S} y_{i,0,0,k} + \Sigma_{k=1}^{S-1} y_{i,0,1,k} \right).$$

.

(x) Expected switching rate

$$ESR = \mu \Sigma_{i=0}^{\infty} y_{i,1,0,s+1}.$$

.

(xi) Server busy probability

$$SRB = \Sigma_{i=0}^{\infty} \Sigma_{k=s+1}^{S} y_{i,1,0,k} + \Sigma_{i=0}^{\infty} \Sigma_{k=1}^{S-1} y_{i,1,1,k}.$$

.

3.5 Cost Analysis

The Expected Total Cost (ETC) is defined as

$$ETC = c_1 ESR + c_2 EIL + c_3 ENC + c_4 ECL_P + c_5 ECL_R + c_6 END + c_7 ERR$$

where

c_1 = switching cost for production
c_2 = holding cost of inventory /unit/unit time
c_3 = holding cost of the customers in the orbit/unit/unit time
c_4 = cost due to loss of primary customers/unit/unit time
c_5 = cost due to loss of retrial customers/unit/unit time
c_6 = cost due to service/unit/unit time
c_7 = repair cost for the server.

4 Numerical and Graphical Illustrations

This section aims to provide a detailed description of the numerical and graphical experiments that were conducted to analyse the effects of changes in various parameters on the performance measures and expected total cost.

- Table 1 displays the nature of expected interruption rate(EIR), overall rate of retrials(ORR), successful rate of retrials(SRR) and server busy probability(SRB) with an increase in the customer arrival rate λ. As the arrival rate increases, we expect an increase in the number of customers in the system. This will lead to an increase in the overall retrial rate, server busy probability, and successful retrial rate. Naturally, there is a chance to increase the service interruption rate. The table explains this well. Figure 1(a) shows the variations in the expected total cost with variations in the parameter λ, keeping other parameters fixed. From the graph, it is clear that the minimum expected total cost is attained when $\lambda = 1.2$.

- Table 2 analyses different performance measures and expected total cost by varying the service rate μ. As shown in the table, an increase in the service rate μ decreases the expected interruption rate, overall rate of retrials, and server busy probability. Customers get quick service, which leads to an increase in the successful rate of retrials. Graphical representation with variations in μ (see Fig. 1(b)) obtains the expected minimum cost of 36.3547 at $\mu = 2.5$.
- Arriving customers go to the orbit because of the server interruption, server busy or no inventory. So as the value of γ increases, there will be an increase in the server busy probability,successful rate of retrials, overall rate of retrials and interruption rate as in Table 3. In Fig. 1(c), the convexity of the graph is obtained, and the minimum expected total cost is attained at $\gamma = 0.4$.
- An increase in the production rate α brings more customers to the service station. This is illustrated in Table 4. This will increase the expected interruption rate, the successful rate of retrials, and the server busy probability. But the overall rate of retrials decreases. The expected minimum cost is obtained for $\alpha = 1.6$ as in Fig. 1(d).
- Table 5 shows that with the increase in the retrial rate θ, the overall rate of retrials and successful rate of retrials increases. This in turn decreases the server busy probability and expected interruption rate. Graphical representation is shown in Fig. 2(a).
- The variations in δ are similar to those of λ and γ as we can see from Table 6. As more customers are retrying for service, the overall rate of retrials increases to a great extent. Figure 2(b) shows the variations in ETC with respect to changes in the value of δ. The expected minimum cost 73.8573 is attained for $\delta = 0.6$.
- Table 7 and 8 analyze the nature of expected number of customers too, in addition to the other performance measures. As the interruption rate δ_1 increases, ENC also increases. This is because as the interruption becomes more frequent, the expected service time of a customer increases. This will reduce the server busy probability and successful rate of retrials(See Table 7). At the same time, an increase in the repair rate δ_2 brings an increase in these rates because the server becomes active in a shorter time after an interruption which in turn leads to an increase in the service completion rate(See Table 8). This will reduce the expected number of customers in the system. Graphical illustrations for variations in δ_1 and δ_2 to obtain the minimum ETC are shown in Fig. 2(c) and Fig. 2(d) respectively.

4.1 Optimization of (s, S) Pair

The optimum value of (s, S) pair is obtained by considering suitable parameter values. Here we calculated the optimum value of expected total cost by varying the maximum inventory level S and the inventory level s at which production starts. The optimum (s, S) pair which minimizes the expected total cost is found to be (7,27) and the corresponding ETC is 5936.7 (Table 9).

Table 1. Variations in λ.

λ	EIR	ORR	SRR	SRB
1.1	0.9685	3.8014	0.7181	0.3228
1.2	0.9918	4.1359	0.7322	0.3306
1.3	1.0136	4.4864	0.7466	0.3378
1.4	1.0338	4.8475	0.7607	0.3446
1.5	1.0524	5.2145	0.7744	0.3508
1.6	1.0694	5.5837	0.7873	0.3565
1.7	1.0849	5.9521	0.7994	0.3616
1.8	1.0990	6.3177	0.8107	0.3663
1.9	1.1118	6.6791	0.8210	0.3706

$S = 50, s = 5, \alpha = 1.5, \delta = 0.7, \theta = 2, \mu = 3, \gamma = 0.8, \beta = 1.5, \delta_1 = 3, \delta_2 = 2.5$

Table 2. Variations in μ.

μ	EIR	ORR	SRR	SRB
2.1	1.1435	5.6694	0.6008	0.3812
2.2	1.1332	5.6154	0.6223	0.3777
2.3	1.1229	5.5622	0.6432	0.3743
2.4	1.1126	5.5100	0.6635	0.3709
2.5	1.1025	5.4586	0.6833	0.3675
2.6	1.0923	5.4080	0.7026	0.3641
2.7	1.0822	5.3584	0.7213	0.3607
2.8	1.0722	5.3096	0.7395	0.3574
2.9	1.0623	5.2616	0.7572	0.3541

$S = 50, s = 5, \alpha = 1.5, \lambda = 1.5, \theta = 2, \gamma = 0.8, \beta = 1.5, \delta = 0.7, \delta_1 = 3, \delta_2 = 2.5$

Table 3. Variations in γ.

γ	EIR	ORR	SRR	SRB
0.1	0.8945	2.1390	0.5210	0.2982
0.2	0.9079	2.3103	0.5406	0.3026
0.3	0.9260	2.5578	0.5672	0.3087
0.4	0.9488	2.9145	0.6020	0.3163
0.5	0.9749	3.3842	0.6434	0.3250
0.6	1.0020	3.9451	0.6879	0.3340
0.7	1.0282	4.5650	0.7323	0.3427
0.8	1.0524	5.2145	0.7744	0.3508
0.9	1.0741	5.8718	0.8129	0.3580

$S = 50, s = 5, \alpha = 1.5, \lambda = 1.5, \theta = 2, \gamma = 0.8, \beta = 1.5, \delta = 0.7, \delta_1 = 3, \delta_2 = 2.5$

Table 4. Variations in α.

α	EIR	ORR	SRR	SRB
1.1	1.0479	5.2217	0.7715	0.3493
1.2	1.0496	5.2191	0.7726	0.3499
1.3	1.0510	5.2169	0.7736	0.3503
1.4	1.0521	5.2152	0.7743	0.3507
1.5	1.0530	5.2138	0.7753	0.3513
1.6	1.0538	5.2126	0.7753	0.3513
1.7	1.0543	5.2116	0.7756	0.3514
1.8	1.0548	5.2108	0.7759	0.3516
1.9	1.0552	5.2101	0.7761	0.3517

$S = 50, s = 5, \lambda = 1.5, \mu = 3, \gamma = 0.8, \beta = 1.5, \theta = 2, \delta = 0.7, \delta_1 = 3, \delta_2 = 2.5$

Table 5. Variations in θ.

θ	EIR	ORR	SRR	SRB
1.1	1.0602	4.5215	0.7575	0.3534
1.2	1.0601	4.6296	0.7607	0.3534
1.3	1.0597	4.7261	0.7633	0.3532
1.4	1.0589	4.8130	0.7654	0.3530
1.5	1.0580	4.8921	0.7671	0.3527
1.6	1.0569	4.9648	0.7687	0.3523
1.7	1.0558	5.0323	0.7701	0.3519
1.8	1.0546	5.0958	0.7715	0.3515
1.9	1.0535	5.1563	0.7729	0.3512
2	1.0524	5.2145	0.7744	0.3508

$S = 50, s = 5, \lambda = 1.5, \delta = 0.7, \alpha = 1.5, \mu = 3, \gamma = 0.8, \beta = 1.5, \delta_1 = 3, \delta_2 = 2.5$

Table 6. Variations in δ.

δ	EIR	ORR	SRR	SRB
0.1	0.9500	2.6857	0.5901	0.3167
0.2	0.9554	2.8086	0.6007	0.3185
0.3	0.9633	2.9800	0.6154	0.3211
0.4	0.9747	3.2284	0.6360	0.3249
0.5	0.9914	3.6045	0.6658	0.3305
0.6	1.0160	4.2031	0.7095	0.3387
0.7	1.0524	5.2145	0.7744	0.3508
0.8	1.1051	7.0801	0.8703	0.3684
0.9	1.1779	11.2774	1.0089	0.3926

$S = 50, s = 5, \lambda = 1.5, \alpha = 1.5, \theta = 2, \mu = 3, \gamma = 0.8, \beta = 1.5, \delta_1 = 3, \delta_2 = 2.5$

Table 7. Variations in δ_1.

δ_1	ENC	EIR	ORR	SRR	SRB
2.2	2.4876	0.8683	4.9753	0.8618	0.3943
2.4	2.5204	0.9185	5.0408	0.8382	0.3827
2.6	2.5511	0.9657	5.1023	0.8158	0.3714
2.8	2.5800	1.0103	5.1601	0.7946	0.3608
3	2.6072	1.0524	5.2145	0.7744	0.3508
3.2	2.6329	1.0923	5.2658	0.7552	0.3413
3.4	2.6572	1.1300	5.3144	0.7369	0.3324
3.6	2.6801	1.1658	5.3603	0.7195	0.3238
3.8	2.7019	1.1999	5.4038	0.7028	0.3158

$S = 50, s = 5, \alpha = 1.5, \lambda = 1.5, \theta = 2, \mu = 3, \gamma = 0.8, \beta = 1.5, \delta = 0.7, \delta_2 = 2.5$

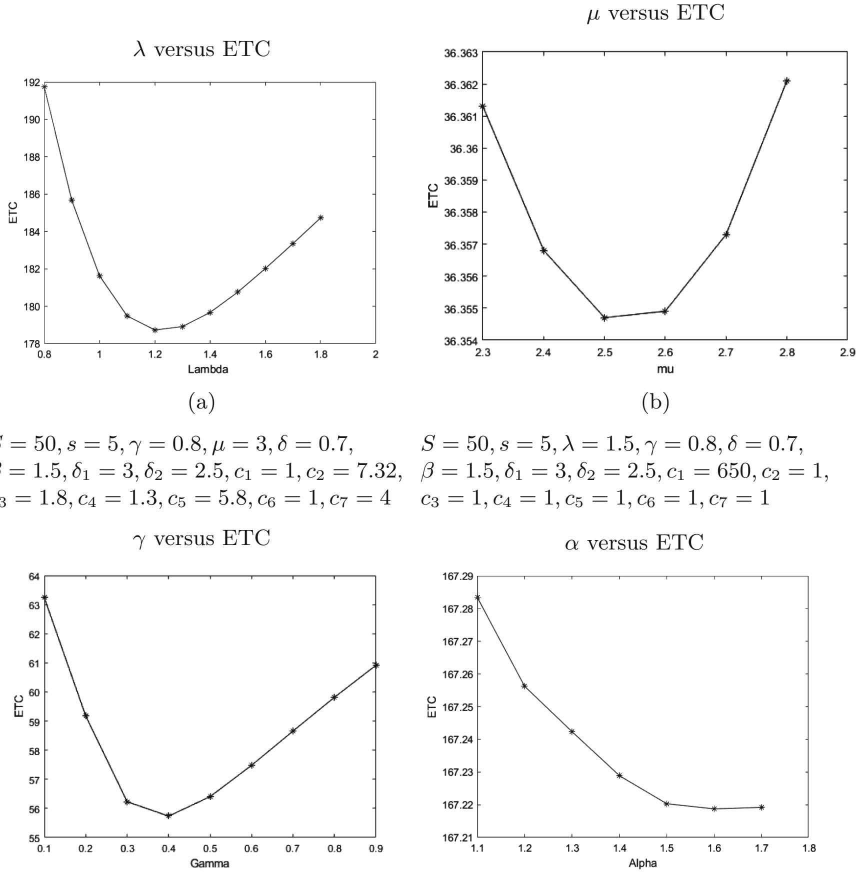
Table 8. Variations in δ_2.

δ_2	ENC	EIR	ORR	SRR	SRB
2.2	2.6606	0.9914	5.3212	0.7325	0.3305
2.4	2.6242	1.0330	5.2485	0.7611	0.3443
2.6	2.5910	1.0709	5.1820	0.7870	0.3570
2.8	2.5605	1.1015	5.1210	0.8104	0.3685
3	2.5324	1.1373	5.0649	0.8317	0.3791
3.2	2.5066	1.1665	5.0131	0.8512	0.3888
3.4	2.4826	1.1934	4.9652	0.8691	0.3978
3.6	2.4604	1.2183	4.9208	0.8855	0.4061
3.8	2.4398	1.2414	4.8796	0.9007	0.4138

$S = 50, s = 5, \alpha = 1.5, \lambda = 1.5, \theta = 2\mu = 3, \gamma = 0.8, \beta = 1.5, \delta = 0.7, \delta_1 = 3$

Table 9. $\lambda = 1.5, \mu = 3, \beta = 2, \delta = 0.7, \delta_1 = 2, \delta_2 = 3, \alpha = 1.5, c_1 = 1, c_2 = 1, c_3 = 2000, c_4 = 10, c_5 = 1, c_6 = 1, c_7 = 1$

s S	25	26	27	28	29	30
5	5938.0	5937.8	5937.7	5937.6	5937.6	5937.7
6	5937.1	5937.0	5936.9	5936.9	5936.9	5937.0
7	5937.0	5936.9	**5936.7**	5936.8	5936.8	5936.9
8	5937.4	5937.2	5937.1	5937.0	5937.1	5937.2
9	5938.0	5937.8	5937.6	5937.6	5937.6	5937.7
10	5938.7	5938.4	5938.3	5938.1	5938.1	5938.2



(a)

$S = 50, s = 5, \gamma = 0.8, \mu = 3, \delta = 0.7,$
$\beta = 1.5, \delta_1 = 3, \delta_2 = 2.5, c_1 = 1, c_2 = 7.32,$
$c_3 = 1.8, c_4 = 1.3, c_5 = 5.8, c_6 = 1, c_7 = 4$

(b)

$S = 50, s = 5, \lambda = 1.5, \gamma = 0.8, \delta = 0.7,$
$\beta = 1.5, \delta_1 = 3, \delta_2 = 2.5, c_1 = 650, c_2 = 1,$
$c_3 = 1, c_4 = 1, c_5 = 1, c_6 = 1, c_7 = 1$

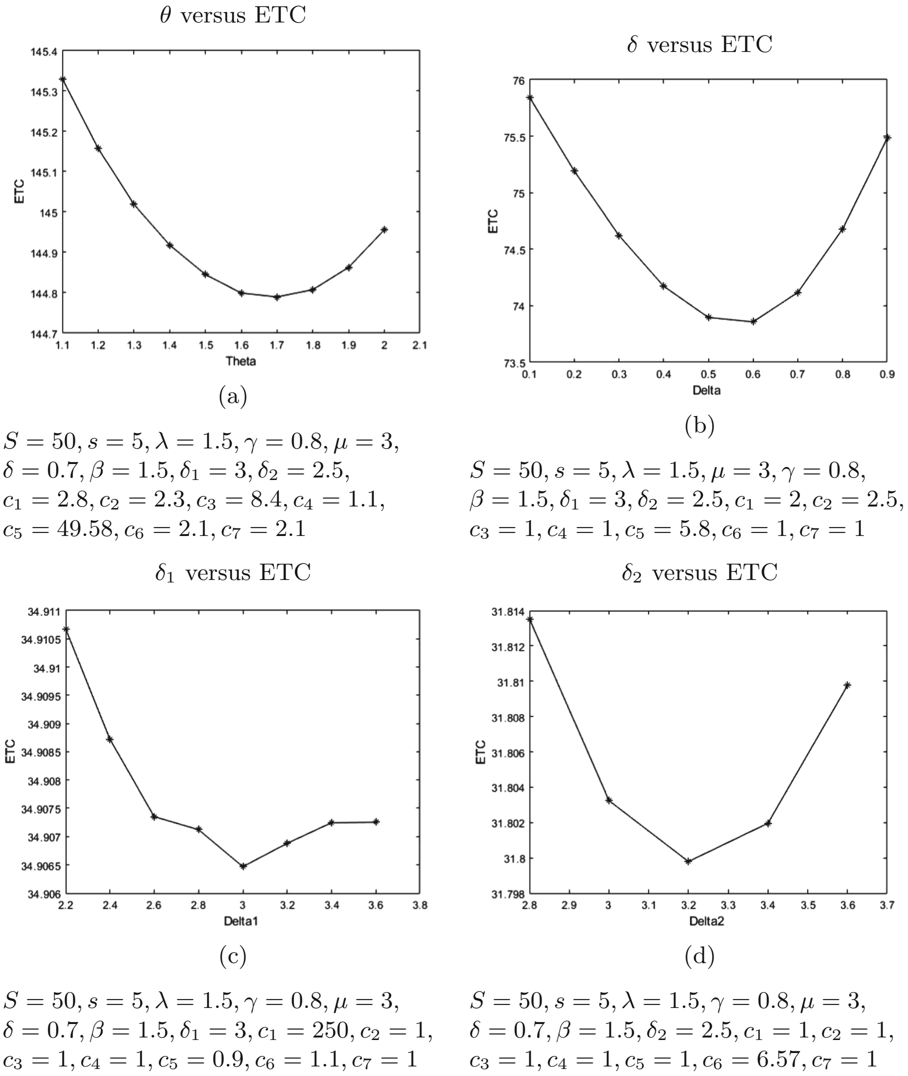
(c)

$S = 50, s = 5, \lambda = 1.5, \mu = 3, \delta = 0.7,$
$beta = 1.5, \delta_1 = 3, \delta_2 = 2.5, c_1 = 30,$
$c_2 = 2.2, c_3 = 2, c_4 = 1, c_5 = 1,$
$c_6 = 1.3, c_7 = 1$

(d)

$S = 50, s = 5, \lambda = 1.5, \gamma = 0.8, \mu = 3,$
$\delta = 0.7, \beta = 1.5, \delta_1 = 3, \delta_2 = 2.5,$
$c_1 = 1.35, c_2 = 1.057, c_3 = 49,$
$c_4 = 7.3, c_5 = 4.2, c_6 = 5.2, c_7 = 2.32$

Fig. 1. Variation of ETC with respect to various parameters



(a)

$S = 50, s = 5, \lambda = 1.5, \gamma = 0.8, \mu = 3,$
$\delta = 0.7, \beta = 1.5, \delta_1 = 3, \delta_2 = 2.5,$
$c_1 = 2.8, c_2 = 2.3, c_3 = 8.4, c_4 = 1.1,$
$c_5 = 49.58, c_6 = 2.1, c_7 = 2.1$

(b)

$S = 50, s = 5, \lambda = 1.5, \mu = 3, \gamma = 0.8,$
$\beta = 1.5, \delta_1 = 3, \delta_2 = 2.5, c_1 = 2, c_2 = 2.5,$
$c_3 = 1, c_4 = 1, c_5 = 5.8, c_6 = 1, c_7 = 1$

(c)

$S = 50, s = 5, \lambda = 1.5, \gamma = 0.8, \mu = 3,$
$\delta = 0.7, \beta = 1.5, \delta_1 = 3, c_1 = 250, c_2 = 1,$
$c_3 = 1, c_4 = 1, c_5 = 0.9, c_6 = 1.1, c_7 = 1$

(d)

$S = 50, s = 5, \lambda = 1.5, \gamma = 0.8, \mu = 3,$
$\delta = 0.7, \beta = 1.5, \delta_2 = 2.5, c_1 = 1, c_2 = 1,$
$c_3 = 1, c_4 = 1, c_5 = 1, c_6 = 6.57, c_7 = 1$

Fig. 2. Variation of ETC with respect to various parameters

5 Conclusion

This paper addressed a production inventory system with different production rates, retrial of customers and server interruptions. Different production rates were considered to minimize the customer's loss during stock-out period. Essential performance measures were derived and a suitable cost function was constructed. Numerical and graphical illustrations were conducted to analyze the total costs concerning the variations in the parameters. An optimum (s, S) pair which minimizes the ETC is also calculated. This work can be extended further by considering the Markovian Arrival Process (MAP) in the place of the Poisson process.

References

1. Berman, O.,Kim, E., Shimshack, D.G.: Deterministic approximations for inventory management at service facilities. IIE Trans. **25**(5), 98–104 (1993). (2018)
2. Jose,K.P., Salini S. Nair: Analysis of two production inventory systems with buffer, retrials an different production rates. J. Indust. Eng. Int. **13**, 369–380 (2017)
3. Jose,K.P., Salini S. Nair: A production inventory system with different rates of production and retrials. Int. J. Math. Arch. **9**(12), 30–40
4. Krishnamoorthy A., Jose, K.P.: Three production inventory systems with service, loss and retrial of customers. Int. J. Inform. Manage. Sci. **19**(3), 367–389 (2008)
5. Krishnamoorthy, A., Pramod, P.K., Deepak, T.G.: On a queue with interruptions and repeat or resumption of service. Nonlinear Anal. Theory Methods Appl. **71**(12), 1673–1683 (2009)
6. Krishnamoorthy, A., Nair, S.S., Viswanath C.N.: An inventory model with server interruptions: Oper. Res. Int. J. **12**, 151–171 (2012)
7. Neuts, M.F.: Matrix-Geometric solutions in Stochastic models: An algorithmic approach (1981)
8. Neuts, M.F., Rao, B,M. : Numerical investigation of a multiserver retrial model. Queue. Syst. **7**(2), 169–189 (1990)
9. Rashid, R., Hoseini, S.F., Gholamian, M.R., Feizabadi, M.: Application of queueing theory in production inventory optimization. J. Indust. Eng. Int. **11**(4), 485–494 (2015)
10. Rejitha, K.R., Jose, K.P.: A queueing inventory system with MAP, retrials and different replenishment rates. Int. J. Pure Appl. Math. **117**(11), 289–297 (2017)
11. Salini.S.Nair, Jose, K.P.: A MAP/PH/1 Production inventory model with varying service rates. Int. J. Pure Appl. Math. **117**(12), 373–380 (2017)
12. Nair, S.S., Jose, K.P.: A *PH* distributed production inventory model with different modes of service and *MAP* arrivals. In: Joshua, V.C., Varadhan, S.R.S., Vishnevsky, V.M. (eds.) Applied Probability and Stochastic Processes. ISFS, pp. 263–279. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-5951-8_16
13. Sandhya, E., Sreenivasan, C., Skaria, S., Nair, S.S.: An explicit solution for an inventory model with server interruption and retrials. In: Dudin, A., Nazarov, A., Moiseev, A. (eds.) ITMM 2022. CCIS, vol. 1803, pp. 149–161. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-32990-6_13

Effectiveness of N-Policy in Managing a Multi-server Stochastic Inventory System with Service Time: A Comprehensive Analysis

N. J. Thresiamma[1] and K. P. Jose[2](✉)

[1] Government Polytechnic College, Muttom 685527, Kerala, India
[2] PG and Research Department of Mathematics, St. Peter's College, Kolenchery 682311, Kerala, India
kpjspc@gmail.com

Abstract. Stochastic models for inventory management are powerful tools that help to optimize inventory decisions and reduce inventory risk. This paper presents an N-policy in a multi-server continuous review (s, S) stochastic inventory system with service time. The system comprises c homogeneous parallel servers, activated based on predetermined queue length thresholds. Inter-arrival time, service time, and lead time are assumed to follow exponential distributions. The study establishes a necessary and sufficient condition for system stability. The Matrix-Geometric method is used to study the characteristics of this model. The performance measures of the system are found, and an appropriate cost function associated with it is developed. The performance evaluation and optimization issues are also discussed and presented both numerically and graphically. Through rigorous numerical calculations, the optimal (s,S) pair is determined for specific parameters. This comprehensive analysis enhances our understanding of stochastic models in inventory management, offering valuable insights into system dynamics and facilitating informed decision-making for efficient inventory control and risk reduction.

Keywords: Inventory · Service time · Multi-Server · N-Policy · Matrix-Geometric Method

Introduction

Effective inventory management is crucial for shaping a business's performance, directly impacting profitability and customer satisfaction. Business managers employ various strategies to maximize profits. For instance, restaurants and

The authors acknowledge the financial support provided by the FIST Program, Dept. of. Science and Technology, Govt. of India, to the PG & Research Dept. of Mathematics through SR/FST/College-2018- XA 276(C).

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 161–176, 2024.
https://doi.org/10.1007/978-3-031-65385-8_12

supermarkets often assign extra staff to cater to peak-time customers while utilizing these employees for other tasks during off-peak hours. Deciding when to deploy servers and when to assign them to different roles plays a pivotal role in determining the profitability of such establishments. This work introduces a control policy for managing multi-server inventory in businesses of this nature.

Considerable research exists on multi-server queuing systems. Yadavalli et al. [13] focused on a multi-server queuing system with continuous review of perishable inventory, combining queuing theory with inventory management. Krishnamoorthy et al. [6] analyzed multi-server queuing inventory systems, emphasizing scenarios with two servers and deriving product form solutions for steady-state distributions. Jose and Beena [4] studied a retail production inventory system with two heterogeneous vacationing servers. Wang et al. [12] analyzed priority multi-server retrial inventory queues with finite queueing and orbit spaces. To enhance multi-server queuing inventory systems, Jeganathan et al. [2] introduced two types of multiserver service facilities. Recently, Samouylov et al. [9] analyzed multi-server queuing systems with flexible priorities, considering N independent identical servers and an infinite-capacity buffer.

The N-policy concept, originating in queueing literature in 1963, finds wide application in service control, including modeling production systems. Yadin and Naor [14] introduced the N-policy to minimize operational costs, and Artalejo [1] compared N, T, and D policies in an M/G/1 queueing system. Tian et al. [11] studied threshold-type vacation policies in multiserver queuing systems, and Krishnamoorthy et al. [5] extended the N-policy concept to (s, S) inventory systems. Subsequently, Jose and Thresiamma [3,10] extended this concept to production inventory systems and retrial inventory systems.

This paper unfolds in four parts: a model description, steady-state analysis, computation of system performance measures, and numerical and graphical illustrations. The study concludes by identifying the optimal (s, S)pair and determining the optimum cost for a specific set of parameters, offering valuable insights for businesses seeking to enhance their inventory management strategies. This information enables businesses to make informed decisions and optimize their operational costs in the context of multi-server inventory systems.

1 Description of the Model

This study examines a continuous review (s, S) multi-server stochastic inventory system with positive service time and lead time. The system consists of 'c' homogeneous parallel servers. Customers are admitted to the service facilities on a first-come first-served basis. The system employs an N-policy, operating as follows.: The d^{th} server, $1 \leq d \leq c$, becomes active when N_d customers accumulate in the system and have at least d items in the inventory, and is available till the queue length falls below N_{d-1}, where $N_d, 1 \leq d \leq c$ is a pre-fixed manageable level. One of the servers continues to be active until the system is empty or the inventory reaches zero. This model is built upon several foundational assumptions, including:

- The arrival of customers forms a Poisson distribution with a rate λ.
- The lead time follows an exponential distribution with a rate β.
- The service time follows an exponential distribution with a rate μ.

Key notations in this model include:

$$\begin{aligned}
N(t) &: \text{Number of customers in the system at time } t.\\
J(t) &: \text{The server status at time } t.\\
J(t) &= \begin{cases} 0, & \text{if no servers are available at time } t.\\ d, & \text{if } d \text{ servers are available at time } t;\ \text{for } d = 1, 2, \ldots, c. \end{cases}\\
I(t) &: \text{Inventory level at time } t.\\
X(t) &= (N(t), J(t), I(t))
\end{aligned}$$

$\{X(t); t \geq 0\}$ forms a Continuous Time Markov Chain with a state space represented by $\bigcup_{i=0}^{\infty} L_i$, where

$$\begin{aligned}
&L_0 = \{(0, 0, k) : k = 0, 1, \ldots, S\}\\
&\quad \text{for } 1 \leq i < N_1;\\
&L_i = \{(i, 0, k) : k = 0, 1, \ldots, S\} \bigcup \{(i, 1, k) : k = 1, 2, \ldots, S\}\\
&\quad \text{for } N_{d-1} \leq i < N_d; \quad d = 2, 3, \ldots, c;\\
&L_i = \{(i, 0, 0)\} \bigcup \{(i, j, k) : j = k; k = 1, 2, \ldots, d-1\}\\
&\quad \bigcup \{(i, j, k) : k = j, j+1, \ldots, S; j = d-1, d\}\\
&\quad \text{for } i \geq c:\\
&L_i = \{(i, 0, 0)\} \bigcup \{(i, j, k) : k = j; k = 1, 2, \ldots, c-1\}\\
&\quad \bigcup \{(i, c, k) : k = c, c+1, \ldots, S\}
\end{aligned}$$

The infinitesimal generator G of the process is a block tri-diagonal matrix and has the form:

$$G = \begin{array}{c} \underline{0} \\ \underline{1} \\ \vdots \\ \underline{N_c - 1} \\ \underline{N_c} \\ \underline{N_c + 1} \\ N_c + 2 \\ \vdots \end{array} \left[\begin{array}{ccccccccc} A_{1,0} & A_{0,0} & & & & & & & \\ A_{2,1} & A_{1,1} & A_{0,1} & & & & & & \\ & \ddots & \ddots & \ddots & & & & & \\ & & A_{2,N_c-1} & A_{1,N_c-1} & A_0, N_c - 1 & & & & \\ & & & A_{2,N_C} & A_1 & A_0 & & & \\ & & & & A_2 & A_1 & A_0 & & \\ & & & & & A_2 & A_1 & A_0 & \\ & & & & & & \ddots & \ddots & \ddots \end{array}\right]$$

where

$$[A_{0,0}](i,j) = \begin{cases} \lambda, & \text{if } i=j, i=1,2,\ldots,S+1 \\ 0, & \text{otherwise} \end{cases}$$

For $k = 1,\ldots,N_1-2$

$$[A_{0,k}](i,j) = \begin{cases} \lambda, & \text{if } i=j, i=1,2,\ldots,2S+1 \\ 0, & \text{otherwise} \end{cases}$$

For $d = 1,\ldots,c$

$$[A_{0,N_d-1}](i,j) = \begin{cases} \lambda, & \text{if } i=j, i=1,2,\ldots,S+1 \\ \lambda, & \text{if } j=(i+(d-1))-S, \\ & i=S+2,\ldots,2S+2-d \\ 0, & \text{otherwise} \end{cases}$$

For $d = 1,\ldots,c-1$; $k = N_d,\ldots,N_{d+1}-2$

$$[A_{0,N_k}](i,j) = \begin{cases} \lambda, & \text{if } i=j, i=1,2,\ldots,2S+1-d \\ 0, & \text{otherwise} \end{cases}$$

$$[A_{0}](i,j) = \begin{cases} \lambda, & \text{if } i=j, i=1,2,\ldots,S+1 \\ 0, & \text{otherwise} \end{cases}$$

$$[A_{1,0}](i,j) = \begin{cases} -(\lambda+\beta), & \text{if } i=j, i=1,2,\ldots,s+1 \\ -\lambda, & \text{if } i=j, i=s+2,\ldots,S+1 \\ \beta, & \text{if } j=S+1, i=1,2,\ldots,s+1 \\ 0, & \text{otherwise} \end{cases}$$

For $k = 1,\ldots,N_1-1$

$$[A_{1,k}](i,j) = \begin{cases} -(\lambda+\beta), & \text{if } i=j, i=1,\ldots,s+1 \\ -\lambda, & \text{if } i=j, i=s+2,\ldots,S+1 \\ -(\mu+\lambda+\beta), & \text{if } i=j, i=S+2,\ldots,S+s+1 \\ -(\lambda+\mu), & \text{if } i=j, i=S+s+2,\ldots,2S+1 \\ \beta, & \text{if } j=S+1 \,\&\, i=1,2,\ldots,s+1; \\ & \text{if } j=2S+1 \,\&\, i=S+2,\ldots,S+1+s \\ 0, & \text{otherwise} \end{cases}$$

For $k = N_d,\ldots,N_{d+1}-1$; $\quad d = 1,\ldots,c-1$

$$[A_{1,k}](i,j) = \begin{cases} -(\lambda+\beta), & \text{if } i=j=1 \\ -((i-1)\mu+\lambda+\beta), & \text{if } i=j, i=2,\dots,d+1 \\ -(d\mu+\lambda+\beta), & \text{if } i=j, i=d+2,\dots,s+1 \\ -(\lambda+d\mu), & \text{if } i=j, i=s+2,\dots,S+1 \\ -(\lambda+(d+1)\mu+\beta), & \text{if } i=j, \\ & i=S+2,\dots,S+1+s-d \\ -(\lambda+(d+1)\mu), & \text{if } i=j, \\ & i=S+s+1-d,\dots,2S+1-d \\ \beta, & \text{if } j=S+1, i=1,2,\dots,s+1 \\ \beta, & \text{if } j=2S+1-d, \\ & i=S+2,\dots,S+s+1-d \\ 0, & \text{otherwise} \end{cases}$$

$$[A_1](i,j) = \begin{cases} -\lambda-\beta, & \text{if } i=j=1 \\ -((i-1)\mu+\lambda+\beta), & \text{if } i=j, i=2,\dots,c+1 \\ -(c\mu+\lambda+\beta), & \text{if } i=j, i=c+2,\dots,s+1 \\ -(\lambda+c\mu), & \text{if } i=j, i=s+2,\dots,S+1 \\ \beta, & \text{if } j=S+1\&i=1,\dots,s+1 \\ 0, & \text{otherwise} \end{cases}$$

$$[A_{2,1}](i,j) = \begin{cases} \mu, & \text{if } j=i-(S+1), i=S+2,\dots,2S+1 \\ 0, & \text{otherwise} \end{cases}$$

For $k=2,\dots,N_1-1$

$$[A_{2,k}](i,j) = \begin{cases} \mu, & \text{if } j=1 \text{ and } i=S+2 \\ \mu, & \text{if } j=i-1, i=S+3,\dots,2S+1 \\ 0, & \text{otherwise} \end{cases}$$

For $d=1,2,\dots,c-1$;

$$[A_{2,Nd}](i,j) = \begin{cases} (i-1)\mu, & \text{if } j=i-1, i=2,\dots,d+1 \\ d\mu, & \text{if } j=i+S-d, i=d+2,\dots,S+1 \\ (d+1)\mu, & \text{if } j=i, i=S+2,\dots,2S+1-d \\ 0, & \text{otherwise} \end{cases}$$

For $k=N_1+1,\dots,N_2-1$

$$[A_{2,k}](i,j) = \begin{cases} \mu, & \text{if } j=i-1, i=2,\dots,S+1 \\ 2\mu, & \text{if } j=2, i=S+2 \\ 2\mu, & \text{if } j=i-1, i=S+3,\dots,2S \\ 0, & \text{otherwise} \end{cases}$$

For $k=N_d+1,\dots,N_{d+1}-1, \quad d=2,3,\dots,c-1$

$$[A_{2,k}](i,j) = \begin{cases} (i-1)\mu, & \text{if } j = i-1, i = 2, \ldots, d \\ d\mu, & \text{if } j = i-1, i = d+1, \ldots, S+1 \\ (d+1)\mu, & \text{if } j = d+1, i = S+2 \\ (d+1)\mu, & \text{if } j = i-1, i = S+3, \ldots, 2S+1-d \\ 0, & \text{otherwise} \end{cases}$$

$$[A_{2,Nc}](i,j) = \begin{cases} (i-1)\mu, & \text{if } j = i-1, i = 2, \ldots, c+1 \\ c\mu, & \text{if } j = i+S-c, i = c+2, \ldots, S+1 \\ 0, & \text{otherwise} \end{cases}$$

$$[A_2](i,j) = \begin{cases} (i-1)\mu, & \text{if } j = i-1, i = 2, \ldots, c+1 \\ c\mu, & \text{if } j = i-1, i = c+2, \ldots, S+1 \\ 0, & \text{otherwise} \end{cases}$$

2 Steady State Analysis

Consider the finite generator matrix $A = A_0 + A_1 + A_2$. The entries of the matrix A are characterized by the following form.

$$[A](i,j) = \begin{cases} -\beta, & \text{if } i = j = 1 \\ -((i-1)\mu + \beta), & \text{if } i = j, i = 2, c+1 \\ -(c\mu + \beta), & \text{if } i = j, i = c+2, \ldots, s+1 \\ -(c\mu), & \text{if } i = j, i = s+1, \ldots, S \\ \beta, & \text{if } j = S+1, i = 1, 2, \ldots, s+1 \\ (i-1)\mu, & \text{if } j = i-1, i = 2, \ldots, c+1 \\ (c\mu, & \text{if } j = i-1, i = c+2, \ldots, S+1 \\ 0, & \text{otherwise} \end{cases}$$

Theorem 1. *The steady-state probability vector* $\pi_A = (\pi_0, \pi_1, \ldots, \pi_S)$ *corresponding to the generator matrix* $\boldsymbol{A} = A_0 + A_1 + A_2$ *is given by* $\pi_j = \psi_j \pi_0$, *where*

$$\psi_j = \begin{cases} \prod_{k=1}^{j} \frac{(\beta+(k-1)\mu)}{k\mu}, & j = 1, \ldots, c \\ \left(\frac{c\mu+\beta}{c\mu}\right)^{j-c} \psi_c, & j = c+1, \ldots, s+1 \\ \psi_{s+1}, & j = s+2, s+3, \ldots, S \end{cases}$$

$$\pi_0 = \frac{\beta}{\psi_{s+1}(c\mu + (S-s)\beta)}$$

Proof: A satisfies the equations $\pi_A A = 0$ *and* $\pi_A e = 1$.

$$\pi_A A = 0 \implies$$

$$\begin{aligned}
-\beta\pi_0 + \mu\pi_1 &= 0, \\
-(k\mu+\beta)\pi_k + (k+1)\mu\pi_{k+1} &= 0, \qquad k = 1, 2, \ldots, c-1 \\
-(c\mu+\beta)\pi_k + c\mu\pi_{k+1} &= 0, \qquad k = c, c+1, \ldots, s \\
-c\mu\pi_k + c\mu\pi_{k+1} &= 0, \qquad k = s+1, \ldots, S \\
\beta(\pi_0 + \pi_1 + \ldots + \pi_s) - \pi_S c\mu &= 0.
\end{aligned} \tag{1}$$

Solving the system of Eqs. (1) *and using the normalizing condition* $\pi_A e = 1$, *one obtains the required result.*

Theorem 2. *The process* $\{X(t)|t \geq 0\}$ *is stable if and only if* $\lambda < \mu(c(1-\pi_0) + l\pi_0)$, *where*

$$l = \sum_{t=1}^{c-1}(c-t)\prod_{j=1}^{t}\frac{\beta + (j-1)\mu}{j\mu} \text{ and}$$

$$\pi_0 = \frac{\beta}{\psi_{s+1}(c\mu + (S-s)\beta)}.$$

Proof: Since the process $\{X(t)|t \geq 0\}$ *is a level-independent QBD process for* $i \geq N_c + 1$, *it will be stable if and only if* $\pi_A A_0 e < \pi_A A_2 e$ *(see Neuts [8]). Here* $\pi_A A_0 e = \lambda$ *and* $\pi_A A_2 e = \mu(c(1-\pi_0) + l\pi_0)$.

Using Theorem 1, the result is obtained.

2.1 The Steady State Probability Vector of G

Let the steady- state probability vector $\mathbf{x}$ of G be partitioned according to the levels as $\mathbf{x} = (x_0, x_1, ..., x_{N_c}, ...)$. The steady state solution takes the form (refer to Latouche and Ramaswami [7].)

$$x_{N_c+1+j} = x_{N_c+1}R^j : j \geq 1,$$

where R is the minimal nonnegative solution of the matrix quadratic equation

$$R^2 A_2 + RA_1 + A_0 = 0.$$

R can be calculated from the iterative procedure (see Neuts [8])

$$R_{n+1} = -(R_n^2 A_2 + A_0)A_1^{-1}$$

Also $\mathbf{x}$ satisfies the equations $\mathbf{x}G = 0$ and $\mathbf{x}e = 1$.
This leads to the following system of equations

$$\begin{aligned}
x_0 A_{1,0} + x_1 A_{2,1} &= 0, \\
x_{i-1}A_{0,i-1} + x_i A_{1,i} + x_{i+1}A_{2,i+1} &= 0, \quad 1 \leq i \leq N_c - 1. \\
x_{N_c-1}A_{0,N_c-1} + x_{N_c}A_1 + x_{N_c+1}A_2 &= 0, \\
x_{N_c}A_0 + x_{N_c+1}(A_1 + RA_2) &= 0, \\
\textstyle\sum_{i=0}^{N_c+1} x_i e + x_{N_c+1}(I-R)^{-1}e &= 1
\end{aligned} \tag{2}$$

Solving the system of Eq. (2) yields the vector $\mathbf{x}$.

3 System Performance Measures

The steady-state probability vector for the system enables the calculation of various measures of system effectiveness. The components of the vector $\mathbf{x}$ are partitioned as follows: $\mathbf{x}$ as

$$\begin{aligned}
&x_0 = (x(0,0,0), x(0,0,1), \ldots, x(0,0,S))\\
&\quad \text{for } 1 \leq i \leq N_1 - 1,\\
&x_i = (x(i,0,0), x(i,0,1), ..., x(i,0,S), x(i,1,1), \ldots, x(i,1,S))\\
&\quad \text{for } N_d \leq i \leq N_{d+1} - 1; d = 1, \ldots, c-1\\
&x_i = (x(i,0,0), x(i,1,1), \ldots, x(i,d,d), x(i,d,d+1), \ldots, x(i,d,S),\\
&\qquad x(i,d+1,d+1), (x,d+1,d+2), \ldots, x(i,d+1,S))\\
&\quad \text{for } i \geq N_c,\\
&x_i = (x(i,0,0), x(i,1,1), \ldots, x(i,c,c), x(i,c,c+1), x(i,c,S))
\end{aligned}$$

With these notations, essential system performance measures are defined and detailed below.

1 Expected Number of customers in the system

$$\begin{aligned}
EC = \sum_{i=1}^{\infty} i x_i e = \sum_{i=1}^{N_1-1} i x_i e + \sum_{d=1}^{c-1} \sum_{i=N_d}^{N_{d+1}-1} i x_i e\\
+ N_c x_{N_c} e + x_{N_c+1}((N_c)(I-R)^{-1} + (I-R)^{-2})e.
\end{aligned}$$

2 Expected inventory level

$$\begin{aligned}
EI = \sum_{k=1}^{S} k x(0,0,k) + \sum_{i=1}^{N_1-1} \sum_{k=1}^{S} \sum_{j=0}^{1} k x(i,j,k)\\
+ \sum_{d=1}^{c-1} \sum_{i=N_d}^{N_{d+1}-1} (\sum_{k=1}^{d} k x(i,k,k) + \sum_{k=d+1}^{S} k x(i,d,K) + \sum_{k=d+1}^{S} k x(i,d+1,k))\\
+ \sum_{i=N_c}^{\infty} (\sum_{k=1}^{c-1} k x(i,k,k) + \sum_{k=c}^{S} k x(i,c,k)).
\end{aligned}$$

3 Expected reorder level

$$\begin{aligned}
ER = \mu \sum_{i=1}^{N_1-1} x(i,1,s+1) + c\mu \sum_{i=N_c}^{\infty} x(i,c,s+1)\\
+ \sum_{d=1}^{c-1} (d\mu \sum_{i=N_d}^{N_{d+1}-1} x(i,d,s+1)) + (d+1)\mu \sum_{i=N_d}^{N_{d+1}-1} x(i,d+1,s+1)).
\end{aligned}$$

4 Expected departure rate

$$ED = \mu \sum_{i=1}^{N_1-1} \sum_{k=1}^{S} x(i,1,k) + \sum_{d=1}^{c-1} \sum_{i=N_d}^{N_{d+1}-1} (\sum_{k=1}^{d} k\mu x(i,k,k)$$
$$+ d\mu \sum_{k=d+1}^{S} x(i,d,k) + (d+1)\mu \sum_{k=d+1}^{S} x(i,d+1,k))$$
$$+ \sum_{i=N_c}^{\infty} (\sum_{k=1}^{c-1} k\mu x(i,k,k) + c\mu \sum_{k=c}^{S} kx(i,c,k)).$$

5 Probability that exactly d servers are busy

i d=0:

$$P_{idle} = \sum_{i=0}^{N_1-1} \sum_{k=0}^{S} x(i,o,k) + \sum_{i=N_1}^{\infty} x(i,0,0).$$

ii d=1:

$$P_{1busy} = \sum_{i=1}^{N_1-1} \sum_{k=1}^{S} x(i,1,k) + \sum_{i=N_1}^{\infty} x(i,1,1).$$

iii $1 < d < c$:

$$P_{dbusy} = \sum_{i=N_{d-1}}^{N_{d+1}-1} \sum_{k=d}^{S} x(i,d,k) + \sum_{i=N_{d+1}}^{\infty} x(i,d,d).$$

iv d=c:

$$P_{busy} = \sum_{i=N_{c-1}}^{\infty} \sum_{k=c}^{S} x(i,c,k).$$

6 Expected switching rate of servers

$$ES = \lambda \left(\sum_{d=1}^{c} \sum_{k=d}^{S} x(N_d - 1, d-1, k) \right) + \beta \left(\sum_{d=1}^{c} \sum_{i=N_d}^{\infty} x(i, d-1, d-1) \right).$$

3.1 Expected Total Cost

The expected total cost (ETC) is defined as:

$$ETC = c_1 \cdot EC + c_2 \cdot EI + c_3 \cdot ER + c_4 \cdot ED + c_5 \cdot ES,$$

where,

c_1 : holding cost of the customer/unit time,
c_2 : holding cost of inventory/unit/unit time,
c_3 : ordering cost/order/unit time,
c_4 : cost of service/unit/unit time,
c_5 : switching cost for the server.

4 Numerical and Graphical Illustrations

This section provides a description of the numerical experiments that were conducted to investigate the effects of variations in various parameters on the performance measures and expected total cost. The values presented in Tables 1 to 5 correspond to calculations performed for a system with four servers ($c = 4$) and threshold levels set as $N_1 = 4$, $N_2 = 7$, $N_3 = 10$, and $N_4 = 12$. These results provide insights into the system's behavior under different configurations, shedding light on the effects of parameter variation on its performance and cost dynamics.

In Table 1, the variations in performance measures and Expected Total Cost (ETC) with respect to maximum inventory level S are highlighted. The analysis reveals the following observations as S increases from 22 to 29:

- The Expected number of customers, Expected re-order rate, expected switching rate of servers, and the probability that all servers are idle exhibit a consistent decrease. This trend is natural since S represents the maximum inventory.
- The Expected inventory, expected departure rate, and the probability that all servers are busy also experience a decrease, aligning with expectations.
- The expected total cost shows a convex variation, as illustrated in Fig. 1(a). For the specific set of parameter values described in Table 1, the minimum cost is attained at $S = 25$.

In Table 2, variations concerning the reorder level 's' are highlighted. The analysis reveals the following observations as s increases from 5 to 11

- The Expected number of customers, Expected inventory, Expected re-order rate and expected switching rate of servers exhibit a slight increase. This upward trend is expected since as s increases, the on-hand inventory and related characteristics also increase.
- There is a slight decrease in the expected switching rate of servers. This decrease can be attributed to the reduced likelihood of servers becoming idle due to a lack of inventory. There are negligible variations in the probability that the servers are idle and busy. This suggests that changes in the parameter 's' have minimal impact on the states of the servers.
- The expected total cost demonstrates a convex variation, as depicted in Fig. 1(b). For the specific set of parameter values outlined in Table 2, the minimum cost is achieved at s=7.

Table 3 illustrates the changes in performance measures as the arrival rate λ ranges from 0.6 to 1.3

- An increase in the arrival rate corresponds to a significant rise in the expected number of customers, leading to an increase in the departure rate and the busy status of servers. The increased expected departure rate results in a decrease in expected inventory and the probability of servers being idle. This reduction in inventory triggers an increase in the re-order rate.

- The expected total cost displays a convex pattern, as illustrated in Fig. 1(c). For the specific parameter values detailed in Table 3, the minimum cost is attained at $\lambda = 0.9$.

Table 4 illustrates the changes in performance measures as the service rate μ ranges from 1 to 1.7

- An increase in the service rate leads to a substantial decrease in the expected number of customers, expected inventory, and the busy server status. Simultaneously, the departure rate increases, consequently causing an uptick in the reorder rate.
- The expected total cost exhibits a convex pattern, as depicted in Fig. 1(d). For the specific parameter values outlined in Table 5, the minimum cost is achieved at $\mu = 1.3$.

Table 5 illustrates the changes in performance measures as the replenishment rate β ranges from 0.6 to 1.4

- An increase in the replenishment rate corresponds to a significant rise in the expected inventory, leading to an increase in the departure rate and hence the number of customers in the system reduces.
- The expected total cost displays a convex pattern, as illustrated in Fig. 1(e). For the specific parameter values detailed in Table 5, the minimum cost is attained at $\beta = 0.9$.
  Table 6 depicts variations in the Expected Total Cost as the parameters s and S simultaneously change, considering the set of specified parameter values detailed beneath Table 6. The corresponding three-dimensional plot (Fig. 1(f)) illustrates this variation, revealing a convex pattern. The optimal (s,S) pair for the system within the specified parameter values is determined to be (4,28), representing the point of minimum cost. Observations across the tables indicate that changes in parameters have a noticeable impact on various performance measures and the Expected Total Cost. Tables revealing convex variations in expected total cost suggest opportunities for optimization.

Table 1. Variations in Performance Measures and ETC w.r.t S

S	EC	EI	ER	ED	ES	PBI	PB	ETC
22	5.7372	12.7072	0.6077	7.4288	0.3494	0.2239	0.0488	16.4505
23	5.7356	13.2255	0.5826	7.4566	0.3468	0.2223	0.0490	16.4385
24	5.7340	13.7420	0.5596	7.4822	0.3443	0.2209	0.0493	16.4313
25	5.7327	14.2571	0.5383	7.5059	0.3421	0.2195	0.0495	**16.4284**
26	5.7314	14.7707	0.5185	7.5278	0.3400	0.2183	0.0497	16.4294
27	5.7302	15.2832	0.5002	7.5482	0.3380	0.2172	0.0499	16.4339
28	5.7291	15.7945	0.4831	7.5672	0.3362	0.2161	0.0501	16.4414
29	5.7280	16.3049	0.4671	7.5849	0.3345	0.2151	0.0503	16.4518

$s = 7 : \lambda = 1 : \mu = 1.3 : \beta = 0.8 c = 4 : c_1 = 2 : c_2 = 0.1 : c_3 = 2 : c_4 = 0.1 : c_5 = 5 :$

Table 2. Variations in Performance Measures and ETC w.r.t s

s	EC	EI	ER	ED	ESR	PBi	PBB	ETC
5	6.9294	12.0936	0.3447	7.5649	0.2007	0.1786	0.0855	45.5834
6	6.9263	12.2193	0.3662	7.5663	0.2006	0.1785	0.0855	45.5790
7	6.9236	12.3385	0.3877	7.5674	0.2006	0.1785	0.0855	**45.5773**
8	6.9214	12.4512	0.4092	7.5683	0.2006	0.1784	0.0855	45.5774
9	6.9195	12.5575	0.4308	7.5690	0.2006	0.1784	0.0855	45.5785
10	6.9178	12.6573	0.4524	7.5697	0.2006	0.1784	0.0854	45.5805
11	6.9163	12.7508	0.4741	7.5703	0.2005	0.1783	0.0854	45.5829

$S = 25 : \lambda = 1.8 : \mu = 1.3 : \beta = 0.4 c = 4 : c_1 = 6 : c_2 = c_3 = c_4 = 0.1 : c_5 = 4 :$

Table 3. Variations in Performance Measures and ETC w.r.t λ

λ	EC	EI	ER	ED	ESR	PBi	PBB	ETC
0.6	5.2198	14.9248	0.4738	6.3439	0.2453	0.2911	0.0383	19.8398
0.7	5.4016	14.7279	0.4897	6.5843	0.2677	0.2628	0.0418	19.8322
0.8	5.5570	14.5658	0.5007	6.7718	0.2902	0.2402	0.0453	19.8282
0.9	5.6932	14.4291	0.5079	6.9202	0.3130	0.2216	0.0488	**19.8267**
1	5.8153	14.3116	0.5122	7.0388	0.3363	0.2062	0.0524	19.8274
1.1	5.9265	14.2088	0.5142	7.1338	0.3600	0.1931	0.0560	19.8297
1.2	6.0296	14.1175	0.5144	7.2101	0.3841	0.1819	0.0597	19.8335
1.3	6.1265	14.0356	0.5132	7.2711	0.4087	0.1723	0.0635	19.8386

$S = 25 : s = 7 : \mu = 1.2 : \beta = 0.8 : c = 4 : c_1 = 0.5 : c_2 = c_3 = 1 : c_4 = 0.25 : c_5 = 1$

Table 4. Variations in Performance Measures and ETC w.r.t μ

μ	EC	EI	ER	ED	ESR	PBi	PBB	ETC
1	6.0041	14.4463	0.4541	6.0507	0.3262	0.1781	0.0600	44.7063
1.1	5.9051	14.3742	0.4842	6.5541	0.3310	0.1924	0.0558	44.6509
1.2	5.8153	14.3116	0.5122	7.0388	0.3363	0.2062	0.0524	44.6215
1.3	5.7327	14.2571	0.5383	7.5059	0.3421	0.2195	0.0495	**44.6127**
1.4	5.6561	14.2095	0.5627	7.9565	0.3482	0.2325	0.0471	44.6204
1.5	5.5846	14.1680	0.5856	8.3916	0.3545	0.2450	0.0449	44.6411
1.6	5.5174	14.1317	0.6073	8.8119	0.3611	0.2571	0.0431	44.6725
1.7	5.4540	14.0999	0.6277	9.2182	0.3678	0.2688	0.0414	44.7124

$S = 25 : s = 7 : \lambda = 1 : \beta = 0.8 : c = 4 : c_1 = c_2 = 2c_3 = 1 : c_4 = 0.5 : c_5 = 1$

Table 5. Variations in Performance Measures and ETC w.r.t β

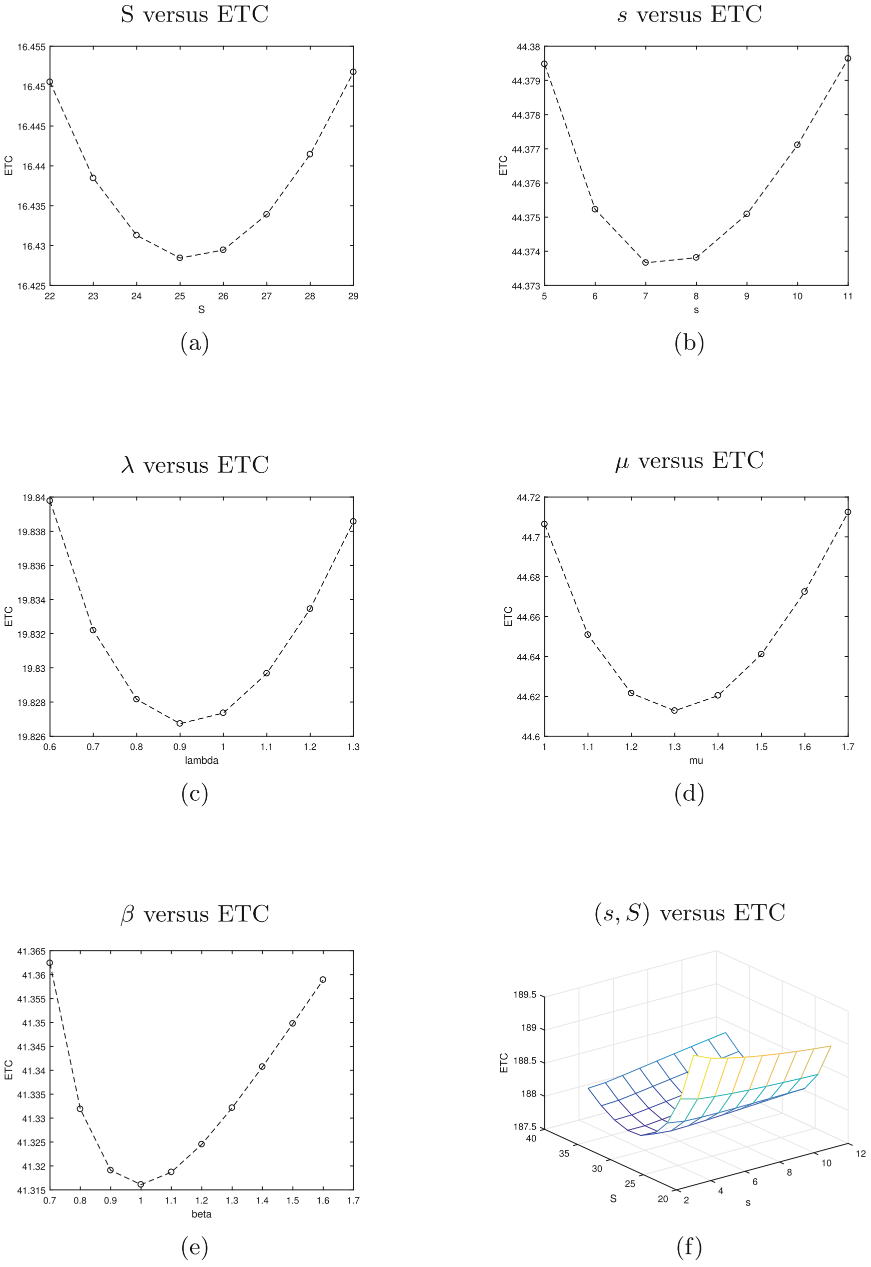
β	EC	EI	ER	ED	ESR	PBi	PBB	ETC
0.6	5.8623	13.7850	0.4655	6.9496	0.3257	0.2135	0.0520	28.3989
0.7	5.8345	14.0696	0.4908	7.0003	0.3315	0.2093	0.0522	28.3673
0.8	5.8153	14.3116	0.5122	7.0388	0.3363	0.2062	0.0524	28.3562
0.9	5.8012	14.5214	0.5305	7.0690	0.3403	0.2038	0.0526	**28.3554**
1	5.7905	14.7063	0.5465	7.0935	0.3437	0.2020	0.0528	28.3602
1.1	5.7820	14.8711	0.5605	7.1137	0.3467	0.2005	0.0531	28.3679
1.2	5.7751	15.0195	0.5730	7.1308	0.3493	0.1993	0.0533	28.3770
1.3	5.7694	15.1543	0.5841	7.1454	0.3516	0.1983	0.0535	28.3868
1.4	5.7645	15.2774	0.5941	7.1580	0.3537	0.1974	0.0537	28.3969

$S = 25 : s = 7 : \lambda = 1 : \mu = 1.2 : c = 4 : c_1 = 4 : c_2 = 0.1 : c_3 = 1 : c_4 = 0.4 : c_5 = 1$

Table 6. Variations in ETC w.r.t (s, S)

S	s							
	3	4	5	6	7	8	9	10
20	188.713	188.642	188.610	188.586	188.568	188.553	188.541	188.530
22	188.258	188.215	188.206	188.203	188.203	188.206	188.209	188.213
24	188.001	187.977	187.984	187.995	188.008	188.022	188.037	188.051
26	187.885	187.875	187.892	187.913	187.935	187.957	187.980	188.001
28	187.872	**187.871**	187.896	187.924	187.952	187.981	188.009	188.036
30	187.937	187.943	187.973	188.006	188.039	188.072	188.104	188.135
32	188.061	188.073	188.107	188.143	188.180	188.216	188.252	188.286
34	188.233	188.249	188.286	188.325	188.365	188.404	188.442	188.479

$\mu = 1.3 : \beta = 0.3 : \lambda = 1.5 : c = 4 : c_1 = 18 : c_2 = 0.5 : c_3 = 0.1 : c_4 = 2 : c_5 = 2 :$



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 1. Variation of ETC with respect to various parameters

Conclusion

A continuous review (s, S) stochastic inventory system with $c, (c > 1)$ servers is studied in this work. The N-policy is implemented for the service facility across multiple stages. A condition necessary and sufficient for system stability is derived, and the Matrix Geometric Method is utilized for system analysis. Key performance measures and a cost function based on these metrics are developed. The system's performance is numerically assessed and graphically visualized. Specifically, the optimal (s, S) pair for a 4-server system with threshold stages $N_1 = 2 : N_2 = 8 : N_3 = 12 : N_4 = 16$ is determined. For the given parameter values and costs $\mu = 1.3 : \beta = 0.3 : \lambda = 1.5 : c_1 = 18 : c_2 = 0.5 : c_3 = 0.1 : c_4 = 2 : c_5 = 2$: the optimum value is found to be 187.871 and $(4, 8)$ is the optimal (s, S).pair. This study suggests potential extensions to inventory systems involving multi-production units or production inventory systems with multiple servers.

References

1. Artalejo, J.: A unified cost function for m /g/1 queueing systems with removable server. Trabajos de Investigacion Operativa **7**, 95–104 (1992)
2. Jeganathan, K., et al.: Analysis of interconnected arrivals on queueing-inventory system with two multi-server service channels and one retrial facility. Electronics **10**(5) (2021). https://www.mdpi.com/2079-9292/10/5/576
3. Jose, K.P., Thresiamma, N.J.: N-policy on a retrial inventory system. In: Dudin, A., Nazarov, A., Moiseev, A. (eds.) Information Technologies and Mathematical Modelling. Queueing Theory and Applications: 21st International Conference, ITMM 2022, Karshi, Uzbekistan, October 25–29, 2022, Revised Selected Papers, pp. 200–211. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-32990-6_17
4. Jose, K.P., Beena, P.: On a retrial production inventory system with vacation and multiple servers. Int. J. Appl. Comput. Math. **6**(4), 1–17 (2020)
5. Krishnamoorthy, A., Narayanan, V.C., Deepak, T., Vineetha, P.: Control policies for inventory with service time. Stochastic Anal. Appl. **24**(4), 889–899 (2006). https://doi.org/10.1080/07362990600753635, https://doi.org/10.1080/07362990600753635
6. Krishnamoorthy, A., M.R.S.D.: Analysis of multi-server queueing system. Adv. Oper. Res. **2015** (2015)
7. Latouche, G., Ramaswami, V.: Introduction to matrix analytic methods in stochastic modelling. Philadelphia: SIAM
8. Neuts, M.: Matrix-Geometric Solutions in Stochastic Models - an Algorithmic Approach. John Hopkins University Press
9. Samouylov, K., Dudina, O., Dudin, A.: Analysis of multi-server queueing system with flexible priorities. Mathematics **11**(4) (2023). https://www.mdpi.com/2227-7390/11/4/1040
10. Thresiamma, N.J., Jose, K.P.: N-policy for a production inventory system with positive service time. Information Technologies and Mathematical Modelling. Queueing Theory and Applications, pp. 52–66 (2022)

11. Tian, N., Zhang, G.: A two threshold vacation policy in multiserver queueing systems. Europ. J. Oper. Res. **168**, 153–163 (01 2006). https://doi.org/10.1016/j.ejor.2004.01.053
12. Wang, F.F., Bhagat, A., Chang, T.M.: Analysis of priority multi-server retrial queueing inventory systems with MAP arrivals and exponential services. Opsearch **54**(1), 44–66 (2017). https://doi.org/10.1007/s12597-016-0270-9
13. Yadavalli, V., Sivakumar, B., Arivarignan, G., Adetunji, O.: A multi-server perishable inventory system with negative customer. Comput. Ind. Eng. **61**(2), 254–273 (2011). https://doi.org/10.1016/j.cie.2010.07.032
14. Yadin, M., Naor, P.: Queueing systems with a removable service station. OR **14**, 393–405 (1963). https://doi.org/10.2307/3006802

Statistical Evaluation of Input Flow Intensity in the Presence of an Interfering Parameter

Gurami Tsitsiashvili(✉)

Institute for Applied Mathematics, Far Eastern Branch of Russian Academy Sciences,
Radio Street 7, Vladivostok, Russia
guram@iam.dvo.ru

Abstract. In this paper, we determine the number of visits of the animals to the salt lake from observations from a camera trap. The main difficulty of the task is that the same animal can be registered and counted several times as different animals. The total number of animals that came to solonets is determined by the formula for the stationary distribution of the number of applications in a queuing system with an infinite number of devices. The interfering parameter is the intensity of service. The problem is solved using the ergodicity of the service process. In systems with different intensity of the flow of animals, the intensity ratio is determined by the ratio of the number of observations from the camera trap.

Keywords: Interfering parameter · Ergodicity · Queuing system

1 Introduction

Estimating the parameters of a statistical sample in the presence of an interfering parameter is an important statistical task [1,2]. This problem has found its application in quantum physics, where usually interfering parameters are called hidden parameters [3]. It was also solved when creating quantum computers [4]. Methods for solving the problem of interfering parameters were used in the analysis of tiger tracks in Primorye in 2005. [5]. In this article, the probability of detecting a trace is selected as an interfering parameter. Our choice of the interfering parameter allows us to use the theorem on colouring points of the Poisson flow [6] to estimate the parameter of the Poisson distribution of the relative number of flow points in a given area. We consider the problem of assessing the intensity of visits by animals to camera traps set for animals at feeding sites or salt shakers. Here, the interference parameter characterizes the average time spent by the animal on the salt shaker.

The problem of processing data on the time of stay and frequency of visits by animals to camera traps has recently become increasingly relevant. Photos and videos of automatic registration of animals (data of camera traps) allow us to

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 177–187, 2024.
https://doi.org/10.1007/978-3-031-65385-8_13

identify some rare species well, for example, large cats (tiger, leopard) by their individual colouring (a combination of stripes and spots on the animal's skin). For photos and videos of the most valuable hoofed animals (raisins, roe deer, elk), personal identification of animals is not possible. An exception may be a short period (one and a half to two months a year), when adult males (bulls) may be distinguished by the configuration of the horns. Therefore, camera traps record the approach of such animals, but without their individualization. This makes it difficult to account the number of animals recorded by the camera trap for some time interval. To overcome this difficulty, it is necessary to use the elements of queuing theory in relation to a system with an infinite number of servers. In particular, such a system arises when counting the approach of animals to salt pans. If we assume that an animal can approach the salt shaker once a day and its successive approaches will be daily until saturation, then the time of the animal's stay at the salt shaker can be interpreted as service time, and the number of channels may be considered infinite, since the competition of animals for such a resource can be neglected. And, finally, the time of the animal's stay at the salt cellar with a sufficient margin of error may be considered the same for different periods of time (months).

This allows us to obtain information about the degree of general use of salt by animals to meet their needs, which is comparable to the number of customers in a queuing system with an infinite number of servers. In some cases, for such a system, it is not difficult to calculate the stationary distribution and to establish an ergodicity in the sense of an equality of the average value (number of animals) in the ensemble and the average value along a long trajectory. The first characteristic may be calculated from the stationary distribution of the process characterizing the number of customers in the queuing system (on the salt marsh). And the second characteristic may be calculated from observations using a camera trap.

Counting the total number of animals cannot be done without serious errors if it is performed according to the frequency of registration with a camera trap. The reason is that it is impossible to prove the difference of hoofed animals, therefore, the same animal may be registered several times and counted as different animals. Fixed information on the date and time of stay does not help in these calculations, which is also recorded by a camera trap. It is necessary to remove the interfering parameter and to build data on animal encounters as a relative number that actually coincides with the total number of registrations of all animals by month. These parameters express the average values of daily and seasonal visits of saline animals [7–9].

The second task of estimating the parameters of the Poisson flow in the presence of an interfering parameter is the task of processing winter animal tracks in a certain area. The peculiarity of this problem is the fact that each trace may be detected with some probability. This probability is an interfering parameter and depends on the thickness of the snow cover, on the organization of trail monitoring, on the financial circumstances, etc. Therefore, instead of estimating the parameters of the Poisson flow characterizing the number of traces in some districts, it is proposed to estimate the ratio of the Poisson flow parameter characterizing the number of traces in a certain district to the Poisson flow parameter characterizing the number of traces in the entire area and assumed to be large. Thus, the estimation of the parameters of the Poisson flow in the presence of an interfering parameter may be realized from long-term observations or in the presence of a large average number of observed traces.

Since both of these tasks are related to animal observations, it begs to consider another problem about the possible movements of an animal in some area. This problem may be solved in an optimization formulation assuming that the minimum number of intersections of the continuous trajectory of the animal through the boundaries of the districts of the studied area to the boundary of the entire area is sought. This problem is closely related to the problem of estimating the parameters of the Poisson distribution of the number of animal tracks in a district of a certain aria. To solve this problem, it is required to use graph-theoretic methods instead of probabilistic ones. Using these methods, we can zone the districts of a certain area according to the minimum number of intersections of a continuous curve starting in a certain district to the boundary of the entire area.

2 Final Distributions in the Queuing System $M|M|\infty$

Consider a queuing system $M|M|\infty$ with an infinite number of servers, the intensity of the Poisson input flow λ and the intensity of service μ. Denote $x(t)$ the number of customers in this system at time t. It is known that the stochastic process $x(t)$ is the death and birth process. In terms of Markov process theory $x(t)$ is ergodic and its final distribution is Poisson [10, ch. III, §3] with the parameter $\rho = \lambda/\mu$:

$$\lim_{t\to\infty} P(x(t) = k) = \frac{e^{-\rho}\rho^k}{k!},\ k = 0, 1, \dots \tag{1}$$

This final distribution has the mean ρ and the variance ρ.

At the physical level of rigour, the ergodicity of the process $x(t)$ may be determined as follows: the average value of the process over the ensemble is equal to the average value of the trajectory. From the point of view of the probability theory, this equality means the law of large numbers for the process $x(t)$. This equality may be represented in terms of the convergence in probability $\xrightarrow{P}$ as follows [11, ch. V, §2]:

$$A(T) = \frac{\int_0^T x(t)dt}{T} \xrightarrow{P} \rho,\ T \to \infty. \tag{2}$$

Remark 1. Assume that there are two queuing systems $M|M|\infty$ with input flows intensities $\lambda_1,\ \lambda_2$ and the same service intensity μ. Suppose that numbers of customers in these systems are described by the death and birth processes $x^1(t),\ x^2(t)$. Then from the formula (2) we have

$$\frac{A^1(T)}{A^2(T)} \xrightarrow{P} \frac{\lambda^1}{\lambda^2},\ T \to \infty. \tag{3}$$

It is possible to extend this statement onto queuing systems $M|G|\infty$ also.

Remark 2. It is worthy to say that the process $x(t)$ is a ladder function of t with a unit height of steps. Therefore, the steps going up are located at the arrival moments of the input flow. Consequently the ratio between the total number of such steps $X(T)$ on the segment $[0, T]$ and T tends to λ with $T \to \infty$ by probability: $\dfrac{X(T)}{T} \xrightarrow{P} \lambda$. However, it is technically quite difficult to use this remark, since it is necessary to organize the observation of the process $x(t)$ on the segment $[0, T]$.

3 Estimation of the Relative Number of Animal Approaches to the Camera Trap

First Model of Observations. Assume that a Poisson flow of points with intensity λ is a set on the half-interval $[0, n)$. These points determine the moments when customers arrive to a queuing system. Let's split the half-interval $[0, n)$ into non intersecting half-intervals $[0, 1),\ [1, 2),\ ldots, [n-1, n)$. Denote $\xi_1, \xi_2, \dots, \xi_n$ independent and identically distributed random variables that determine the number of flow points in the selected half-intervals. Now let each point of the input Poisson flow that falls into the half-interval $[i-1, i)$, generates one new point in each of the half-intervals $[i, i+1), \dots, [i+m-1)$ (moments of fixation by the surveillance device). Then the random number of points generated by all points of the Poisson flow in the half-interval $[0, n)$ is determined by the equality

$$N_n = \sum_{k=1}^{m} k\xi_{n-k+1} + m \sum_{k=m+1}^{n} \xi_{n-k+1}.$$

Let for some finite positive numbers $M,\ \Lambda$ the inequalities $m < M,\ \lambda < \Lambda$ take place. Denote by the icon $\xrightarrow{P}$ a convergence in probability. Then we obtain the following relations:

$$\frac{\sum_{k=1}^{m} k\xi_{n-k+1}}{n} \xrightarrow{P} 0,\ \frac{\sum_{k=m+1}^{n} \xi_{n-k+1}}{n-m} \xrightarrow{P} \lambda,\ n \to \infty.$$

It follows from this [12, ch. 1, §3] that

$$S_n = \frac{N_n}{n} \xrightarrow{P} m\lambda,\ n \to \infty. \tag{4}$$

Thus, the S_n statistic is a consistent estimate of the product $m\lambda$.

Denote

$$A(m,n)=-\frac{m-1}{2n}\to 0;\ B(m,n)=\frac{1}{2n}-\frac{2\,m}{3n}+\frac{1}{6mn}\to 0,\ n\to\infty,$$

and get mathematical expectation and variance of random variable N_n

$$EN_n=mn\lambda(1+A(m,n)),\ varN_n=m^2n\lambda(1+B(m,n)).$$

Hence the equalities follow

$$ES_n=m\lambda(1+A(m,n)),\ var\frac{N_n}{\sqrt{n}}=m^2\lambda(1+B(m,n)). \tag{5}$$

From the Bieneme-Chebyshev and Cauchy-Bunyakovsky inequalities, we obtain the bound that for any $\varepsilon>0$

$$P\left(\left|\frac{N_n}{n}-m\lambda\right|>\varepsilon\right)\le\frac{2\,m^2\lambda}{n\varepsilon^2}\left(B(m,n)+\frac{\lambda(m-1)^2}{4n}\right). \tag{6}$$

Remark 3. If the parameter λ is known, then the parameter m may be estimated by the random value S_n. And vice versa, if the parameter m is known, then the parameter λ may be estimated by the random value S_n. In this formulation of the problem, we are not talking about estimating the number of animals that form the flow of customers. We consider only flows of their arrivals.

Remark 4. If S_n^i are consistent estimates of the parameters $\lambda^i,\ i=1,2$, then for general unknown parameter m, the ratio S_n^1/S_n^2 is a consistent estimate of the ratio λ^1/λ^2. This property may be extended to any finite number of parameters λ^i.

Let the random sequences $N_{n,t},\ t=1,\dots,l$ are independent and coincide by distribution with $N_n,\ n=1,\dots$ Denote $R_{n,l}=\frac{1}{l-1}\sum_{t=1}^{l}\left(\frac{N_{n,t}}{\sqrt{n}}-\sum_{j=1}^{l}\frac{N_{n,j}}{\sqrt{n}}\right)^2$, then from the consistency property of the estimate $R_{n,l}$ [12, ch. 1,§3] we get

$$R_{n,l}\xrightarrow{P}var\frac{N_n}{\sqrt{n}}\to m^2\lambda,\ n,l\to\infty.$$

Remark 5. Thus, the statistics $S_n,\ R_{n,l}$ are, for $n,l\to\infty$, consistent estimates of the quantities $m\lambda,\ m^2\lambda$ and therefore, with their help, it is possible to construct consistent estimates of the parameters $m,\ \lambda$. However, for such an assessment, it is necessary to put not one, but several devices in the vicinity of those places where animals receive the resource they need (for example, salt).

Let $\eta_k,\ k = 1, \ldots, n$ is the number of points generated by the points of the Poisson flow from the semi-intervals $[k-m, k-m+1), \ldots, [k-1, k)$. Then the equalities are fulfilled

$$\eta_k = \sum_{i=1}^{k} \xi_i,\ 1 \le k \le m;\ \eta_k = \sum_{i=k-m+1}^{k} \xi_i,\ m+1 \le k \le n. \tag{7}$$

Remark 6. It is easy to obtain from the central limit theorem the weak convergence of random variables $\dfrac{N_n - nm\lambda}{\sqrt{nm^2\lambda}}$ to the standard normal distribution for $n \to \infty$.

Remark 7. Since the parameter m is unknown, it will not display $\eta_k,\ 1 \le k \le n$, via $\xi_k,\ 1 \le k \le n$.

Second model of observations. Let us now consider another stochastic model of a queuing system and observations of it. Assume that almost certainly random variables $\eta_k^i \le c < \infty$ and each customer/animal is observed ξ_k^i times at a camera trap with an average value $E\xi_k^i = b$. Denote $\nu^i(T)$ the total number of customers arriving queuing system at the segment $[0, T]$ and put $D^i(T)$ the total number of camera trap observations at $[0, T]$. Then we have almost surely the following inequality

$$\sum_{k=1}^{\nu^i(T-c)} \eta_k^i \le D^i(T) \le \sum_{k=1}^{\nu^i(T)} \eta_k^i.$$

Using Wald identity [13] it is possible to obtain the relation

$$b\lambda^i(T-c) \le ED^i(T) \le b\lambda^i(T),\ i = 1, 2.$$

Consequently the following limit formula is true

$$\frac{ED^1(T)}{ED^2(T)} \to \frac{\lambda^1}{\lambda^2},\ T \to \infty. \tag{8}$$

This formula allows to estimate the ratio λ^1/λ^2 by observations of queuing systems 1, 2.

Dividing Input flow into Few Sub Flows. Here $\lambda_1 T,\ \lambda_2 T$ allow to approximate the parameters of flows intensities $\lambda_1,\ \lambda_2$, respectively. These parameters express the average values of the number of animals arrivals to the system associating with the salt marsh. In real conditions, animals may arrive the salt system in groups. For example, males usually come alone, and females - with groups of animals. Therefore, there is a need to build a Poisson flow model, each point of which corresponds to a certain group of animals. But since the camera trap records the number of animals in a group, it is possible to consider several flows instead of a Poisson flow with a group intake. The customer in each of these flows corresponds to a fixed number of incoming animals.

Let's consider the simplest example when it contains one animal and the probability of containing more than one animals are analysed. Using the colourization theorem of the points of the Poisson flow [6], it is possible to construct two independent Poisson flows from the original flow. Each customer in one of these flows contains one animal, and each customer in the other flow contains more than one animal. Then, for each of these flows, similar considerations and calculations may be carried out. This simple example may be extended by assuming that each customer of the Poisson input flow corresponds with some probability to a certain number of animals: one, two, three, etc.

4 Method of Eliminating the Interfering Parameter in Statistics of Poisson Flow of Animal Traces

This section examines the analysis of traces of rare animals in different districts of a certain area. The results obtained in this section show that the analysis of traces of rare animals requires more careful processing of the results obtained. This becomes especially important when traces are examined over a large area and data collection turns into an industrial statistics procedure. The considered problem arises when comparing data on the number of animal tracks in the snow, when the interfering parameter takes values caused by different meteorological and economic characteristics in different districts and in different years.

Let the Lebesgue-measurable and disjoint regions be distinguished on the plane $G_k,\ k = 1, \dots, n$. Poisson point flow $\mathit{\Pi}$ with continuous intensity $\lambda(x)$ is given, and the relations are fulfilled

$$\overline{\lambda}_k = \int_{G_k} \lambda(x)dx < \infty,\ k = 1, \dots, r,\ \overline{\lambda} = \sum_{k=1}^{r} \overline{\lambda}_k.$$

Denote $\mathit{\Lambda}_k = \dfrac{\overline{\lambda}_k}{\overline{\lambda}}$ and consider the flow $\mathit{\Pi}_m$ with the intensity function $m\lambda(x)$. Let each point of the flow $\mathit{\Pi}_m$ be independent of other points and, from its coordinates with probability p, enter the flow $\overline{\mathit{\Pi}}_m$. Then, the flow $\overline{\mathit{\Pi}}_m$ due to the point colouring theorem of the Poisson flow [6] is Poisson with intensity $pm\lambda(x)$. Therefore, the number n_k of flow points $\overline{\mathit{\Pi}}_m$ in the subdomain G_k has a Poisson distribution with the parameter $pm\overline{\lambda}_k$, and the sum $n = \sum\limits_{k=1}^{r} n_k$ has a Poisson distribution with the parameter $pm\overline{\lambda}$.

Theorem 1. *The convergence in probability of the random variable* $N_k = \dfrac{n_k}{n}$ *to the parameter* $\mathit{\Lambda}_k$ *is valid when* $m \to \infty$.

Proof. Due to the properties of the Poisson distribution, the relations are fulfilled

$$En_k = Var\ n_k = pm\overline{\lambda}_k,\ k = 1, \dots, r;\ En = Var\ n = pm\overline{\lambda}. \tag{9}$$

It follows from the equalities (13) that

$$Var\,\frac{n_k}{pm\overline{\lambda}_k}=\frac{1}{pm\overline{\lambda}_k},\ Var\frac{n}{pm\lambda}=\frac{1}{pm\overline{\lambda}}.$$

From Chebyshev's inequality, we obtain that for any $\varepsilon,\ 0<\varepsilon<1$, and for $m\to\infty$

$$P\left(1-\varepsilon\le\frac{n_k}{pm\overline{\lambda}_k}\le 1+\varepsilon\right)\ge 1-\frac{1}{pm\overline{\lambda}_k\varepsilon^2}\to 1,\ k=1,\dots,r, \tag{10}$$

$$P\left(1-\varepsilon\le\frac{n}{pm\overline{\lambda}}\le 1+\varepsilon\right)\ge 1-\frac{1}{pm\overline{\lambda}\varepsilon^2}\to 1.$$

Using Formulas (13) and (10), it is not difficult for any $\varepsilon,\ 0<\varepsilon<\frac{1}{2}$, to obtain the inequality

$$P\left(1-2\varepsilon\le\frac{N_k}{\Lambda_k}\le 1+4\varepsilon\right)\ge P\left(\frac{1-\varepsilon}{1+\varepsilon}\le\frac{N_k}{\Lambda_k}\le\frac{1+\varepsilon}{1-\varepsilon}\right)\ge 1-\frac{1}{pm\overline{\lambda}_k\varepsilon^2}-\frac{1}{pm\overline{\lambda}\varepsilon^2}. \tag{11}$$

From the inequality $\Lambda_k\le 1$, and from Formula (11), we obtain the relation

$$P\left(-2\varepsilon\le N_k-\Lambda_k\le 4\varepsilon\right)\ge 1-\frac{1}{pm\lambda_k\varepsilon^2}-\frac{1}{pm\lambda\varepsilon^2}\to 1,\ m\to\infty. \tag{12}$$

This proves the statement of Theorem 1.

Therefore, the relation N_k, for $m\to\infty$, is a consistent estimate of the parameter Λ_k, free of probability p, playing the role of an interfering parameter in this problem.

Remark 8. Using Inequalities (12), we can assume that $p=p(m)=m^{-\delta},\ 0\le\delta<1$, and establish convergence by probability N_k to Λ_k at $m\to\infty$.

5 Setting the Problem of the Shortest Paths to the Border of Some Aria

A map of a certain region with areas highlighted on it is considered. It is required to zone the regions of the region according to the degree of proximity to the outer border of the region. The degree of proximity of the district to the outer border of the region means the minimum number of inter-district borders that must be crossed to get from it to the outer border of the region. It is assumed that the region is represented on the map by a flat graph consisting of polygons [14] – [16]. This problem arises when trying to plot possible animal trajectories to the boundary of a certain area.

The problem is to build an algorithm for zoning a certain area and listing the shortest paths to the outer boundary. It is assumed that this algorithm will be based on the procedure of hierarchical classification of districts and on the concept of a graph dual to a planar one. It is expected that the zoning algorithm will lead to the construction of all possible shortest paths from the districts to the outer border of the region.

Suppose that there is a finite set of $U_0 = \{u\}$ bounded polygons on the plane. By a polygon (simple) we mean a connected and disconnected set of points bounded by some closed and non-self-intersecting polyline [14, pp. 749-752.]. We assume that these polygons can only have common borders or parts of them. Then the boundaries of these polygons form a planar graph $\mathit{\Gamma}$, the vertices of which are the break points of the boundary poly lines, and the edges are rectilinear sections of the poly lines.

As a result of this construction, the polygons themselves become faces of a planar graph $\mathit{\Gamma}$ [15, Chapter 1]. Due to the limitation of the number of polygons from the set U_0, the outer face of the graph $\mathit{\Gamma}$ is not included in U_0. Each face $g \in G_0$ is mapped to a subset of the faces $S(g) \subseteq G_0$, touches the boundaries with the face G. Let's call $S(u)$ a set of adjacent faces u. A face adjacent to the outer faces is called a boundary, and not adjacent to the outer faces is called an inner one.

Let's denote $V_0 \subseteq U_0$ the set of boundary faces and $U_1 = U_0 \setminus V_0$ - the set of inner faces. Recursively define an algorithm for hierarchical classification of faces

$$U_{k+1} = \{u \in U_k : \ S(u) \subseteq U_k\}, \ V_k = U_k \setminus U_{k+1}. \tag{13}$$

We extend this recursion to the moment n, when for the first time $U_{n+1} = \oslash$ or $U_{n+1} = U_n$. As a result, we get the reward $U_0 \subset U_1 \subset \ldots \subset U_n$ and consider the set of subsets $V_k = U_k \setminus U_{k+1}, \ k = 0.1, \ldots, n-1, \ V_n = U_n$, moreover $U_k = \bigcup\limits_{j=k}^{n} V_j, \ k = 0, \ldots, n.$

Lemma 1. *The equality $U_{n+1} = U_n$ is impossible.*

Proof. We will conduct a proof of this statement from the contrary. For any face $u_k \in U_k, \ k > 0$, there exists such a $\varepsilon > 0$, that ε - the neighbourhood of the face u_k is completely contained in the set $u_k \bigcup S(u_k)$. Here, ε - neighbourhood is understood as the set of all points on the plane, the distances from which to the set $u_k \bigcup S(u_k)$ does not exceed ε.

Since the set of faces $U_k, \ k > 0$, is a collection of bounded polygons, it is possible to determine the maximum point x_k of its projection on the abscissa axis x. Let this maximum point belong to the projection on the x axis of the polygon-face $u_k \in U_k$. But then the face u_k adjoins the faces of the set $S(u_k) \subseteq U_k$ and consequently ε - the neighbourhood of the face u_k is contained in the set of faces U_k. The last statement leads to the fact that x_k cannot be the maximum projection point of the set of faces of U_k on the axis of the abscissa x.

Remark 9. From the formula (4) it follows that for all $k \leq n, \ u_k \in V_k$ there is $u_{k-1} \in S(u_k) \bigcap V_{k-1}$ so that for all $j > 1$ we have $S(u_k) \bigcap V_{kj} = \oslash$.

Lemma 2. *1) Assume that $k \leq n$, $u_k \in V_k$, then there is a face $u_{k-1} \in V_{k-1}$ so that $u_{k-1} \in S(u_k)$.*
2) For any $j > 1$ the intersection $S(u_k) \cap V_{k-j} = \oslash$.

Proof. Indeed, otherwise $S(u_k) \subseteq U_k$ and therefore the inclusion of $u_k \in_k \subseteq U_k$ is not performed. Therefore, the statement 1) is true.
Statement 2) follows from the formula (4) defining the sequence of sets of faces U_k, $k = 0, 1, \ldots, n$. Indeed, if the face $u_k \in V_k \subseteq U_k$, then the ratio $S(u_k) \subseteq U_{k-1} = \bigcup_{t=k-1}^{n} V_t$ and means $S(u_k) \bigcap V_{k-j} = \oslash$, $j > 1$, because from the formula (7) it follows that $U_{k-1} \bigcap V_{k-j} = \oslash$, $j > 1$.

An algorithm has been developed for the hierarchical classification of the faces of a bounded and connected planar graph (without an outer face), in which the faces are polygons and the edges are fragments of the boundaries between them. The algorithm allows us to build shortest paths between the face and the outer boundary of the entire graph containing the minimum number of intersecting boundaries. This algorithm is based on the allocation of edges connecting faces from the sets V_k, V_{k+1} and was applied (together with geographers V.N. Bocharnikov and S.M. Krasnopeev) for identification of possible migration routes of the Amur tiger between hunting farms Primorsky Krai.

6 Discussion

The consideration given in the article can be expanded if animal movements between camera traps are known. In this case, it is possible to build a queuing network model. It should also be noted that the definition of ergodicity used in the work, as the equality of the ensemble average and the trajectory average, is usually considered in the theory of random processes as the law of large numbers. And the proof of this law is possible with the help of the central limit theorem for an ergodic stochastic process. It should also be noted that instead of queuing systems $M|M|\infty$ we can consider a more general system $M|G|\infty$, in which the random service time has a fairly general distribution. It should be noted that the colouring theorem The use of Poisson flow points makes it possible to expand the range of statistical tasks related to observations of animal movements in a certain area. This theorem also provides an opportunity to formulate and solve an optimization problem on a planar graph. This problem naturally arises when considering the possible trajectories of animal movement. A further development of this topic is the analysis of the passage of animals to the salt marshes according to information from camera traps installed on trails approaching the salt marshes.

7 Conclusion

In this article the intensity of the flow of animal customers coming to salt shaker estimates. It is assumed that these animals constantly return to the salt shaker for several days, after which they leave the salt shaker. This system is considered as a $M|M|\infty$ queuing system with an infinite number of servers. For this, a formula is known that expresses the intensity of the input flow through the stationary distribution of the number of customers in the system and through the intensity of the service. This formula can be applied to data on the fixation of animals on camera traps. Assuming that the intensity of service is an interfering parameter, the ratio of the intensities of input flows in two systems with the same intensity of service is estimated. The reference to the theorem on the points of colouring of the Poisson flow stimulated and proved convenient for considering new problems in dividing the flow of animals into flows consisting of separate groups of animals.

References

1. Cox, D., Hinckley, D.: Theoretical statistics. Moscow: Mir. 560 p. (1978) (in Russian)
2. Young, G.A., Smith, R.L.: Essentials of Statistical Inference. Cambridge University Press, Cambridge, UK (2005)
3. Holevo, A.S.: Statistical structures of quantum mechanics and hidden parameters. Moscow: Znanie. 32 p. (1985) (In Russian)
4. Holevo, A.S.: Quantum Systems, Channels, Information. ICNMO, Moscow (2010). (In Russian)
5. Tsitsiashvili, G.S., Bocharnikov, V.N., Krasnopeev, S.M.: Method of eliminating the interfering parameter in the statistics of the Poisson flow of points. Bulletin of TSU, Management, computer engineering and Informatics. **62**(1), 101–106 (2023). (In Russian)
6. Kingman, J.: Poisson processes. Moscow: ICNMO. 136 p. (2007) (In Russian)
7. Lukarevsky, V.S., Lukarevsky, S.V.: Estimation of the number of the Far Eastern leopard (PANTHERA PARDUS) in Russia. Zoological J. **98**(5), 567–577 (2019). (In Russian)
8. Ogurtsov, S.S.: Review of software for processing data from camera traps: the latest novelties, working with video and GIS. Nature Conserv. Res. Zapovednaya nauka. **4**(2), 95–124 (2019). (In Russian)
9. Kalinkin, Y.N.: Daily activity of ungulates on the salt flats of the Altai Reserve. field research in the Altai Biosphere Reserve. Zoology. no **5**, 6–14 (2023). (In Russian)
10. Ivchenko, G.I., Kashtanov, V.A., Kovalenko, I.N.: Theory of Queuing: A Textbook For Universities. Higher School, Moscow (1982). (In Russian)
11. Kalashnikov, V.V.: Qualitative Analysis of The Behavior of Complex Systems. Nauka, Moscow (1978). (In Russian)
12. Borovkov, A.A.: Mathematical Statistics. Nauka, Evaluation of parameters. Hypothesis testing. Moscow (1984). (In Russian)
13. Borovkov, A.A.: Course of probability theory. Nauka, Moscow (1972). (In Russian)
14. The polygon. The Mathematical Encyclopedia. Vol. 3. Moscow: The Soviet Encyclopedia. 1982. (In Russian)
15. Prasolov, V.V.: Elements of Combinatorial And Differential Topology. ICNMO, Moscow (2004). (In Russian)
16. Harari, F., Norman, R.Z., Cartwright, D.: Structural Models: An Introduction To The Theory Of Oriented Graphs. Wiley, New York (1965)

The Numerical Analysis of the Time-Scale Shortest Queue Model Under the Dobrushin Mean-Field Approach

Sergey A. Vasilyev(✉), Mohamed A. Bouatta, Shahmurad K. Kanzitdinov, and Galina O. Tsareva

Peoples' Friendship University of Russia named after Patrice Lumumba (RUDN University), 6 Miklukho-Maklaya St, Moscow 117198, Russian Federation
vasilyev-sa@rudn.ru
http://www.rudn.ru

Abstract. 5G/6G networks are a next technological step in the field of telecommunications. 5G/6G networks provide the implementation of the required quality of communication with the growth of subscriber devices and lack of frequency bands. The application of queuing theory methods to analyze network performance is very important at the design, implementation and operation stages, as it is necessary to ensure a high return on investment that will be directed to the introduction of this new technology. Consequently, the attention of 5G/6G researchers is particularly focused on the analysis of the shortest queue problem which is widely used as balancing mechanisms in time-scale queueing system (TSQS). In this paper we employ simulation analysis of the TSQS evolution dynamics under the supposition that there are the large number of identical single-service devices and it is suppose this number increases indefinitely. It is assumed that all single-service devices have identical exponentially distributed service time with a finite mean value and a finite service intensity. It is supposed that there is a Poisson incoming stream of arriving requests with a finite intensity and TSQS fulfills a service discipline so that for each incoming request is provided a random selection a server device from random selected m-set server devices that has the s-th shortest queue size. The evolution of TSQS states can be represent by solutions of a system of differential equations of infinite degree.

We investigate the Cauchy problem for this singularly perturbed system with a small parameter. We use the mean-field approach and formulate the Cauchy problem for the truncated singularly perturbed finite order system of differential equations and the initial condition problem for the singularly perturbed nonlinear first order partial differential equation with a small parameter.

We construct an analytical solution of the initial condition problem for the singularly perturbed nonlinear first order partial differential equation and apply a high-order non-uniform grid scheme for numerical analysis

This paper has been supported by the RUDN University Strategic Academic Leadership Program.

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 188–201, 2024.
https://doi.org/10.1007/978-3-031-65385-8_14

of the solutions of the truncated singularly perturbed Cauchy problem. We use the numerical scheme with different sets which gives to evaluate the impact of a small parameter in time-scaling processes for TSQS. This grid scheme demonstrates good convergence of the solutions of the truncated singularly perturbed Cauchy problem when a small parameter tend to zero. The final outcome of our numerical simulations shows that this TSQS can support execution of the services with a high incoming flow of requests.

Keywords: Countable Markov chains · Time-scale network analysis · Singular perturbed systems of differential equations · Numerical analysis of the Cauchy problem · Layer-adapted piecewise uniform Shishkin-type meshes

1 Introduction

The research of time-scale queueing systems (TSQS) is very significant because of the application of 5G/6G networks to the telecommunications market requires to investigate not only the complicated analytical methods of the queuing theory [3,16] but apply modern numerical methods [1,6].

The researching of TSQS with a huge number of identical server devices and shortest queue disciplines [7,15] associated with the Caushy problems for infinite degree systems of ordinary differential equations.

The problem of the time-scaling TSQS especially attracts attention among the methods of TSQS analysis [4] since it is an effective technique for analyzing large-scale complex systems. If there is a purpose to investigate the scale in time invariance of TSQS, then there is an opportunity to study the transformation properties of solutions of differential equations which define the dynamics of the system changes over time. The time-scaling change is similarity transformations of the solutions of a system of differential equations with a time-scaling parameter and it form a group of time-scale transformations of TSQS. The time-scaling methods often identify with the use of a small time-scaling parameter in TSQS models [2,8].

The mean-field theory is a very important modeling tool in queue theory lately [12,13]. The mean-field theory is a mathematical apparatus used to model and analyze large-scale systems consisting of many interacting elements. Each element of the system interacts with the rest of the elements in this approach, but instead of taking into account all these interactions, they are averaged or simplified to a single average value, called the average mean-field. The mean-field theory allows to analyze complex systems in which interactions between elements play an important role. The main idea of the mean-field theory is that the interactions between the elements of the system can be averaged or approximated using an average value. This makes it easier to model and analyze the system, since it is necessary to take into account only one average field, and not all possible interactions. The theory of the mean field also allows us to study the special properties of the system, that is, properties that arise as a result of the interaction of elements and cannot be explained only by their individual characteristics. This allows you to understand how the system as a whole functions and what properties it exhibits.

In recent times, it is proposed layer-adapted methods for the numerical simulation analysis of singularly perturbed systems differential equations which are modifications of non-uniform mesh methods [5,9].

In the papers [10,11] we studied the shortest queue system model and showed that this system had evolution dynamics which was described by the solutions of the Cauchy problem of the singularly perturbed Tikhonov type system of infinite order differential equations. We used the truncation method for this system and applied a high-order non-uniform grid scheme for numerical analysis of the solutions of the truncated singularly perturbed Cauchy problem. This numerical scheme demonstrated good convergence of solutions of the singularly perturbed Cauchy problem when a small parameter tend to zero $\varepsilon \to 0$.

In this paper we use a high-order non-uniform grid scheme for the numerical simulation of the evolution dynamics of TSQS with n-identical single-service devices in the case $n \to \infty$. The TSQS evolution dynamics can be obtained as a solution $x_k^{s,m}(t)$ of a infinite degree system of differential equations. We study the singularly perturbed Cauchy problem for this system of differential equations with a small parameter. We apply the Dobrushin mean-field approach for this singularly perturbed Cauchy problem and formulate the truncated Cauchy problem for the finite order system of differential equations and the initial condition problem for the nonlinear first order partial differential equation. We use an analytical method and find the solution of the initial condition problem for the nonlinear first order partial differential equation. We apply a high-order non-uniform grid scheme for numerical analysis of the truncated Cauchy problem. We show that the grid scheme demonstrate good convergence of the solutions that describe TSQS evolution when a small parameter $\varepsilon \to 0$. The results of the numerical simulation illustrate that this TSQS can carry out the execution of services when there is a high incoming flow of requests.

2 The Shortest Time-Scale Queueing System with a Small Parameter

In the works [10,11] we study the shortest queue problem for TSQS with FCFS n-identical single-service devices with its own exponentially distributed service times. We admit that the value of TSQS average service time is $\bar{t} = 1/\mu$. Thus, the parameter $\mu > 0$ is a TSQS service intensity. We admit that there is the Poisson arrival process with the rate of requests $n\lambda > 0$ in this model. We assume that we can select m identical TSQS-devices immediately and randomly for each arrival request and then we can choice one single device among m selected devices immediately that has the s-th shortest ($(m-s)$-th longest) queue length in the choice moment for each arrival request. If there is more than one device with the s-th shortest queue size, the choice between them is made randomly and the request is sent to the chosen device after selection immediately.

We use the vector of the functions $\mathbf{x}^{s,m}(t) = \{x_k^{s,m}(t)\}_{k=0}^{\infty}$, where the functions $x_k^{s,m}(t)$ are the shares of the identical devices that have the queue lengths not less than the value k at the moment $t \geq 0$.

These functions $x_k^{s,m}(t)$ become deterministic when the infinite limit tend to infinity $n \to \infty$. Thus, we can find the functions $x_k^{s,m}(t)$ by solving the Cauchy problem for the infinite system of differential equations with small parameter $\varepsilon > 0$ in the such form [10,11]:

$$\begin{cases} \varepsilon^{\rho_k} \dot{x}_k^{s,m}(t) = \mu[x_{k+1}^{s,m}(t) - x_k^{s,m}(t)] - \lambda \Delta h_{s,m}(x_k^{s,m}(t)), \\ x_0^{s,m}(t) = 1,\, x_k^{s,m}(0) = x_k^0 \geq 0,\, x_k^0 \geq x_{k+1}^0,\, k \geq 1,\, t \geq 0, \end{cases} \tag{1}$$

where $\mathbf{x}^0 = \{x_k^0\}_{k=0}^{\infty}$ $(x_0^0 = 1, x_k^0 \geq x_{k+1}^0, x_k^0 \geq 0, x_k^0 \in \mathbf{R})$ is a non-increasing sequence and $\{\rho_k\}_{k=1}^{\infty}$ $(\rho_k \geq 0,\, \rho_k \in \mathbf{R})$ is a numerical sequence which we apply for the time transformation in the form $\bar{t}_k = \varepsilon^{-\rho_k} t$.

The function $\Delta h_{s,m}(x_k^{s,m}(t))$ has the form for $1 \leq s \leq m$ $(s, m \in \mathcal{N})$:

$$\Delta h_{s,m}(x_k^{s,m}(t)) = \left[h_{s,m}(x_k^{s,m}(t)) - h_{s,m}(x_{k-1}^{s,m}(t))\right],$$

$$h_{s,m}(x_k^{s,m}(t)) = \sum_{l=0}^{s-1} \sum_{p=0}^{l} \frac{(-1)^{l-p} m!}{p!(m-l)!(l-p)!} (x_k^{s,m}(t))^{m-p},$$

$$\Delta h_{s,m}(x_k^{s,m}(t)) = \sum_{l=0}^{s-1} \sum_{p=0}^{l} \frac{(-1)^{l-p} m!}{p!(m-l)!(l-p)!} [(x_k^{s,m}(t))^{m-p} - (x_{k-1}^{s,m}(t))^{m-p}].$$

The Cauchy problem (1) can be represented in this way:

$$\begin{cases} \dot{\mathbf{x}} + \mathbf{M}(\mathbf{x}(t), \mu, \varepsilon, \rho) = \mathbf{F}(\mathbf{x}(t), \lambda, \varepsilon, \mathbf{R}), \\ \mathbf{x}(0) = \mathbf{x}^0,\ t \geq 0, \end{cases} \tag{2}$$

$$\mathbf{x}(t) = (x_1^{s,m}(t), x_2^{s,m}(t), \ldots, x_n^{s,m}(t), \ldots),\ x_0^{s,m}(t) = 1,\ t \geq 0,$$

$$\mathbf{M} = (M_1, M_2, \ldots, M_n, \ldots),\ \mathbf{F} = (F_1, F_2, \ldots, F_n, \ldots),$$

$$M_k = \varepsilon^{-\rho_k} \mu \left(x_k^{s,m}(t) - x_{k+1}^{s,m}(t)\right),\ F_k = \varepsilon^{-\rho_k} \lambda \Delta h_{s,m}(x_k^{s,m}(t)),\ k \geq 1,$$

$$\mathbf{x}(0) = (x_1^{s,m}(0), \ldots, x_n^{s,m}(0), \ldots),\ \mathbf{x}^0 = (x_1^0, x_2^0, \ldots, x_n^0, \ldots),$$

$$x_k^0 \geq x_{k+1}^0,\ k \geq 1,\ \mathbf{R} = (\rho_1, \rho_2, \ldots, \rho_n, \ldots).$$

3 Dobrushin Mean-Field Approach and Time-Scale Queueing Systems Model with a Small Parameter

We apply the Dobrushin mean-field approach in this section. If assume that $\mathbf{R} = (\rho_1, \rho_2, \ldots, \rho_n, \rho_{n+1}, \rho_{n+1}, \ldots)$, $\rho_k = \rho_{n+1}$, $k > n+1$ for (2), we can write the n-order finite system of differential equations with n-finite conditions and use the Dobrushin mean-field approach for the next $n+1, n+2, \ldots$ equations in the form:

$$\begin{cases} \dot{\tilde{x}} + \tilde{\mathbf{M}}(\tilde{\mathbf{x}}(t), \mu, \varepsilon, \tilde{\mathbf{R}}) = \tilde{\mathbf{F}}(\tilde{\mathbf{x}}(t), \lambda, \varepsilon, \tilde{\mathbf{R}}), \\ \tilde{\mathbf{x}}(0) = \tilde{\mathbf{x}}^0, \\ \tilde{\mathbf{L}}_{n+1}(\mu, \varepsilon, \rho_{n+1}) x_{n+1}^{s,m}(t) = 0, \\ x_{n+1}^{s,m}(0, \eta) = x_{n+1}^0(\eta),\ x_n^0 \geq x_{n+1}^0(\eta), \\ t \geq 0, \eta \geq 0, \end{cases} \tag{3}$$

where

$$\tilde{\mathbf{x}}(t) = (x_1^{s,m}(t), x_2^{s,m}(t), \ldots, x_n^{s,m}(t)),\ x_0^{s,m}(t) = 1,\ t \geq 0,$$

$$\tilde{\mathbf{M}} = (M_1, M_2, \ldots, M_n),\ \tilde{\mathbf{F}} = (F_1, F_2, \ldots, F_n),$$

$$M_k = \varepsilon^{-\rho_k} \mu \left(x_k^{s,m}(t) - x_{k+1}^{s,m}(t) \right),\ F_k = \varepsilon^{-\rho_k} \lambda \mathit{\Delta} h_{s,m}(x_k^{s,m}(t)),\ k \in \overline{1, n},$$

$$\tilde{\mathbf{x}}(0) = (x_1^{s,m}(0), \ldots, x_n^{s,m}(0)),\ \tilde{\mathbf{x}}^0 = (x_1^0, x_2^0, \ldots, x_n^0),$$

$$x_k^0 \geq x_{k+1}^0,\ k \in \overline{1, n-1},\ \tilde{\mathbf{R}} = (\rho_1, \rho_2, \ldots, \rho_n, \rho_{n+1}),$$

$$\tilde{\mathbf{L}}_{n+1}(\mu, \varepsilon, \rho_{n+1}) x_{n+1}^{s,m}(t) = \varepsilon^{\rho_{n+1}} \frac{\partial x_{n+1}^{s,m}(t, \eta)}{\partial t} - \mu \frac{\partial x_{n+1}^{s,m}(t, \eta)}{\partial \eta} +$$

$$+ \lambda \sum_{l=0}^{s-1} \sum_{p=0}^{l} \frac{(-1)^{l-p} m!}{p!(m-l)!(l-p)!} \frac{\partial (x_{n+1}^{s,n}(t, \eta))^{m-p}}{\partial \eta} =$$

$$= \varepsilon^{\rho_{n+1}} \frac{\partial x_{n+1}^{s,m}(t, \eta)}{\partial t} + \mathit{\Phi}_{\lambda,\mu}^{s,m}(x_{n+1}^{s,m}(t, \eta), \omega_{n+1}^{s,m}(t, \eta)),$$

$$\mathit{\Phi}_{\lambda,\mu}^{s,m}(x_{n+1}^{s,m}(t, \eta), \omega_{n+1}^{s,m}(t, \eta)) = -\mu \frac{\partial x_{n+1}^{s,m}(t, \eta)}{\partial \eta} +$$

$$+ \lambda \sum_{l=0}^{s-1} \sum_{p=0}^{l} \frac{(-1)^{l-p} m!}{p!(m-l)!(l-p)!(m-p)} (x_{n+1}^{s,n}(t, \eta))^{m-p-1} \omega_{n+1}^{s,m}(t, \eta),$$

$$\omega_{n+1}^{s,m}(t, \eta) = \frac{\partial x_{n+1}^{s,m}(t, \eta)}{\partial \eta},$$

where a piecewise continuous function $x_{n+1}^0(\eta)$ $(x_n^0 \geq x_{n+1}^0(\eta), \eta \in [0, +\infty))$ is an initial condition for the nonlinear first order partial differential equation. This Cauchy problem (3) is the Tikhonov problem, if we assume that $\rho_k = 0$, $k = \overline{1, l}$, $\rho_k > 0$, $k = \overline{l+1, n+1}$ $(2 \leq l \leq n)$.

Thus, we have the system with two separate problems and we can consider the auxiliary initial condition problem for the nonlinear first order partial differential equation and find the solution $x_{n+1}^{s,m}$

$$\begin{cases} \tilde{\mathbf{L}}_{n+1}(\mu, \varepsilon, \rho_{n+1}) x_{n+1}^{s,m}(t, \eta) = 0, \\ x_{n+1}^{s,m}(0, \eta) = x_{n+1}^0(\eta),\ x_n^0 \geq x_{n+1}^0(\eta), \\ t \geq 0, \eta \geq 0. \end{cases} \tag{4}$$

It means that we can consider the problem of finding an integral surface S_0 of the Eq. (4) that passing through a given initial curve $P_0(\eta)$, $\eta \geq 0$ ($t = 0$, $x^{s,m}_{n+1}(0,\eta) = x^0_{n+1}(\eta)$). Thus, we can obtain a uniquely differentiable function $x^{s,m}_{n+1}(t,\eta)$ of the variables $t \geq 0, \eta \geq 0$ where this surface S_0 must be uniquely projected onto the plane $x^{s,m}_{n+1}(t,\eta) = 0$ of the variables $t \geq 0, \eta \geq 0$.

We can rewrite the problem (4) in the form of the initial condition problem for a system of two quasi-linear equations

$$\begin{cases} \varepsilon^{\rho_{n+1}} \frac{\partial x^{s,m}_{n+1}}{\partial t} + \Phi^{s,m}_{\lambda,\mu} = 0, \\ \varepsilon^{\rho_{n+1}} \frac{\partial \omega^{s,m}_{n+1}}{\partial t} + \frac{\partial \Phi^{s,m}_{\lambda,\mu}}{\partial \omega^{s,m}_{n+1}} \frac{\partial \omega^{s,m}_{n+1}}{\partial \eta} + \frac{\partial \Phi^{s,m}_{\lambda,\mu}}{\partial x^{s,m}_{n+1}} \omega^{s,m}_{n+1} + \frac{\partial \Phi^{s,m}_{\lambda,\mu}}{\partial \eta} = 0, \\ x^{s,m}_{n+1}(0,\eta) = x^0_{n+1}(\eta),\, x^0_n \geq x^0_{n+1}(\eta), \\ \omega^{s,m}_{n+1}(0,\eta) = \frac{\partial x^0_{n+1}(\eta)}{\partial \eta}, t \geq 0, \eta \geq 0, \end{cases} \tag{5}$$

where the solutions $x^{s,m}_{n+1}(t,\eta), \omega^{s,m}_{n+1}(t,\eta)$ $(t \geq 0, \eta \geq 0)$ have the form [14]:

$$\eta(t,\xi) = \xi - \varepsilon^{-\rho_{n+1}} t \left(\mu - \lambda \sum_{l=0}^{s-1} \sum_{p=0}^{l} \frac{(-1)^{l-p} m!(m-p)}{p!(m-l)!(l-p)!} (x^0_{n+1}(\xi))^{m-p-1} \right) \geq 0,$$

where a parameter $\xi \in [0, +\infty)$ is a non-negative real value.

We use the exponentially decreasing function $x^0_{n+1}(\eta) = a\, exp(-\sigma\eta) \leq x^0_n$ for our model, where we admit that $0 < a \leq x_n$, $\sigma > 0$ are positive parameters. The solution of the problem (5) has the form in this case:

$$\begin{cases} x^{s,m}_{n+1}(\eta(t,\xi)) = a\, e^{-\sigma\eta(t,\xi)}, \\ \omega^{s,m}_{n+1}(\eta(t,\xi)) = -a\sigma\, e^{-\sigma\eta(t,\xi)}, \\ \eta(t,\xi) = \xi - \varepsilon^{-\rho_{n+1}} t[\mu - \lambda\, a\, B_{s,m}(\sigma,\xi)] \geq 0, \\ \eta \geq 0,\, \xi \geq 0, \end{cases} \tag{6}$$

where

$$B_{s,m}(\sigma,\xi) = \sum_{l=0}^{s-1} \sum_{p=0}^{l} \frac{(-1)^{l-p} m!(m-p)}{p!(m-l)!(l-p)!} e^{-\sigma(m-p-1)\xi}.$$

Now we can apply this solution $x^{s,m}_{n+1}(t,\eta)$ of the problem (5) for finding the solution $\tilde{\mathbf{x}}(t)$ of the problem (3).

4 Numerical Analysis of Time-Scale Queueing Systems Model with a Small Parameter Using Dobrushin Mean-Field Approach

We apply a piecewise-uniform grid $\bar{\Xi}_\tau[0, T_0]$

$$t_0 = 0, t_i < t_{i+1},\, i = \overline{0, N-1},\, t_N = T_0$$

for numerical analysis of the problem (3)

$$\begin{cases} \bar{\Xi}_\tau[0,T_0] = (t_i | t_i = i\tau_1;\ i = \overline{0,K}; \\ t_i = t_K + i\tau;\ i = \overline{1,M}; \\ t_i = t_{K+M} + i\tau_2;\ i = \overline{1,2L}; \\ t_i = t_{K+M+2L} + i\tau;\ i = \overline{1,N-K-M-2L}), \\ \tau_1 = \delta_1/K,\ \tau_2 = \delta_2/2L,\ \delta_j = \bar{D}_j \varepsilon\, ln(\varepsilon^{-1})),\ j = 1,2; \\ \tau = (T_0 - \delta_1 - \delta_2)/(N-K-M-2L), \end{cases} \tag{7}$$

where parameters $\bar{D}_j$ $(j = 1,2)$ are determined analytically by asymptotic estimates of solutions of the Tikhonov problem (3) and $T_{IBL} = t_{K+M} + L\tau_2$ is a point where there is an inner boundary layer. Thus, this piecewise-uniform grid $\bar{\Xi}_\tau$ has K small steps τ_1, $2L$ small steps τ_2, and $(N-K-M-2L)$ big steps τ on the segment $[0,T_0]$.

We apply a finite-difference approximation of the problem (3) in the form:

$$\begin{cases} \tilde{\mathbf{x}}_{i+1} = \mathbf{B}_i, \\ i = \overline{1,N-1}, \\ \tilde{\mathbf{x}}_0 = \tilde{\mathbf{x}}^0, \end{cases} \tag{8}$$

$$h_i = t_i - t_{i-1},\ i = \overline{1,N-1},$$

$$x^{s,m}_{k,i} = x^{s,m}_k(t_i),\ x^{s,m}_{0,i} = 1,\ i = \overline{0,N},$$

$$\mathbf{B}_i = (B_{1i}, B_{2i}, \dots, B_{ni}),\ \mathbf{B}_i = \tilde{\mathbf{x}}_i + h_i[\tilde{\mathbf{F}}_i - \tilde{\mathbf{M}}_i],$$

$$F_{ki} = \varepsilon^{-\rho_k} \lambda \sum_{l=0}^{s-1} \sum_{p=0}^{l} \frac{(-1)^{l-p} m!}{p!(m-l)!(l-p)!} [(x^{s,m}_{k-1,i})^{m-p} - (x^{s,m}_{k,i})^{m-p}],$$

$$M_{ki} = \varepsilon^{-\rho_k} \mu \left(x^{s,m}_{k,i} - x^{s,m}_{k+1,i} \right), k \in \overline{1,n},\ i = \overline{0,N}$$

$$B_{ki} = x^{s,m}_{k,i} + h_i \varepsilon^{-\rho_k} [\mu \left(x^{s,m}_{k+1,i} - x^{s,m}_{k,i} \right) +$$

$$+ \lambda (h_{s,m}(x^{s,m}_{k-1,i}) - h_{s,m}(x^{s,m}_{k,i})],\ k = \overline{1,n},\ i = \overline{0,N},$$

$$\tilde{\mathbf{x}}_0 = (x^{s,m}_{1,0}, x^{s,m}_{2,0}, \dots, x^{s,m}_{n,0}), \tilde{\mathbf{x}}^0 = (x^0_1, x^0_2, \dots, x^0_n),\ x^0_k \geq x^0_{k+1},\ k \in \overline{1,n-1},$$

$$x^{s,m}_{n+1,i} = x^{s,m}_{n+1}(\eta(t_i,0)),\ i = \overline{0,N},$$

where we use the value of the parameter $\xi = 0$ for the solution $x^{s,m}_{n+1}(\eta(t,\xi))$.

We apply the the fourth-order Runge-Kutta method for problem (4):

$$\mathbf{g}^1_i = \mathbf{B}(\mathbf{x}_i, t_i),\ \mathbf{g}^2_i = \mathbf{B}(\mathbf{x}_i + \frac{h_i}{2}\mathbf{g}^1_i, t_i + h_i/2),$$

$$\mathbf{g}^3_i = \mathbf{B}(\mathbf{x}_i + \frac{h_i}{2}\mathbf{g}^2_i, t_i + h_i/2),\ \mathbf{g}^4_i = \mathbf{B}(\mathbf{x}_i + h_i \mathbf{g}^3_i, t_i + h_i),$$

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \frac{h_i}{6}(\mathbf{g}^1_i + 2\mathbf{g}^2_i + 2\mathbf{g}^3_i + \mathbf{g}^4_i),$$

where $\mathbf{g}^j_i \in R^n$ $(i = \overline{0,N},\ j = \overline{1,4})$ are vectors.

The result of the simulation analysis of the TSQS evolution $x_k^{s,m}(t)$, when $k = \overline{1,25}$ and $m = 2, 3, 4$, where $1 \leq s \leq m$), is showed in the figures (see Fig. 1-9). The parameters of the model such as the low incoming mode of arrival request rates λ_l, the high incoming mode of arrival request rates λ_h, the service intensity μ, the parameters a, σ, the number of steps of the grid N, the permissible error δ and so on are presented in the Table 1.

Table 1. Simulation parameters

Parameters	Values of the parameters
λ_l	3
λ_h	7
μ	5
n	25
l $(1 \leq k \leq 9)$	9
$\rho_k(1 \leq k \leq 9)$	0
$\rho_k(10 \leq k \leq 25)$	$1/k$
x_k^0	$(28-k)/30,\ k = \overline{1,25}$
a	x_{25}^0
s	1
N	10^4
δ	10^{-6}

The values of the parameters s, m, ε are showed in the captions under the figures (see Figs. 1, 2, 3, 4, 5, 6, 7, 8, and 9).

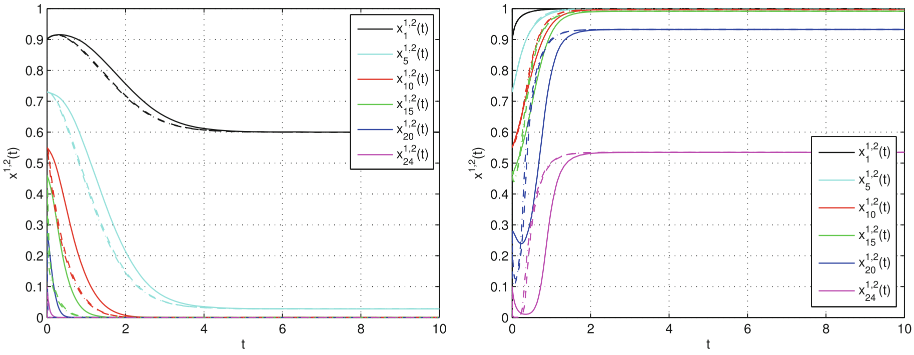


Fig. 1. The function $x_k^{1,2}(t)$ ($\varepsilon = 0.1$ solid line, $\varepsilon = 0.01$ long dash line, $\varepsilon = 0.001$ short dash line, $\lambda = 3$ for the left graph, $\lambda = 7$ for the right graph, $\mu = 5$).

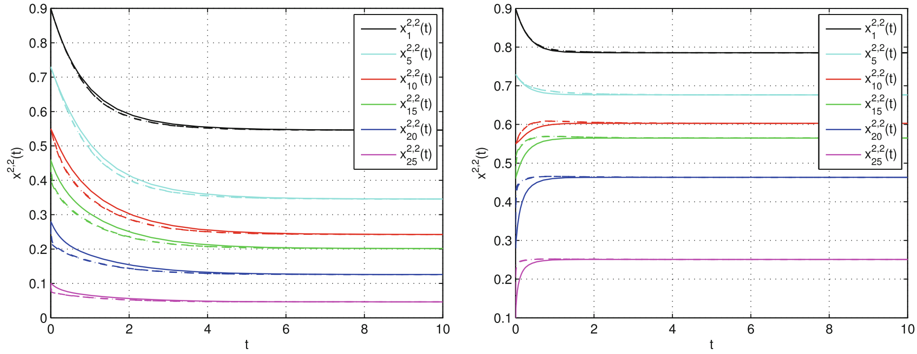


Fig. 2. The function $x_k^{2,2}(t)$ ($\varepsilon = 0.1$ solid line, $\varepsilon = 0.01$ long dash line, $\varepsilon = 0.001$ short dash line, $\lambda = 3$ for the left graph, $\lambda = 7$ for the right graph, $\mu = 5$).

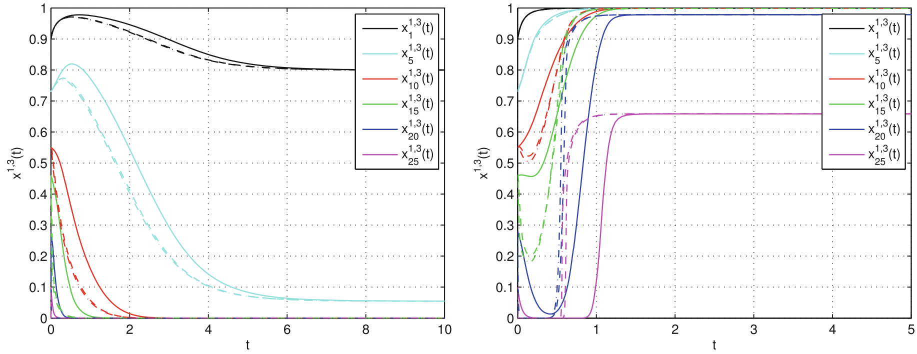


Fig. 3. The function $x_k^{1,3}(t)$ ($\varepsilon = 0.1$ solid line, $\varepsilon = 0.01$ long dash line, $\varepsilon = 0.001$ short dash line, $\lambda = 3$ for the left graph, $\lambda = 7$ for the right graph, $\mu = 5$).

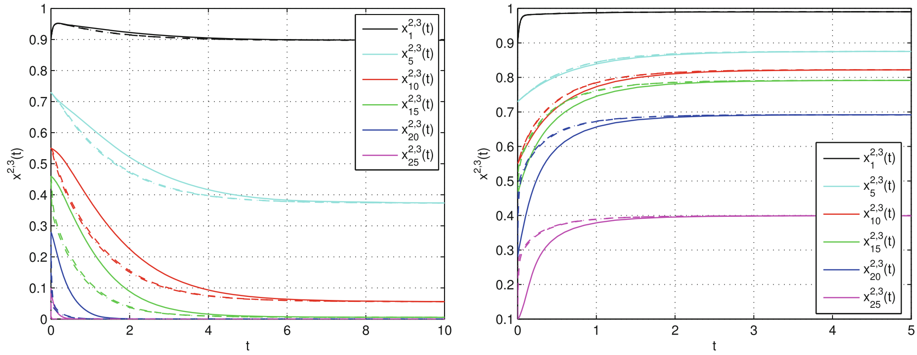


Fig. 4. The function $x_k^{2,3}(t)$ ($\varepsilon = 0.1$ solid line, $\varepsilon = 0.01$ long dash line, $\varepsilon = 0.001$ short dash line, $\lambda = 3$ for the left graph, $\lambda = 7$ for the right graph, $\mu = 5$).

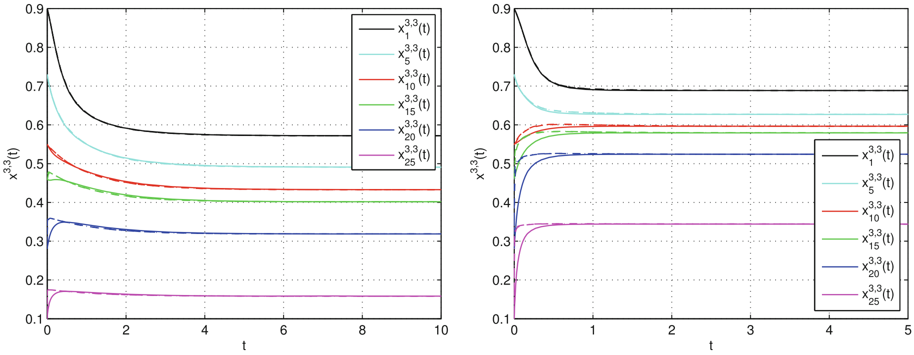


Fig. 5. The function $x_k^{3,3}(t)$ ($\varepsilon = 0.1$ solid line, $\varepsilon = 0.01$ long dash line, $\varepsilon = 0.001$ short dash line, $\lambda = 3$ for the left graph, $\lambda = 7$ for the right graph, $\mu = 5$).

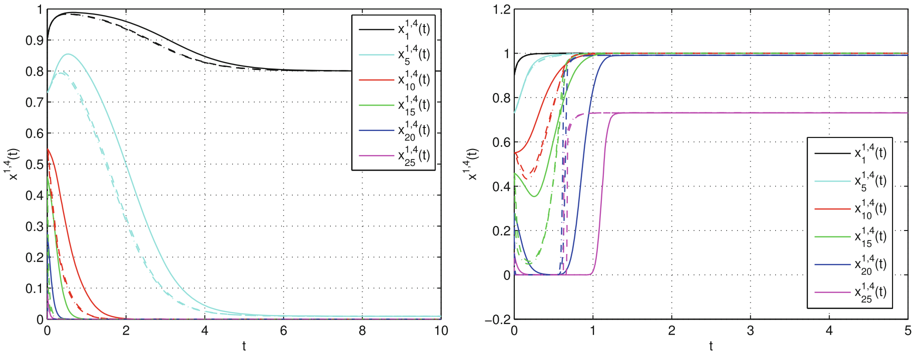


Fig. 6. The function $x_k^{1,4}(t)$ ($\varepsilon = 0.1$ solid line, $\varepsilon = 0.01$ long dash line, $\varepsilon = 0.001$ short dash line, $\lambda = 3$ for the left graph, $\lambda = 7$ for the right graph, $\mu = 5$).

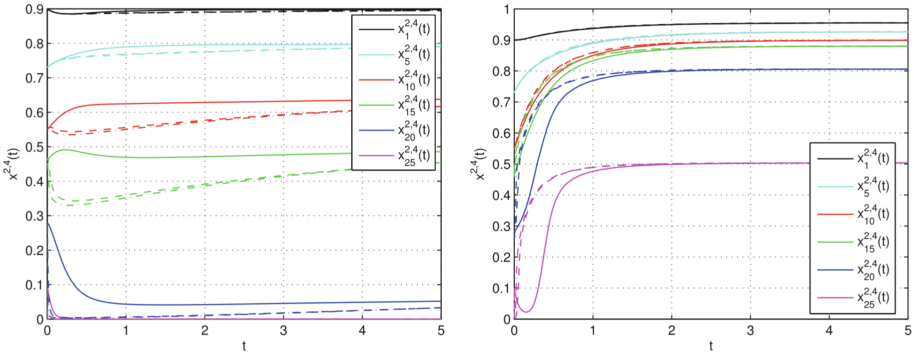


Fig. 7. The function $x_k^{2,4}(t)$ ($\varepsilon = 0.1$ solid line, $\varepsilon = 0.01$ long dash line, $\varepsilon = 0.001$ short dash line, $\lambda = 3$ for the left graph, $\lambda = 7$ for the right graph, $\mu = 5$).

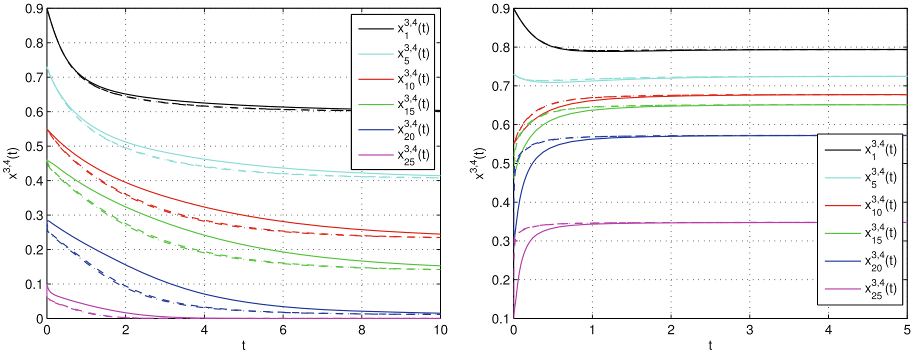


Fig. 8. The function $x_k^{3,4}(t)$ ($\varepsilon = 0.1$ solid line, $\varepsilon = 0.01$ long dash line, $\varepsilon = 0.001$ short dash line, $\lambda = 3$ for the left graph, $\lambda = 7$ for the right graph, $\mu = 5$).

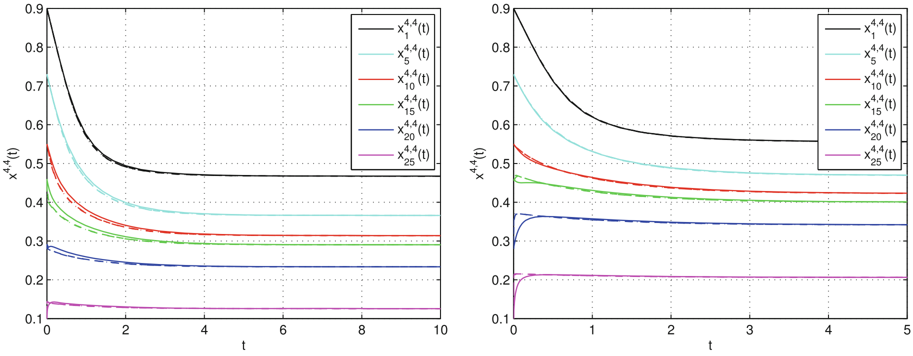


Fig. 9. The function $x_k^{4,4}(t)$ ($\varepsilon = 0.1$ solid line, $\varepsilon = 0.01$ long dash line, $\varepsilon = 0.001$ short dash line, $\lambda = 3$ for the left graph, $\lambda = 7$ for the right graph, $\mu = 5$).

In the Figs. 1, 2, 3, 4, 5, 6, 7, 8, and 9 we demonstrate that the behaviour of the TSQS evolution solution $x_k^{s,m}(t)$ ($m = 2, 3, 4$, $1 \leq s \leq m$) has an unstable service conditions when there are the low incoming mode of arrival request rates $\lambda_l = 3$ and the high incoming mode of arrival request rates $\lambda_h = 7$. We can see the left and inner transition layers. The transitions become more sharper when $\varepsilon \rightarrow 0$.

In the Fig. 1, 3, 6 we can see that the TSQS evolution of the solution $x_k^{1,2}(t)$, $x_k^{1,3}(t)$, $x_k^{1,4}(t)$. The TSQS evolution when $t \rightarrow 10$ has a lower stable service mode when there is the low incoming mode of arrival request rates $\lambda_l = 3$ and it has a higher stable service mode when there is the high incoming mode of arrival request rates $\lambda_h = 7$.

In the Fig. 4, Fig. 5 we can see the results of the numerical stimulation of the TSQS evolution of the solutions $x_k^{2,3}(t)$, $x_k^{3,3}(t)$. It is shown that TSQS has an stable service mode when $t \rightarrow 10$ and when there are the overload conditions similar to the case the Fig. 2. Thus, an increase in the parameter s leads to the stably service. There are only the left transition layers. The left transitions become more sharper when $\varepsilon \to 0$.

In the Fig. 7, 8, and 9 we can see the results of the numerical stimulation of the TSQS evolution of the solutions $x_k^{2,4}(t)$, $x_k^{3,4}(t)$, $x_k^{4,4}(t)$. It is shown that TSQS has an stable service mode when $t \to 10$ and when there are the overload conditions similar to the case the Fig. 2, 4, 5. Thus, an increase in the parameter s leads to the stably service. There are only the left transition layers. The left transitions become more sharper when $\varepsilon \to 0$.

5 Conclusion

Modern 5G/6G telecommunications have stepped far ahead and these technologies are developing in the areas of research of wired and wireless networks, artificial intelligence, the Internet of Things (IoT), the creation of smart cities. Strategic planning in 5G/6G telecommunications, which is the basis for development and financial success, is actively used not only at the level of individual companies, but also entire countries. It is necessary to predict the innovative changes for successful development of the 5G/6G telecommunications technology and for strategic planning during long-term period. The implementation of 5G/6G technology requires special attention of developers of new products in the field of wireless communication, as these technologies will be able to provide higher data transfer speeds, shorter latency, as well as greater energy efficiency compared to 4G technologies currently used. Developers will inevitably face a number of technical problems when designing the 5G/6G architecture, which must cope with a more complex multi-user environment and the use of channels at higher frequencies. The modern studies of theoretical TSQS models makes it possible to understand the problems which will be solved for the implementation of 5G/6G technology.

In this paper we show how the Dobrushin mean-field approach and the numerical methods may be applied for analysis of the evolution of TSQS dynamics. For researching of the evolution dynamics TSQS we study the function $x_k^{s,m}(t)$, which are the shares of the identical devices that have the queue lengths not less than the value k at the moment $t \geq 0$. We use a high-order non-uniform grid scheme of the Shishkin-type for numerical solving of the Cauchy problem for the system of differential equations finite degree using the solution of the initial problem for the nonlinear first order partial differential equation. This equation also describes shock waves in gas dynamics. Thus, we can find general methods for analyzing the behavior of systems in which there is abrupt changes in the solutions of equations that describe such systems. The presence of a small parameter makes it possible to analyze the behavior of solutions in relation to time transformation and to study the problem of time scaling.

This paper has been supported by the RUDN University Strategic Academic Leadership Program (recipient S.A. Vasilyev, mathematical model development, simulation model development, numerical analysis).

References

1. Baddour, A., Malykh, M., Sevastianov, L.: On periodic approximate solutions of dynamical systems with quadratic right-hand side. J. Math. Sci. **261**, 698–708 (2022). https://doi.org/10.1007/s10958-022-05781-4
2. Bushkova, T., Moiseeva, S., Moiseev, A., Sztrik, J., Lisovskaya, E., Pankratova, E.: Using infinite-server resource queue with splitting of requests for modeling two-channel data transmission. Methodol. Comput. Appl. Probab. **24**, 1753–1772 (2022). https://doi.org/10.1007/s11009-021-09890-6
3. Dudin, A., Dudina, O., Dudin, S., Gaidamaka, Y.: Self-Service System with Rating Dependent Arrivals, Mathematics. **10**(3), 297 (2022). https://doi.org/10.3390/math10030297
4. Hu, K., Wang, B., Cao, S., Li, W., Wang, L.: A novel model predictive control strategy for multi-time scale optimal scheduling of integrated energy system. Energy Reports **8**, 7420–7433 (2022). https://doi.org/10.1016/j.egyr.2022.05.184
5. Kaushik, A., Choudhary, M.: A higher-order uniformly convergent defect correction method for singularly perturbed convection-diffusion problems on an adaptive mesh. Alexandria Eng. J. **61**(12), 9911–9920 (2022). https://doi.org/10.1016/j.aej.2022.03.005
6. Kondratyeva, A., et al.: Characterization of dynamic blockage probability in industrial millimeter wave 5g deployments. Future Internet **14**(7), 193 (2022). https://doi.org/10.3390/fi14070193
7. Liu, X., Gong, K., Ying, L.: Steady-state analysis of load balancing with Coxian-2 distributed service times, Naval Res. Logist. **69**(1), 57–75 (2022). https://doi.org/10.1002/nav.21986
8. Nazarov, A., Dudin, A., Moiseev, A.: Pseudo steady-state period in non-stationary infinite-server queue with state dependent arrival intensity. Mathematics. **10**(15), 2661 (2022). https://doi.org/10.3390/math10152661
9. Roul, P.: A fourth-order non-uniform mesh optimal B-spline collocation method for solving a strongly nonlinear singular boundary value problem describing electrohydrodynamic flow of a fluid. Appl. Num. Math. **153**, 558–574 (2020). https://doi.org/10.1016/j.apnum.2020.03.018
10. Silhavy, R., Silhavy, P., Prokopova, Z. (eds.): Data Science and Algorithms in Systems: Proceedings of 6th Computational Methods in Systems and Software 2022, Vol. 2. Springer International Publishing, Cham (2023)
11. Dudin, A., Nazarov, A., Moiseev, A. (eds.): Information Technologies and Mathematical Modelling. Queueing Theory and Applications: 21st International Conference, ITMM 2022, Karshi, Uzbekistan, October 25–29, 2022, Revised Selected Papers. Springer Nature Switzerland, Cham (2023)
12. Vvedenskaya N.D., Dobrushin R.L., Kharpelevich F.I.: Queueing system with a choice of the lesser of two queues - the asymptotic approach. Probl. inform. **32**(1), 15–27 (1996). https://www.mathnet.ru/eng/ppi298
13. Zhu, Z., Ke, J., Wang, H.: A mean-field Markov decision process model for spatial-temporal subsidies in ride-sourcing markets. Transp. Res. Part B: Methodol. **150**, 540–565 (2021). https://doi.org/10.1016/j.trb.2021.06.014
14. Zaytsev, V.F., Polyanin, A.D.: Spravochnik po differencialnym uravneniyam s chastnymi proizvodnymi pervogo poryadka. Izdatel'stvo Fiziko-matematicheskoj literatury (FIZ.-MAT.LIT), Moskva (2003). https://search.rsl.ru/ru/record/01001836334

15. Zhou, X., Shroff, N., Wierman, A.: Asymptotically optimal load balancing in large-scale heterogeneous systems with multiple dispatchers. Perform. Eval. **145**, 102146 (2021). https://doi.org/10.1016/j.peva.2020.102146
16. Zisgen, H.: An approximation of general multi-server queues with bulk arrivals and batch service. Oper. Res. Lett. **50**(1), 57–63 (2022). https://doi.org/10.1016/j.orl.2021.12.006

Research of Adaptive RQ System M/M/1 with Unreliable Server

N. M. Voronina[1(✉)] and S. V. Rozhkova[1,2]

[1] National Research Tomsk Polytechnic University, Lenin Avenue 30, Tomsk, Russia
{vnm,rozhkova}@tpu.ru
[2] National Research Tomsk State University, Lenin Avenue 36, Tomsk, Russia

Abstract. The paper considers a single-line retrial queueing (RQ) system with an unreliable server controlled by a adaptive random multiple access protocol. The study is carried out using the method of asymptotic analysis under conditions of heavy system load. In this paper, the main characteristics of the system were found.

Keywords: Retrial queue · Adaptive random multiple access protocol · Unreliable server

1 Introduction

Unreliable servers can be used in telecommunications, call centers and data networks. For example, faulty hardware or software can lead to network failures, loss of communications, interruptions in data transmission, or incorrect processing of information. Research into unreliable devices in such areas helps to identify the causes of failures and develop methods to prevent them.

RQ queuing system with unreliable server and adaptive random multiple access protocol is a system that combines elements of queuing, unreliable server and adaptive random multiple access protocols in a data network.

In such systems, a large number of customers or clients are served using unreliable devices that may be subject to failures or malfunctions. To ensure efficiency and reliability, such a system can be equipped with an adaptive random multiple access protocol, which provides mechanisms for optimizing the use of available resources and managing data transmission in the face of varying network load, interference, or frequent server failures.

Adaptive Random Multiple Access Protocol is a method of controlling access to a common data link that allows multiple devices to share available resources. It is a form of multiple access protocol that allows devices to compete for access to data communications.

Unlike static protocols, an adaptive random multiple access protocol can change its parameters depending on current network conditions such as load, collisions and delays. This allows you to optimize the use of the available data channel and significantly reduce the likelihood of collisions (a situation in which

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 202–212, 2024.
https://doi.org/10.1007/978-3-031-65385-8_15

two or more devices try to transmit data at the same time, resulting in signal loss).

Many scientific works are devoted to the study of various models of data transmission networks and random multiple access protocols. There are various modifications of access protocols [1–8].

In papers [9–12], authors investigate models with adaptive access protocols. The papers [13–20] consider the study of queuing systems with a dynamic access protocol.

In this paper, we study a single-channel RQ system with an unreliable device controlled by an adaptive access protocol. The server is considered unreliable if it periodically fails and requires time to be restored. Which, accordingly, can lead to a decrease in the efficiency of the system and an increase in waiting time for service.

2 Description of the Mathematical Model

Let's consider an RQ system with an adaptive access protocol, the input of which receives a simple flow of requests with parameter λ. The time for servicing a customer by the server is distributed exponentially with the parameter μ_1. We assume that the server is unreliable. An unreliable device can be in one of the following states: idle, busy or under repair. When a new customer arrives and the server is idle, then the servicing immediately begins.

If at this moment another customer arrives, and the device is busy, then the received customer goes into orbit and waits for the opportunity to occupy the device during the next attempt. After a random delay, a request with intensity $\sigma = 1/T(t)$ again tries to occupy the server for service, where $T(t)$ is the state of the adapter at time t (see Fig. 1).

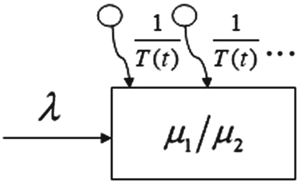


Fig. 1. Model of adaptive retrial queueing system M/M/1 with unreliable server

The working time is distributed exponentially with parameter γ_1, if server is idle and with parameter γ_2, if the server is busy. As soon as a breakdown occurs, the server is sent to repair and the servicing customer goes into the orbit. During repairing, all incoming customers go into the orbit. The recovery time is distributed exponentially with parameter μ_2.

The goal of the research is to study such a system, as well as to determine its main characteristics and to find the throughput of the system and the stationary probability distribution of server states.

Let $i(t)$ be the number of customers in the orbit at time t and $k\left(t\right)$ determine the state of the server as follows:

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

The process of changing adapter states $T(t)$ is defined as follows:

$$T(t+\Delta t)=\begin{cases}T(t)-\alpha\Delta t,\text{if } k(t)=0,\\T(t+\Delta t)=T(t),\text{if } k(t)=1,\\T(t)+\beta\Delta t,\text{if } k(t)=2,\end{cases}$$

where $\alpha>0,\ \beta>0$ are adapter parameters, the values of which are indicated.

If the device is idle, then $T(t)$ decreases linearly with intensity α; if the device is busy, then $T(t)$ does not change; if the device is under repair, then $T(t)$ increases linearly with intensity β.

3 The Method of Asymptotic Analysis

Let us denote

$$P\left(k,\ i,\ T,\ t\right)=\frac{\partial P\{k(t)=k,\ i(t)=i,\ T(t)<T\}}{\partial T}$$

- the probability that at time t the server is in state k and i customers in the orbit.

The probability distribution $P\left(k,\ i,\ T,\ t\right)$ satisfies the following system of equations:

$$\begin{cases}P\left(0,i,T-\alpha\Delta t,\ t+\Delta t\right)=\left(1-\lambda\Delta t\right)\left(1-\frac{i}{T}\Delta t\right)\left(1-\gamma_1\Delta t\right)P\left(0,i,\ T,t\right)\\+\mu_1\Delta tP\left(1,i,T,t\right)+\mu_2\Delta tP\left(2,i,T,t\right)+o\left(\Delta t\right),\\P\left(1,i,T,\ t+\Delta t\right)=\left(1-\lambda\Delta t\right)\left(1-\mu_1\Delta t\right)\left(1-\gamma_2\Delta t\right)P_1\left(1,i,T,\ t\right)\\+\lambda\Delta tP\left(0,i,T,\ t\right)+\frac{i+1}{T}\Delta tP\left(0,i+1,T,t\right)+\lambda P\left(1,i-1,T,t\right)+o\left(\Delta t\right),\\P\left(2,i,\ T+\beta\Delta t,t+\Delta t\right)=\left(1-\lambda\Delta t\right)\left(1-\mu_2\Delta t\right)P\left(2,i,T,t\right)\\+\gamma_1\Delta tP\left(0,i,T,t\right)+\gamma_2\Delta tP\left(1,i-1,T,t\right)+\lambda\Delta tP\left(2,i-1,T,t\right)+o\left(\Delta t\right).\end{cases}$$

Let us compose a system of Kolmogorov differential equations for the stationary probability distribution $P\left(k,\ i,\ T\right)$:

$$\begin{cases} -\alpha\dfrac{\partial P\left(0,i,\ T\right)}{\partial T} = -\left(\lambda+\dfrac{i}{T}+\gamma_1\right)P\left(0,i,T\right)+\mu_1 P\left(1,i,T\right) \\ +\mu_2 P\left(2,i,T\right), \\ \dfrac{\partial P\left(1,i,\ T\right)}{\partial T} = -\left(\lambda+\mu_1+\gamma_2\right)P\left(1,i,\ T\right)+\lambda P\left(0,i,T\right) \\ +\dfrac{i+1}{T}P\left(0,i+1,T\right)+\lambda P\left(1,i-1,T\right), \\ \beta\dfrac{\partial P\left(2,i,\ T\right)}{\partial T} = -\left(\lambda+\mu_2\right)P\left(2,i,T\right)+\gamma_1 P\left(0,i,T\right) \\ +\gamma_2 P\left(1,i-1,T\right)+\lambda P\left(2,i-1,T\right). \end{cases} \tag{1}$$

Let us denote the partial characteristic functions

$$\begin{aligned} H_k(u_1,\ u_2) &= \sum_i e^{-u_1 i}\int\limits_0^\infty e^{-u_2 T}P(k,i,T)dT \\ &= P\{k(t)=k\}M\{e^{-u_1 i(t)-u_2 T(t)}\,|k(t)=k\}. \end{aligned} \tag{2}$$

Functions $H_k(u_1,\ u_2)$ have the following properties:

$$\sum_i e^{-u_1 i}\int\limits_0^\infty e^{-u_2 T}iP(k,i,T)dT = -\frac{\partial H_k(u_1,u_2)}{\partial u_1},$$

$$\sum_i e^{-u_1 i}\int\limits_0^\infty e^{-u_2 T}\frac{\partial P(k,i,T)}{\partial T}dT = u_2 H_k(u_1,u_2),$$

$$\sum_i e^{-u_1 i}\int\limits_0^\infty e^{-u_2 T}\frac{1}{T}P(k,i,T)dT = \int\limits_{u_2}^\infty H_k(u_1,x)dx.$$

Using the Eq. (2) and the properties of characteristic functions from System (1) we obtain:

$$\begin{cases} \left(\alpha u_2-\lambda-\gamma_1\right)H_0\left(u_1,u_2\right)+\displaystyle\int\limits_{u_2}^\infty \frac{\partial H_0(u_1,x)}{\partial u_1}dx+\mu_1 H_1\left(u_1,u_2\right) \\ +\mu_2 H_2\left(u_1,u_2\right)=0, \\ -\left(\lambda(1-e^{-u_1})+\mu_1+\gamma_2+u_2\right)H_1\left(u_1,u_2\right)+\lambda H_0\left(u_1,u_2\right) \\ -e^{u_1}\displaystyle\int\limits_{u_2}^\infty \frac{\partial H_0(u_1,x)}{\partial u_1}dx=0, \\ -\left(\lambda(1-e^{-u_1})+\mu_2+\beta u_2\right)H_2(u_1,u_2)+\gamma_1 H_0\left(u_1,u_2\right)+ \\ \gamma_2 e^{-u_1}H_1\left(u_1,u_2\right)=0. \end{cases}$$

We introduce a parameter

$$\rho = \frac{\lambda}{\mu_1},$$

that characterizes the system load, then we get

$$\begin{cases} \left(\dfrac{\alpha u_2 - \gamma_1}{\mu_1} - \rho\right) H_0\left(u_1, u_2\right) + \dfrac{1}{\mu_1} \int\limits_{u_2}^{\infty} \dfrac{\partial H_0(u_1, x)}{\partial u_1} dx + H_1\left(u_1, u_2\right) \\ + \dfrac{\mu_2}{\mu_1} H_2\left(u_1, u_2\right) = 0, \\ -\left(\rho(1 - e^{-u_1}) + 1 + \dfrac{\gamma_2 + u_2}{\mu_1}\right) H_1\left(u_1, u_2\right) + \rho H_0\left(u_1, u_2\right) \\ - \dfrac{e^{u_1}}{\mu_1} \int\limits_{u_2}^{\infty} \dfrac{\partial H_0(u_1, x)}{\partial u_1} dx = 0, \\ -\left(\rho(1 - e^{-u_1}) + \dfrac{\mu_2 + \beta u_2}{\mu_1}\right) H_2(u_1, u_2) + \dfrac{\gamma_1}{\mu_1} H_0\left(u_1, u_2\right) \\ + \dfrac{\gamma_2}{\mu_1} e^{-u_1} H_1\left(u_1, u_2\right) = 0. \end{cases} \tag{3}$$

There are no exact analytical methods for solving the System (3), so we will find the main characteristics of the adaptive system.

Let's study the System (3) under the condition of heavy load. Let us define the throughput S of an adaptive RQ system as the exact upper bound of system load values ρ for which there is a steady-state regime with $\rho \uparrow S$.

Let us denote

$$\varepsilon = S - \rho.$$

Assuming that $\varepsilon \to 0$ in the System (2), we will perform the following substitutions:

$$\rho = S - \varepsilon,\ u_1 = \varepsilon w_1,\ u_2 = \varepsilon w_2,\ H_k(u_1, u_2) = F_k(w_1, w_2, \varepsilon).$$

Then, we obtain:

$$\begin{cases} F_0\left(w_1, w_2, \varepsilon\right) \left(\dfrac{\alpha \varepsilon w_2 - \gamma_1}{\mu_1} - (S - \varepsilon)\right) + \dfrac{1}{\mu_1} \int\limits_{u_2}^{\infty} \dfrac{\partial F_0(w_1, x, \varepsilon)}{\partial w_1} dx \\ + F_1\left(w_1, w_2, \varepsilon\right) + \dfrac{\mu_2}{\mu_1} F_2\left(w_1, w_2, \varepsilon\right) = 0, \\ F_0\left(w_1, w_2, \varepsilon\right) (S - \varepsilon) - \dfrac{e^{\varepsilon w_1}}{\mu_1} \int\limits_{w_2}^{\infty} \dfrac{\partial F_0(w_1, x, \varepsilon)}{\partial w_1} dx \\ + F_1\left(w_1, w_2, \varepsilon\right) \left((S - \varepsilon)(e^{-\varepsilon w_1} - 1) - 1 - \dfrac{\gamma_2 + \varepsilon w_2}{\mu_1}\right) = 0, \\ F_0\left(w_1, w_2, \varepsilon\right) \dfrac{\gamma_1}{\mu_1} + F_1\left(w_1, w_2, \varepsilon\right) \dfrac{\gamma_2}{\mu_1} e^{-\varepsilon w_1} \\ + F_2(w_1, w_2, \varepsilon)\left((S - \varepsilon)(e^{-\varepsilon w_1} - 1) - \left(\dfrac{\mu_2 + \beta \varepsilon w_2}{\mu_1}\right)\right) = 0. \end{cases} \tag{4}$$

Theorem 1. *The values of the parameters S and y in the adaptive RQ-system are determined by the equalities*

$$S = \frac{\alpha\mu_2 - \beta\gamma_1}{(1-\beta)\gamma_1 + (\alpha+1)\mu_2 + \gamma_2(\alpha+\beta)},$$

$$y = \frac{(\alpha\mu_2 - \beta\gamma_1)((\alpha+\beta)\gamma_2^2 + ((\alpha+1)\mu_2 + \alpha\mu_1 - \beta\gamma_1 + \gamma_1)\gamma_2)}{(\gamma_2(\alpha+\beta) + (\alpha+1)\mu_2 + \gamma_1(1-\beta))(\beta\gamma_2 + \mu_2)} + \frac{\mu_1(\alpha\mu_2 + \gamma_1(1-\beta))}{(\gamma_2(\alpha+\beta) + (\alpha+1)\mu_2 + \gamma_1(1-\beta))(\beta\gamma_2 + \mu_2)}.$$

where $\alpha > 0,\ \beta > 0$ are adapter parameters, the values of which are indicated.

Proof. Let us denote $\lim\limits_{\varepsilon\to 0} F_k(w_1, w_2, \varepsilon) = F_k(w_1, w_2)$ and for $\varepsilon \to 0$, we get

$$\begin{cases} -F_0\left(w_1, w_2\right)\left(\dfrac{\gamma_1}{\mu_1} + S\right) + \dfrac{1}{\mu_1}\displaystyle\int_{w_2}^{\infty} \dfrac{\partial F_0(w_1, x)}{\partial w_1} dx + F_1\left(w_1, w_2\right) \\ + \dfrac{\mu_2}{\mu_1} F_2\left(w_1, w_2\right) = 0, \\ F_0\left(w_1, w_2\right) S - \dfrac{1}{\mu_1}\displaystyle\int_{w_2}^{\infty} \dfrac{\partial F_0(w_1, x)}{\partial w_1} dx - F_1\left(w_1, w_2\right)(1 + \dfrac{\gamma_2}{\mu_1}) = 0, \\ F_0\left(w_1, w_2\right)\dfrac{\gamma_1}{\mu_1} + F_1\left(w_1, w_2\right)\dfrac{\gamma_2}{\mu_1} - F_2(w_1, w_2)\dfrac{\mu_2}{\mu_1} = 0. \end{cases} \tag{5}$$

We will look for the solution $F_k(w_1, w_2)$ of the System (5) in the form:

$$F_k(w_1, w_2) = F_k\Phi(w_1, w_2) = R_k(S, y)\varphi(w_2 + w_1 y). \tag{6}$$

Assuming that the function $\varphi(w)$ is equal to zero at infinity, we obtain

$$\int_{w_2}^{\infty} \frac{\partial F_0(w_1, x)}{\partial w_1} dx = -yR_0(S, y)\varphi(w_2 + yw_1).$$

Then we rewrite the System (5):

$$\begin{cases} -R_0\left(S, y\right)\left(\dfrac{\gamma_1}{\mu_1} + S\right) - \dfrac{y}{\mu_1} R_0\left(S, y\right) + R_1\left(S, y\right) + \dfrac{\mu_2}{\mu_1} R_2\left(S, y\right) = 0, \\ R_0\left(S, y\right) S + \dfrac{y}{\mu_1} R_0\left(S, y\right) - R_1\left(S, y\right)(1 + \dfrac{\gamma_2}{\mu_1}) = 0, \\ R_0\left(S, y\right)\dfrac{\gamma_1}{\mu_1} + R_1\left(S, y\right)\dfrac{\gamma_2}{\mu_1} - R_2(S, y)\dfrac{\mu_2}{\mu_1} = 0. \end{cases} \tag{7}$$

Let us add a normalization condition to the System (7):

$$R_0 + R_1 + R_2 = 1.$$

We obtain:

$$\begin{cases} -R_0\left(S,y\right)\left(\dfrac{\gamma_1}{\mu_1}+S\right)-\dfrac{y}{\mu_1}R_0\left(S,y\right)+R_1\left(S,y\right)+\dfrac{\mu_2}{\mu_1}R_2\left(S,y\right)=0, \\ R_0\left(S,y\right)S+\dfrac{y}{\mu_1}R_0\left(S,y\right)-R_1\left(S,y\right)(1+\dfrac{\gamma_2}{\mu_1})=0, \\ R_0\left(S,y\right)\dfrac{\gamma_1}{\mu_1}+R_1\left(S,y\right)\dfrac{\gamma_2}{\mu_1}-R_2(S,y)\dfrac{\mu_2}{\mu_1}=0. \\ R_0+R_1+R_2=0. \end{cases} \tag{8}$$

Then from the System (8) we find expressions for the stationary distribution of server states:

$$R_0=\frac{\mu_2(\gamma_2+\mu_1)}{(S\mu_1+\gamma_1+\mu_2+y)\gamma_2+(S\mu_2+\mu_2+\gamma_1)\mu_1+\mu_2 y},$$

$$R_1=\frac{\mu_2(S\mu_1+y)}{(S\mu_1+\gamma_1+\mu_2+y)\gamma_2+(S\mu_2+\mu_2+\gamma_1)\mu_1+\mu_2 y},$$

$$R_2=\frac{(S\mu_1+\gamma_1+y)\gamma_2+\gamma_1\mu_1}{(S\mu_1+\gamma_1+\mu_2+y)\gamma_2+(S\mu_2+\mu_2+\gamma_1)\mu_1+\mu_2 y}.$$

To find the values S and y, we sum up all the equations of the System (4) and for $\varepsilon\to 0$, we obtain

$$\begin{aligned} &F_0\left(w_1,w_2,\varepsilon\right)\left(\frac{\alpha\varepsilon w_2}{\mu_1}\right)-\frac{e^{\varepsilon w_1}-1}{\mu_1}\int\limits_{u_2}^{\infty}\frac{\partial F_0(w_1,x,\varepsilon)}{\partial w_1}dx \\ &+F_1\left(w_1,w_2,\varepsilon\right)((S-\varepsilon)(e^{-\varepsilon w_1}-1)+\frac{\gamma_2}{\mu_1}(e^{-\varepsilon w_1}-1)-\frac{\varepsilon w_2}{\mu_1}) \\ &+F_2\left(w_1,w_2,\varepsilon\right)((S-\varepsilon)(e^{-\varepsilon w_1}-1)-\frac{\beta\varepsilon w_2}{\mu_1})=0. \end{aligned}$$

Dividing resulting equation by ε, we get:

$$\begin{aligned} &F_0\left(w_1,w_2,\varepsilon\right)\left(\frac{\alpha w_2}{\mu_1}\right)-\frac{e^{\varepsilon w_1}-1}{\mu_1\varepsilon}\int\limits_{u_2}^{\infty}\frac{\partial F_0(w_1,x,\varepsilon)}{\partial w_1}dx \\ &+F_1\left(w_1,w_2,\varepsilon\right)((S-\varepsilon)\frac{(e^{-\varepsilon w_1}-1)}{\varepsilon}+\frac{\gamma_2}{\mu_1}\frac{(e^{-\varepsilon w_1}-1)}{\varepsilon}-\frac{w_2}{\mu_1}) \\ &+F_2\left(w_1,w_2,\varepsilon\right)((S-\varepsilon)\frac{(e^{-\varepsilon w_1}-1)}{\varepsilon}-\frac{\beta w_2}{\mu_1})=0. \end{aligned}$$

Then using the Taylor expansion, we obtain:

$$F_0\left(w_1, w_2, \varepsilon\right) \frac{\alpha w_2}{\mu_1} - \frac{w_1}{\mu_1} \int\limits_{u_2}^{\infty} \frac{\partial F_0(w_1, x, \varepsilon)}{\partial w_1} dx$$

$$- F_1\left(w_1, w_2, \varepsilon\right)(w_1(S + \frac{\gamma_2}{\mu_1} + \frac{w_2}{\mu_1}) - F_2\left(w_1, w_2, \varepsilon\right)(S w_1 + \frac{\beta w_2}{\mu_1}) = 0.$$

Applying (6) to the equation, we get:

$$\begin{aligned} &\alpha w_2 \varphi(w_2 + w_1 y) R_0(S, y) + y w_1 \varphi(w_2 + w_1 y) R_0(S, y) \\ &- \varphi(w_2 + w_1 y) R_1(S, y)(\mu_1 w_1 S + w_1 \gamma_2 + w_2) \\ &- \varphi(w_2 + w_1 y) R_2(S, y)(\mu_1 w_1 S + \beta w_2) = 0. \end{aligned}$$

Let us write the equation in the form:

$$\begin{aligned} &\alpha w_2 R_0(S, y) + y w_1 R_0(S, y) - R_1(S, y)(\mu_1 w_1 S + w_1 \gamma_2 + w_2) \\ &- R_2(S, y)(\mu_1 w_1 S + \beta w_2) = 0. \end{aligned} \qquad (9)$$

Then we rewrite the Eq. (9):

$$\begin{aligned} &w_1(y R_0(S, y) - \mu_1 S R_1(S, y) - \gamma_2 R_1(S, y) - \mu_1 S R_2(S, y)) \\ &+ w_2(\alpha R_0(S, y) - R_1(S, y) - \beta R_2(S, y)) = 0. \end{aligned}$$

In order to turn the equation into an identity in w_1 and w_2, it is enough to require the following equalities:

$$\begin{cases} y R_0(S, y) - \mu_1 S R_1(S, y) - \gamma_2 R_1(S, y) - \mu_1 S R_2(S, y) = 0, \\ \alpha R_0(S, y) - R_1(S, y) - \beta R_2(S, y) = 0. \end{cases} \qquad (10)$$

By substituting $R_0(S, y)$, $R_1(S, y)$, $R_2(S, y)$ into the System (10), we get:

$$S = \frac{\alpha \mu_2 - \beta \gamma_1}{(1 - \beta)\gamma_1 + (\alpha + 1)\mu_2 + \gamma_2(\alpha + \beta)},$$

$$\begin{aligned} y = &\frac{(\alpha \mu_2 - \beta \gamma_1)((\alpha + \beta)\gamma_2^2 + ((\alpha + 1)\mu_2 + \alpha \mu_1 - \beta \gamma_1 + \gamma_1)\gamma_2)}{(\gamma_2(\alpha + \beta) + (\alpha + 1)\mu_2 + \gamma_1(1 - \beta))(\beta \gamma_2 + \mu_2)} \\ &+ \frac{\mu_1(\alpha \mu_2 + \gamma_1(1 - \beta))}{(\gamma_2(\alpha + \beta) + (\alpha + 1)\mu_2 + \gamma_1(1 - \beta))(\beta \gamma_2 + \mu_2)}. \end{aligned}$$

Definition. Throughput S is the upper limit of those load values $\rho = \frac{\lambda}{\mu_1}$, for which there is the steady-state regime.

The inequality

$$\frac{\lambda}{\mu_1} \leq S$$

determines the condition for the existence of a steady-state regime for the considered adaptive system.

So Theorem 1 is proved.

4 Numerical Example

We consider a system with parameters:

$$\mu_1 = 5, \quad \mu_2 = 2, \quad \gamma_1 = 0.03, \quad \gamma_2 = 0.03, \quad \lambda = 1, \quad \beta = 1.$$

Table 1 shows the values of S and y for a given system for different α.

Table 1. Values of S and y for different α

α	0,2	0,4	0,8	1	5	10	100
S	0,036	0,155	0,314	0,370	0,703	0,779	0,860
y	0,049	0,376	1,426	2,070	18,838	41,502	455,880

According to the data in Table 1, we can conclude that as α increases, the value of throughput S increases, and the value of y also increases significantly.

For adaptive RQ systems, under the limiting condition of heavy system load, random processes $i(t)$ and $T(t)$ are linearly dependent with some parameter y equal to the ratio of linearly dependent random processes $i(t)/T(t)$.

When $\alpha = 0,9$, the throughput of the adaptive RQ system $S = 0,334$ and $y = 2$, which corresponds to the throughput of the dynamic RQ system M/M/1 $S = 0,334$ at $y = 2$, which confirms the asymptotic equivalence of the adaptive and dynamic RQ systems with the simplest incoming flow of customers.

Consequently, adaptive RQ systems are asymptotically, under heavy load conditions, equivalent to dynamic RQ systems with the specified value of the parameter y, calculated in the work [21].

5 Conclusion

In this paper, we study the adaptive RQ-system M/M/1 with an unreliable server. The study was carried out using the method of asymptotic analysis under conditions of heavy system load. As a result, the main characteristics of the system, the stationary distribution of server states, and the throughput of the system under consideration are found.

References

1. Lyubina, T.V., Nazarov, A.A.: Research of dynamic and adaptive RQ-systems with an incoming MMPP flow of applications. Bull. Tomsk State Univ. Manage. Comput. Technol. Inform. **3**(24), 104–112 (2013)
2. Ephremides, A.: Analysis of protocols of multiple access. In: Modelling and Performance Evaluation Methodology, pp. 563–575. Springer, Heidelberg (1984)

3. Nazarov, A.A., Tsoy, S.A.: A general approach to the analysis of Markov models of data communication networks controlled by static random multiple access protocols. Autom. Control. Comput. Sci. **38**(4), 64–75 (2004)
4. Nazarov, A.A., Nikitina, M.A.: Application of Markov chain ergodicity conditions to the study of the existence of stationary regimes in communication networks. Autom. Control. Comput. Sci. **37**(1), 50–55 (2003)
5. Gao, J., Rubin, I.: Analysis of a random-access protocol under long-range-dependent traffic. IEEE Trans. Veh. Technol. **52**(3), 693–700 (2003)
6. Nazarov, A.A., Sudyko, E.A.: Method of asymptotic semiinvariants for studying a mathematical model of a random access network. Probl. Inf. Transm. **46**(1), 86–102 (2010)
7. Nazarov, A.A.: Research of computer communication networks with random multiple access protocols. Bull. Tomsk State Univ. No. **271**, 72–73 (2000)
8. Kuznetsov, D.Y., Nazarov, A.A.: Analysis of a communication network governed by an adaptive random multiple access protocol in critical load. Probl. Inf. Transm. **40**(3), 243–253 (2004)
9. Kuznetsov, D.Y., Nazarov, A.A.: Investigation of communication networks with adaptive random multiple access protocol under heavy load conditions for an infinite number of nodes. Autom. Control. Comput. Sci. **37**(3), 47–55 (2003)
10. Nazarov, A.A., Odyshev, Y.D.: Analysis of a communication network with the adaptive ALOHA protocol for a finite number of stations under overload. Problemy Peredachi Informatsii **36**(3), 83–93 (2000)
11. Anyugu Francis Lin, B., Ye, X., Hao, S.: Adaptive protocol for full-duplex two-way systems with the buffer-aided relaying. IET Commun. **13**(1), 54–58 (2019)
12. Lyubina, T.V., Nazarov, A.A.: Study of an adaptive RQ system with an incoming MMPP flow of customers using the method of asymptotic analysis. Information technologies and mathematical modeling (ITMM-2012): materials of the XI All-Russian scientific and practical conference with international participation (Anzhero-Sudzhensk, 23–24 November, 2012). Kemerovo: Practice, Part 2, pp. 94–99 (2013)
13. Shokhor, S.L.: Distribution of the number of messages in a communication network with channel reservation and a dynamic access protocol. Bull. Tomsk State Univ. **271**, 67–69 (2000)
14. Lyubina, T.V., Nazarov, A.A.: Investigation of a Markovian dynamic RQ-system with conflicts of requests. Bulletin of the Tomsk State University. Management, computer technology and informatics **3**(12), 73–84 (2010)
15. Nazarov A. A., Shokhor S. L.: Comparison of an asymptotic and prelimit model of a communication network with a dynamic protocol of random multiple access. In: Mathematical Modeling and Probability Theory, ed. I.A. Alexandrova and others, Tomsk: Peleng, pp. 233–241 (1988)
16. Gupur, G., Li, X.Z., Zhu, G.T.: The Application of C 0 -semigroup Theory to Dynamic Queueing Systems. Semigroup Forum 62, pp. 205–216 (2001). https://doi.org/10.1007/s002330010030
17. Ouyang H., Nelson B.L.: Simulation-based predictive analytics for dynamic queueing systems. 2017 Winter Simulation Conference (WSC), IEEE, pp. 1716–1727 (2017)
18. Mounce, R.: Existence of equilibrium in a continuous dynamic queueing model for traffic networks with responsive signal control. Transp. Traffic Theory 2009: Golden Jubilee, Springer, Boston, MA, pp. 327–344 (2009)
19. Filipiak, J.: Dynamic routing in a queueing system with a multiple service facility. Oper. Res. **32**(5), 1163–1180 (1984)

20. Nazarov, A.A., Odyshev, Y.D.: Investigation of communications networks with dynamic synchronous ALOHA protocol under heavy load conditions. Avtomat. Vychisl. Tekhn. **35**(1), 77–84 (2001)
21. Voronina, N.M., Rozhkova, S.V., Fedorova, E.A.: Research of Dynamic RQ System M/M/1 with unreliable server. In: Dudin, A., Nazarov, A., Moiseev, A. (eds.) Information Technologies and Mathematical Modelling. Queueing Theory and Applications. ITMM 2022. Communications in Computer and Information Science, vol. 1803, pp. 212–224. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-32990-6_18

On Information in Competing Risks Model of Random Censoring

A. A. Abdushukurov[1(✉)], N. S. Nurmukhamedova[2], and S. Erisbaev[3]

[1] Branch of Moscow State University named after M.V. Lomonosov in Tashkent, Tashkent, Uzbekistan
a_abdushukurov@rambler.ru
[2] National University of Uzbekistan named after Mirzo Ulugbek, Tashkent, Uzbekistan
[3] Karakalpak State University named after Berdakh, Nukus, Uzbekistan

Abstract. Authors consider a Fisher Information in Competing Risks Model of random censoring when distribution function of all risks depends on the same parameter θ. In this paper, they propose the decomposition formulas of information of model also Cramer-Rao type inequalities for unbiased estimators and some of its improvement. At the end of the paper, there are some useful recommendations based on use of decomposition formulas.

Keywords: Fisher information · Distribution family · Hazard rate · Censorship · Competing risks · Incomplete observation

1 Introduction

R.A. Fisher first proposed a concept of information in statistics in 1925 [1]. Now it is known in literature as a Fisher information of family of distribution $\{F_\theta,\ \theta \in \Theta \subseteq R\}$, depending on parameter θ (for simplifying) and density function of $f(x;\theta)$. It is denoted as integral

$$I(\theta) = E_\theta\left(\frac{\partial \log f(\xi;\theta)}{\partial\theta}\right)^2 = \int\limits_{-\infty}^{+\infty}\left(\frac{\partial \log f(x;\theta)}{\partial\theta}\right)^2 f(x;\theta)dx, \tag{1}$$

ξ is a random variable with regular density $f(x;\theta)$. For more properties of information (1) we can recommend monography [2] of Zaks Sh. Efron and Jonstone [3] gave another decomposition of information (1) in terms of the hazard rate function $\mu(x;\theta) = f(x;\theta)/(1 - F(x;\theta))$, where $F(x;\theta)$ is a distribution function corresponding for density $f(x;\theta)$ as

$$I_\xi(\theta) = E_\theta\left(\frac{\partial \log \mu(\xi;\theta)}{\partial\theta}\right)^2,\ \theta \in \Theta. \tag{2}$$

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 213–227, 2024.
https://doi.org/10.1007/978-3-031-65385-8_16

In this paper we generalize a concept of Fisher information to Competing Risks Model (CRM). CRM one can meet in queueing theory when consider reliability of series system. A series system is a configuration such that, if any one of the system components fails, the entire system fails. In papers [4–15] authors are considering several statistical problems of efficiency of functioning of server systems. They are calculating Fisher information or Cramer-Rao lowerbound of queueing model under certain regularity conditions.

2 Competing Risks Model

We can generalize these conceptions in case when dealing with CRM [4-6] of incomplete observation including random censorship. For a fixed natural number k and $i = 1, ..., k$, let $\{X^{(1)}, X^{(2)}, ..., X^{(k)}\}$ aggregation of independent random variables having a continuous distribution functions $\{F^{(1)}(x_1;\theta), F^{(2)}(x_2;\theta), ..., F^{(k)}(x_k;\theta)\}$ correspondingly and depending on common parameter θ, $\theta \in \Theta \subseteq R$, $x = (x_1, ..., x_k) \in R^k$. We are interested in observing random variables $X = \min\left(X^{(1)}, ...X^{(k)}\right)$ and indicators $\delta^{(i)} = I\left(X = X^{(i)}\right)$, $i = \overline{1,k}$, $\delta^{(1)} + ... + \delta^{(k)} = 1$. Let's denote functions

$$f^{(i)}(t;\theta) = \frac{\partial F^{(i)}(t;\theta)}{\partial t},\ \lambda^{(i)}(t;\theta) = \frac{f^{(i)}(t;\theta)}{\overline{F^{(i)}(t;\theta)}},\ \overline{F^{(i)}(t;\theta)} = 1 - F^{(i)}(t;\theta),\ i = 1, ..., k.$$

For example, in reliability theory, one can consider a physical system of k component sub-systems connected in series. The system fails when one of the sub-systems fail. Then the failure time X of system coincides with the failure time of its first component.

Thus, in CRM we are interested not only in observing vector $\left(X, \delta^{(1)}, ..., \delta^{(k)}\right)$ but also in pairs $\left(X, \delta^{(i)}\right)$, $i = 1, ..., k$. Consider the sub-distributions $H^{(i)}(t;\theta) = P_\theta(X \leq t, \delta^{(i)} = 1)$, $i = \overline{1,k}$, where

$$H(x;\theta) = H^{(1)}(x;\theta) + ... + H^{(k)}(x;\theta) = \int\limits_{-\infty}^{x} h(t;\theta)dt = P_\theta(X \leq x)$$

with $h(x;\theta) = h^{(1)}(x;\theta) + ... + h^{(k)}(x;\theta)$ and sub-density

$$h^{(i)}(x;\theta) = f^{(i)}(x;\theta) \prod_{l=1 l\neq i}^{k} \left(1 - F^{(l)}(x;\theta)\right),\ i = 1, ..., k.$$

Introduce regression functions

$$m^{(i)}(x;\theta) = P_\theta(\delta^{(i)} = 1/X = x) = E_\theta(\delta^{(i)}/X = x),\ i = 1, ..., k;\ (x;\theta) \in R \times \Theta,$$

where $m^{(1)}(x;\theta) + ... + m^{(k)}(x;\theta) = 1$. It is not difficult to verify that

$$H^{(i)}(t;\theta) = \int\limits_{-\infty}^{t} m^{(i)}(u;\theta)dH(u;\theta),\ \ i = 1, ..., k.$$

Let us denote hazard rate function $\lambda(x;\theta)$ of distribution function $H(x;\theta)$, then it is easy to verify that

$$\lambda^{(i)}(x;\theta) = m^{(i)}(x;\theta)\lambda(x;\theta), \quad i = 1, ..., k.$$

Consider regularity conditions of sub-densities as:

(I) The sets $N^{(i)} = \left\{t : h^{(i)}(t;\theta) > 0\right\}$, $i = \overline{1,k}$ do not depend on the parameter θ and $\bigcap\limits_{i=1}^{k} N^{(i)} \neq \emptyset$.

(II) The derivatives $\frac{\partial^m h^{(i)}(x;\theta)}{\partial\theta^m}$, $m = 1, 2;$ $i = 1, ..., k$ exist for all $\theta \in \Theta$.

(III) $\int\limits_{-\infty}^{+\infty} \left|\frac{\partial^m h^{(i)}(x;\theta)}{\partial\theta^m}\right| dx$, $m = 1, 2;$ $i = 1, ..., k$.

Let $I_{(X,\delta^{(1)},...,\delta^{(k)})}(\theta)$, $I_X(\theta)$ and $I^{(i)}_{(X,\delta^{(i)})}(\theta)$ $i = \overline{1,k}$ be Fisher information corresponding to the random vector $\left(X, \delta^{(1)}, ..., \delta^{(k)}\right)$, variable X and the pairs $\left(X, \delta^{(i)}\right)$, $i = 1, ..., k$.

$$I_{(X,\delta^{(1)},...,\delta^{(k)})}(\theta) = E_\theta\left(\frac{\partial \log h\left(X, \delta^{(1)}, ..., \delta^{(k)}\right)}{\partial\theta}\right)^2, \theta \in \Theta,$$

$$I_X(\theta) = E_\theta\left(\frac{\partial \log h\left(X;\theta\right)}{\partial\theta}\right)^2, \theta \in \Theta,$$

$$I^{(i)}_{(\delta^{(i)}/X)}(\theta) = \int\limits_{-\infty}^{+\infty} \left(\frac{\partial \log m^{(i)}(t;\theta))}{\partial\theta}\right)^2 m^{(i)}(t;\theta)dH(t;\theta), \quad i = 1, ..., k.$$

In papers [21,22] we prove the following result.

Theorem 1. *[21] Suppose conditions (I)–(III). Then for any $\theta \in \Theta$.*

$$(A) \quad I_{(X,\delta^{(1)},...,\delta^{(k)})}(\theta) = \sum_{i=1}^{k} I^{(i)}_{(X,\delta^{(i)})}(\theta). \tag{3}$$

$$(B) \quad I_{(X,\delta^{(1)},...,\delta^{(k)})}(\theta) = I_X(\theta) + \sum_{i=1}^{k} I^{(i)}_{(\delta^{(i)}/X)}(\theta). \tag{4}$$

In the papers [21,22] we established the Cramer-Rao type inequality in CRM. Note that in case of simple random censoring situation analogous results were established in paper of authors [16] and used in [17–20].

Theorem 2. *[22] Suppose that conditions of theorem 1 hold and for differentiable parameter function $\varphi(\theta)$ there exist an unbiased estimator $\widehat{\varphi}\left(X, \delta^{(1)}, ..., \delta^{(k)}\right)$ and the differentiation with respect to θ under integral sign*

$E_\theta \widehat{\varphi} = \varphi(\theta)$ is permissible for all $\theta \in \Theta$. Let $\varphi'(\theta)$ denote the derivative of $\varphi(\theta)$ with respect to θ. Then

$$Var_\theta \left\{ \widehat{\varphi} \right\} \geq \frac{\left(\varphi'(\theta)\right)^2}{I_{(X, \delta^{(1)}, \dots, \delta^{(k)})}(\theta)} \ \textit{for any} \ \theta \in \Theta. \tag{5}$$

An improved version of result (5) will be derived by application of the Walker's [23] inequality. Walker [23] obtained an improved version of the Caushy-Schwarts inequality.

Theorem 3. *[23] If ξ and η are random variables defined on a probability space $(\mathbf{X}, F, P)$ with finite second moments, then*

$$\left[E(\xi\eta)\right]^2 \leq E\xi^2 E\eta^2 - \left[|E\xi| \sqrt{Var\eta} - |E\eta| \sqrt{Var\xi} \right]^2 \tag{6}$$

In particular, if random variable η has mean zero but positive variance, then from (6) it follows that

$$\left[E(\xi\eta)\right]^2 \leq E\xi^2 E\eta^2 - |E\xi|^2 E\eta^2 = Var\xi E\eta^2$$

Hence

$$E\xi^2 \geq \left[E\xi\right]^2 + \frac{|E\xi\eta|^2}{E\eta^2} \tag{7}$$

Then we have

Theorem 4. *Under the conditions of theorem 2 the following inequality holds*

$$E_\theta \left[\widehat{\varphi}\left(X, \delta^{(1)}, ..., \delta^{(k)}\right) - \varphi(\theta) \right]^2 \geq \left\{ E_\theta \left[\widehat{\varphi}\left(X, \delta^{(1)}, ..., \delta^{(k)}\right) - \varphi(\theta) \right] \right\}^2$$

$$+ \frac{\left(\varphi'(\theta)\right)^2}{I_{(X, \delta^{(1)}, \dots, \delta^{(k)})}(\theta)} \ \textit{for any} \ \theta \in \Theta. \tag{8}$$

Proof. In inequality (7) we put

$$\xi = \widehat{\varphi}\left(X, \delta^{(1)}, ..., \delta^{(k)}\right) - \varphi(\theta), \quad \eta = \sum_{i=1}^{k} \delta^{(i)} \frac{\partial \log h^{(i)}(X; \theta)}{\partial \theta}.$$

Then for any $\theta \in \Theta$

$$E_\theta \eta = \sum_{i=1}^{k} E_\theta \left[\delta^{(i)} \frac{\partial \log h^{(i)}(X; \theta)}{\partial \theta} \right] = \sum_{i=1}^{k} \int_{-\infty}^{+\infty} \frac{\partial \log h^{(i)}(x; \theta)}{\partial \theta} h^{(i)}(x; \theta) dx =$$

$$\frac{\partial}{\partial \theta} \sum_{i=1}^{k} H^{(i)}(+\infty; \theta) dx = \frac{\partial}{\partial \theta} H^{(i)}(+\infty; \theta) = \frac{\partial}{\partial \theta}(1) = 0, \tag{9}$$

and

$$E_\theta\left[\xi\eta\right] = E_\theta\left[\widehat{\varphi}\left(X, \delta^{(1)}, ..., \delta^{(k)}\right) - \varphi(\theta)\right]\left[\sum_{i=1}^{k} \delta^{(i)} \frac{\partial \log h^{(i)}(X;\theta)}{\partial\theta}\right]$$

$$= \sum_{i=1}^{k} E_\theta\left[\widehat{\varphi}\left(X, \delta^{(1)}, ..., \delta^{(k)}\right) \frac{\partial \log h^{(i)}(X;\theta)}{\partial\theta}\right]$$

$$= \frac{\partial}{\partial\theta} \sum_{i=1}^{k} \int_{-\infty}^{+\infty} \widehat{\varphi}\left(x, y^{(1)}, ..., y^{(k)}\right) h^{(i)}(x;\theta)dx$$

$$= \frac{\partial}{\partial\theta} \int_{-\infty}^{+\infty} \widehat{\varphi}\left(x, y^{(1)}, ..., y^{(k)}\right) dH(x;\theta) = \varphi'(\theta). \tag{10}$$

Now (8) follows from (8)–(10). This completes the proof. Let's denote the probabilities

$$p^{(i)}(\theta) = P_\theta\left(\delta^{(i)} = 1\right) = P_\theta\left(X = X^{(i)}\right), \quad i = 1, ..., k,$$

where $p^{(1)}(\theta) + p^{(2)}(\theta) + ... + p^{(k)}(\theta) = 1, \ \theta \in \Theta$.

Definition. CRM follows a proportional hazards model (PHM) if we have the presentations for all $x \in R$

$$1 - F^{(i)}(x;\theta) = \left(1 - H(x;\theta)\right)^{p^{(i)}(\theta)}, \quad i = 1, ..., k. \tag{11}$$

Note that from (11) follows proportionality of hazard rate functions

$$\lambda^{(i)}(x;\theta) = p^{(i)}(\theta)\lambda(x;\theta), \quad i = 1, ..., k,$$

where $\lambda(x;\theta) = h(x;\theta)/\left(1 - H(x;\theta)\right)$ [5,6].

In PHM from (4) we can obtain corollary that

$$I_{(X,\delta^{(1)},...,\delta^{(k)})}(\theta) = I_X(\theta) + \sum_{i=1}^{k}\left(\frac{\partial \log p^{(i)}(\theta)}{\partial\theta}\right)^2 p^{(i)}(\theta), \quad \theta \in \Theta.$$

But, in particular, if all probabilities $p^{(i)}(\theta)$ are constants depending on parameter θ, $\left(p^{(i)}(\theta) = \beta^{(i)}, \ i = \overline{1,k}\right)$, i.e.

$$\frac{\partial \log p^{(i)}(\theta)}{\partial\theta} = 0, \quad i = 1, ..., k,$$

then the information on vector $\left(X, \delta^{(1)}, ..., \delta^{(k)}\right)$ coincides with information only on minima $X = \min\left(X^{(1)}, ..., X^{(k)}\right)$:

$$I_{(X,\delta^{(1)},...,\delta^{(k)})}(\theta) = \ I_X(\theta), \quad \theta \in \Theta.$$

At the end of paper we consider random censorship, when $k = 2$,

$$1 - F^{(1)}(x;\theta) = e^{-x/\theta},\ 1 - F^{(2)}(x;\theta) = e^{-x/\lambda} = \Big(1 - F^{(1)}(x;\theta)\Big)^{\theta/\lambda}$$

$$= \Big(1 - F^{(1)}(x;\theta)\Big)^{\beta(\theta)},\quad x \geq 0,\ \theta, \lambda > 0.$$

Here

$$1 - F^{(1)}(x;\theta) = (1 - H(x;\theta))^{p^{(1)}(\theta)},$$

$$1 - F^{(2)}(x;\theta) = (1 - H(x;\theta))^{p^{(2)}(\theta)},$$

$$\beta(\theta) = \frac{\theta}{\lambda},\ p^{(1)}(\theta) = \frac{1}{1+\beta(\theta)},\ p^{(2)}(\theta) = \frac{\beta(\theta)}{1+\beta(\theta)}.$$

Thus, distribution function $F^{(2)}$ does functionally not depend on θ, but depends through on indicators $\delta^{(1)}$ and $\delta^{(2)}$. Consequently, full information on triplet $\big(X, \delta^{(1)}, \delta^{(2)}\big)$ is equal

$$I_{(X,\delta^{(1)},...,\delta^{(k)})}(\theta) = I_X(\theta) + p^{(1)}(\theta)\bigg(\frac{\partial \log p^{(1)}(\theta)}{\partial\theta}\bigg)^2 + p^{(2)}(\theta)\bigg(\frac{\partial \log p^{(2)}(\theta)}{\partial\theta}\bigg)^2.$$

Now, if $\beta(\theta) = \beta$ -const (not depending on θ), then

$$\frac{\partial \log p^{(1)}(\theta)}{\partial\theta} = \frac{\partial \log p^{(2)}(\theta)}{\partial\theta} = 0,\ \ \theta \in \Theta$$

and information of triplet $\big(X, \delta^{(1)}, \delta^{(2)}\big)$ is contained only in minimal $X = \min\big(X^{(1)}, X^{(2)}\big)$.

Remark. Note that authors [24] give some numerical calculations of Fisher information in uninformative censoring case when distribution function of censors are not dependent on parameter θ and compared with Fisher information of complete sample.

3 Gibrid Censoring in CRM

Let's consider $\{X_1, ..., X_m\}$ where $X_j = \min\Big(X_j^{(1)}, ..., X_j^{(k)}\Big)$, $j = 1, ..., n$, is the sample of size n on independent observations of $X_j^{(i)}$, $i = 1, ..., k$. Consider ordered variables $X_{1n} < X_{2n} < ... < X_{nn}$ of sample $\{X_1, ..., X_m\}$. Introduce a sample of indicators $\Big\{\delta_j^{(1)}, ..., \delta_j^{(k)},\ j = 1, ..., n\Big\}$ corresponding to order statistics $\{X_{jn},\ j = 1, ..., n\}$. Let's consider vector $\mathbf{Y}^{(m)} = (Y_{1n}, ..., Y_{mn})$, $m = 1, ..., n$ with $Y_{jn} = (X_{jn}, \delta_{jn}^{(1)}, ..., \delta_{jn}^{(k)})$. Suppose that all considered random variables are positive and consider a fixed numbers $T_1, T_2 \in (0; +\infty)$, $T_1 < T_2$. Consider a lifetime experiment, which for a numbers $s < r < n$ stopped at time $T_{sn} = \min\{\max\{X_{\tau n}, T_1\}, T_2\}$.

Let's say the j-th object (or individual) with possible competing survival times $\left\{X_j^{(1)}, ..., X_j^{(k)}\right\}$ fails after time T_1 up to time $\min\{X_{\tau n}, T_2\}$ if s-th individual fails up to time T_2.Then the possible situations are following:

$$(I) \quad 0 < X_{sn} < X_{\tau n} < T_1 < T_2;$$

$$(II) \quad 0 < X_{sn} < T_1 < X_{\tau n} < T_2;$$

$$(III) \quad 0 < X_{sn} < T_1 < T_2 < X_{\tau n};$$

$$(IV) \quad 0 < T_1 < X_{sn} < X_{\tau n} < T_2;$$

$$(V) \quad 0 < T_1 < X_{sn} < T_2 < X_{\tau n};$$

$$(VI) \quad 0 < T_1 < T_2 < X_{sn} < X_{\tau n};$$

By considering these situations we know that experiment can be stopped at times $T_1, X_{\tau n}, T_2, X_{\tau n}, T_2$ and X_{sn}. Then the situations (I), (III) and (V) corresponds to type I censoring and others to the type II censoring. Hence, we deal with gibrid censoring situation where stopping time T_{sn} is Markovian stopping time. Above six situations may be described by pair (T_{sn}, τ_n) where

$$(T_{sn}, \tau_n) = \begin{cases} (T_1, \nu_1), \text{ for situation } (I), \\ (X_{rn}, r), \text{ for situation } (II) \text{ and } (IV), \\ (T_2, \nu_2), \text{ for situation } (III) \text{ and } (V), \\ (X_{sn}, s), \text{ for situation } (VI). \end{cases}$$

Here τ_n is common number of failed objects, v_1 and $v_2$are number of objects failed at times T_1 and T_2. Then the joint density function of vector (T_{sn}, τ_n) is

$$p_n(\mathbf{Y}^{(\tau)}; \theta) = \frac{n!}{(n-\tau)!} \prod_{l=1}^{\tau_n} \prod_{i=1}^{k} \{[h^{(i)}(X_{ln}; \theta)]^{\delta_{ln}^{(i)}}\}[1 - H(T_{\tau_n}; \theta)]^{n-\tau_n}. \qquad (12)$$

Let's introduce sums

$$S_{1n} = \sum_{l=1}^{\tau_n - 1} \sum_{i=1}^{k} \delta_{l\,n}^{(i)} \frac{\partial \log h^{(i)}(X_{l\,n}; \theta)}{\partial \theta}$$

and

$$S_{2n} = \sum_{i=1}^{k} \delta_{l\,n}^{(i)} \frac{\partial \log h^{(i)}(T_{sn}; \theta)}{\partial \theta} + (n - \tau_n) \frac{\partial \log(1 - H(T_{sn}; \theta))}{\partial \theta}.$$

By $M_{S_1S_2}(t_1,t_2)$ we denote the moment generating function of vector (S_{1n},S_{2n}). Then it is clear that Fisher information of considered model calculated as

$$I_{n,\tau}=E_\theta\left\{\left[\frac{\partial\log p_n(\mathbf{Y}^{(\tau)};\theta)}{\partial\theta}\right]^2\right\}=E_\theta(S_{1n}^2)+2E_\theta(S_{1n},S_{2n})+E_\theta(S_{2n}^2),$$

where

$$E_\theta(S_{1n}^2)=\left.\frac{\partial^2M_{S_1S_2}(t_1,t_2)}{\partial t_1^2}\right|_{t_1=t_2=0};$$

$$E_\theta(S_{1n}S_{2n})=\left.\frac{\partial^2M_{S_1S_2}(t_1,t_2)}{\partial t_1\partial t_2}\right|_{t_1=t_2=0};$$

$$E_\theta(S_{2n}^2)=\left.\frac{\partial^2M_{S_1S_2}(t_1,t_2)}{\partial t_2^2}\right|_{t_1=t_2=0};$$

Theorem 5. *Consider parametric function $\varphi(\theta)$, $\theta\in\Theta\subseteq R$ with differential $\varphi'(\theta)$ and the unbiased estimator $\widehat{\varphi}\left(X_{1n},...,X_{\tau n}\right)$ of $\varphi(\theta)$. Suppose that regularity conditions (I)-(III) hold and the differentiation with respect to θ under integral sign in equality $E_\theta\widehat{\varphi}=\varphi(\theta)$ is permissible for all $\theta\in\Theta$. Then*

$$Var_\theta\left\{\widehat{\varphi}\right\}\geq\frac{\left(\varphi'(\theta)\right)^2}{I_{n,\tau}(\theta)}\ \textit{for all}\ \theta\in\Theta.$$

Fisher information in terms of variation series, has been well studied by many authors. Models for censoring types I and II, their hybrids, records, etc. are considered. In the following section of this paper, you can familiarize yourself with the numerical results of calculating information for random right censoring model.

4 Simulation

Let ξ - random variable of be interest to us, defined on the probability space (Ω,A,P), with values in $X\subseteq R^1=(-\infty,\infty)$. Let $F(x),x\in R^1$, denote the distribution function of the random variable ξ. Let us consider the parametric case when the distribution function F is specified up to the parameter θ: $F\left(x;\theta\right)=P_\theta\left(\xi\leq x\right),\ P=\left\{P_\theta,\theta\in\Theta\right\}$ where $\theta=(\theta_1,..,\theta_s)\in\Theta$, Θ-open interval in R^s. In the case when, according to the sample $\xi^{(n)}=\left(\xi_1,...,\xi_n\right)$, which is the result of independent observations of the random variable ξ, it is required to calculate the Fisher information, the formula is used:

$$I(\theta)=\int\limits_{-\infty}^{+\infty}\left(\frac{\partial\ln f(x;\theta)}{\partial\theta}\right)^2dF(x;\theta).$$

However, when the sample $\xi^{(n)}$ is randomly right-censored, certain difficulties arise when calculating Fisher information, since for some distributions this information is not explicitly calculated and one has to use numerical methods. Let's consider uninformative random censoring. Let $G(x)$ be the distribution function of the random variable η. It is easy to see that the random variables ζ_i are independent and identically distributed with the distribution function $H(x;\theta) = 1 - \overline{F(x;\theta)} \cdot \overline{G(x)}$. In general, the distribution function G plays the role of a "nuisance parameter". Fisher information is defined as follows:

$$I(\theta) = I_G(\theta) = \int\limits_{-\infty}^{\infty} \left(\frac{\partial \ln(f(x;\theta))}{\partial \theta} \right)^2 f(x;\theta)\overline{G(x)}dx$$

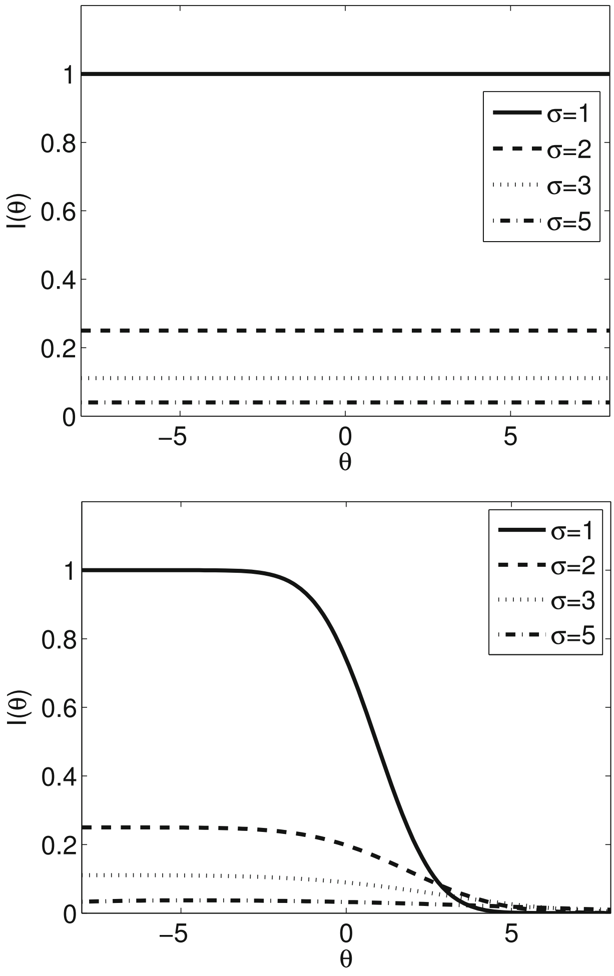


Fig. 1. Fisher information without censoring and with censoring on the right. In the second case, the random variable has a normal distribution with parameters $(0, \sigma^2)$

$$+\int_{-\infty}^{\infty}\left(\frac{\partial \ln(1-F(x;\theta))}{\partial\theta}\right)^2 g(x)(\overline{F(x;\theta)})dx.$$

Let's consider the case when the random variable ξ has a normal distribution with density: $f(x;\theta) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\theta)^2}{2\sigma^2}}$, $x \in (-\infty;\infty)$, $\theta \in (-\infty;\infty)$, $\sigma > 0$, where σ is a known parameter. Let η also have a normal distribution with parameters $(0,1)$ (or $(0,\sigma^2)$). In the following graph you can see a graphical representation of Fisher information (Fig. 1):

The graphs show that in classical models, Fisher information is constant, and with incomplete data, Fisher information decreases with increasing value of the unknown parameter.

Now consider the case when the variance is unknown, i.e. density r.v. distributions ξ has the following form: $f(x;\theta) = \frac{1}{\sqrt{2\pi}\theta}e^{-\frac{(x-a)^2}{2\theta^2}}, x \in (-\infty;\infty), a \in (-\infty;\infty), \theta > 0$: (Fig. 2 and 3))

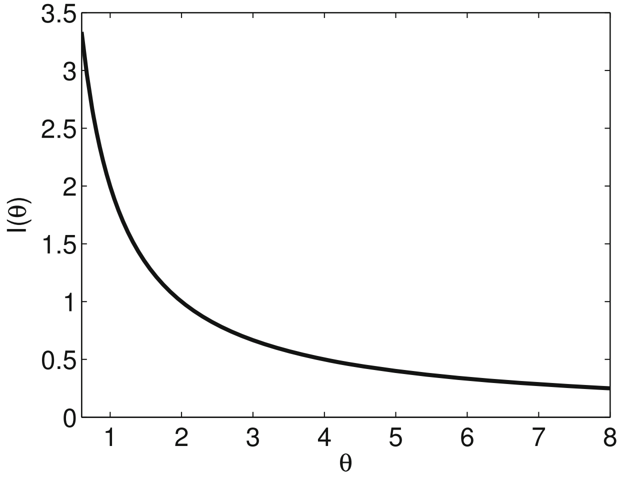


Fig. 2. Fisher information without censoring

Table 1 shows that the value of Fisher's information is greater than that of censoring.

From the graphs we can conclude that Fisher information for complete and incomplete samples can be described by the following inequality $I_{complete}(\theta) \geq I_{incomplete}(\theta)$ (Fig. 4).

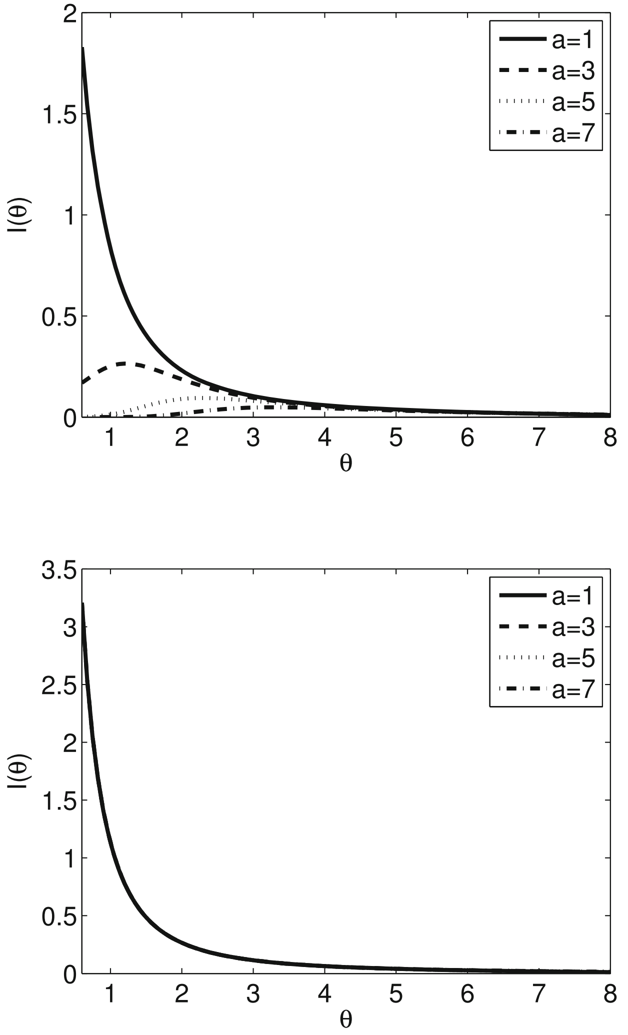


Fig. 3. a) has a normal distribution with parameters (0,1)**; b) has a normal distribution with parameters** (a,1)

Now let's look at how the degree of censoring affects the amount of Fisher information (Table 2).

Analyzing the results presented in Table 1, it can be noted that the higher the degree of censoring is the smaller the amount of information becomes. For example, in the case of a normal distribution with a censoring degree of 50%, the sample contains 81.2% of Fisher's complete information when estimating only the parameter θ_1, 50% when estimating only the parameter θ_2.

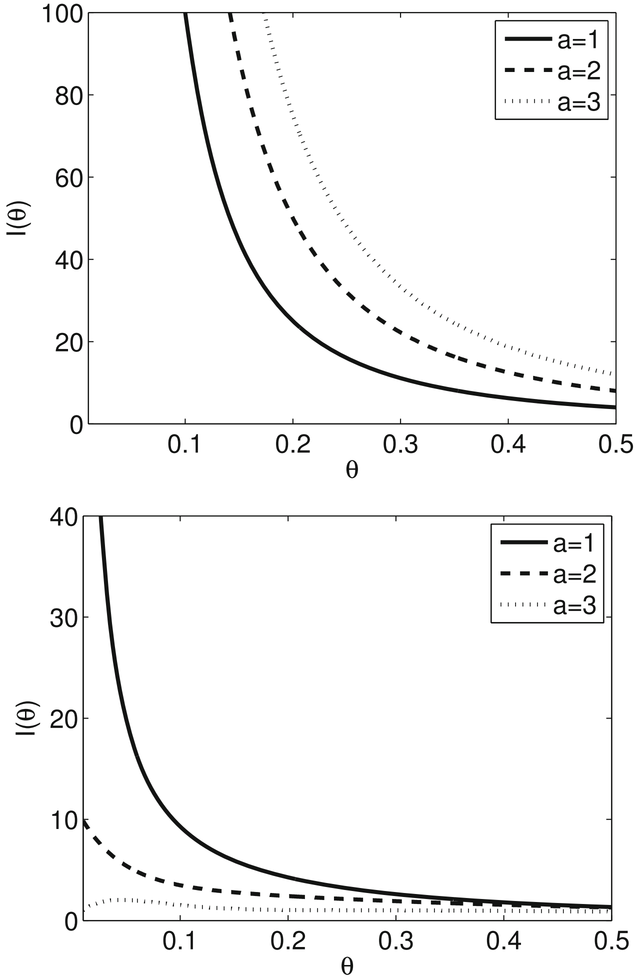


Fig. 4. Fisher information in case of complete data and in case of censoring on the right. In the second case, the random variable has an exponential distribution with parameter 1

Table 1. Fisher information values in uncensored and censored models

Model	Parameter	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$	$\sigma = 5$	$\sigma = 10$
Complete	$I(\theta)$	1	0.25	0.111	0.04	0.01
Censored	$\theta = -8.0$	1.0000	0.2500	0.1109	0.0332	0.0024
	$\theta = -6.4$	1.0000	0.2499	0.1108	0.0364	0.0031
	$\theta = -4.8$	1.0000	0.2495	0.1099	0.0374	0.0040
	$\theta = -3.2$	0.9976	0.2462	0.1073	0.0369	0.0048
	$\theta = -1.5$	0.9566	0.2326	0.1009	0.0351	0.0055
	$\theta = 0.1$	0.7215	0.1956	0.0888	0.0324	0.0060
	$\theta = 1.7$	0.2886	0.1314	0.0703	0.0286	0.0062
	$\theta =3.3$	0.0426	0.0630	0.0479	0.0240	0.0062
	$\theta =4.9$	0.0019	0.0195	0.0270	0.0190	0.0060
	$\theta = 6.5$	0.0000	0.0037	0.0122	0.0140	0.0056
	$\theta = 8.0$	0.0000	0.0005	0.0048	0.0100	0.0052

Similar results were obtained for the gamma distribution: *a)* Gamma distribution with density $f(x;\theta) = \frac{\theta^a}{\Gamma(a)} x^{a-1} e^{-\theta x}$, $x > 0$, $a > 0$, $\theta > 0$:

Table 2. Fisher information quantities in a censored sample observation

Normal distribution		
Degree of censoring	About θ_1 parameter	About θ_2 parameter
5%	0,9930	0,4327
10%	0,9831	1,7116
20%	0,9563	1,4750
30%	0,9206	1,2784
40%	0,8752	1,1198
50%	0,8182	1,0000
60%	0,7467	0,9202
70%	0,6551	0,8800
80%	0,5335	0,8720
Gamma distribution		
Degree of censoring	About θ_1 parameter	About θ_2 parameter
5%	0,9498	0,5542
10%	0,8989	0,5527
20%	0,7989	0,5479
30%	0,6989	0,5393
40%	0,5989	0,5283
50%	0,4989	0,5121
60%	0,3989	0,4888
70%	0,2989	0,4543
80%	0,1989	0,4001
Weibull distribution		
Degree of censoring	About θ_1 parameter	About θ_2 parameter
5%	0,8460	0,4555
10%	0,7411	0,4099
20%	0,5919	0,3644
30%	0,4950	0,3189
40%	0,4342	0,2733
50%	0,4009	0,2278
60%	0,3878	0,1822
70%	0,3860	0,1367
80%	0,3814	0,0911

5 Conclusion

This paper deals with calculation of Fisher information in some models of random censoring including Competing Risks Model. In Competing Risks Model we provided some useful formulas for calculation of information, in hybrid censoring for calculation one can use the moment generating function. Analogously problems considered by authors in some queueing systems in [7–15]. In simple right random censoring simulation shows that if censoring is not informative and does not depend on unknown parameter then information is smaller than in case of no censoring.

References

1. Fisher R.A.: Theory of statistical estimation. In: Proceedings of the Cambridge Philosophical Society, pp. 700–725 (1925)
2. Zacks, Sh.: The theory of statistical inferences. Ohio (1971)
3. Efron, B., Zonstone, I.M.: Fisher Information in terms of the hazard rate. Ann. Stat. **18**(1), 38–62 (1990)
4. Beichert, F., Tittman, P.: Reliability and maintenance. Networks and systems CRC Press. Taylor & Francis Group. A Chapmamn & Hall book
5. Farhal, A.-W.A., Badr, S.A., Abu-Sinada, H.: Abd-Elmongod: analysis of generalized inverted exponential competing risks model in presence of partially observed failure models. Alex. Eng. J. **78**, 74–87 (2023)
6. Shan (Sam) Ma, Z., Krings Axel, W.: Competing Risks Analysis of Reliability, Survivability and Prognostics and Health Management. IEEEAC paper #1626, Final Version Dec. 27, 2007 2008 IEEE
7. Jassim, F.M., Asaedi, H.J.N., Zainab Falih Hamza: Estimating system reliability functions for the generalized exponential distribution with application. Periodicals Eng. Natural Sci. **11**(3), 108–114 (2023)
8. Winick, K.A.: Cramer-Rao lower bounds on the performance of Charge-coupled-device optical position estimators. Optical Soc. Am. **3**(11), 1809–1815 (1986)
9. Kumar, R., Sharma, S.: Time-dependent analysis of a single-server queuing model with discouraged arrivals and retention of reneging customers. RT&A **4**(47), 84–92 (2017)
10. Kumar Singh, S.: Moderate deviations for maximum likelihood estimators from simple server queues. Probability, Uncertainty and Quantitative Risk (2020). https://doi.org/10.1186/s41546-020-00044-z
11. Lin, Z., Zon, Q., Sally Ward, E., Ober, R.J.: Cramer-rao lower bound for parameter estimation in nonlinear systems. IEEE Signal Process. Lett. **12**(12), 855–858 (2005)
12. Tisharsky, P.: Posterior cramer-rao bound for adaptive harmonic retrieval. IEEE Trans. Signal Process. **43**(5), 1299–1308 (1995)
13. Zamir, R.: A Proof a Fisher Information inequality via data processing argument. IEEE Trans. Inf. Theory **44**(3), 1246–1252 (1998)
14. Mageed Demetres Kouvatsos, I.A.: The Impact of Information Geometry on the Analysis of the stable M/G/1 Queue Manifold. In: Proceedings of the 10th International Conference on Operation research and Enterprise Systems (ICORES), pp. 153–160 (2021)
15. Li, C., Okamura, H., Dohi, T.: Parameter estimation on M_t/M/1/K Queueing System with utilization data. IEEE Access. **7**, 42664–42671 (2019)

16. Abdushukurov, A.A., Kim, L.V.: Lower Cramer-Rao and Bhattacharya bounds for random censored observation. J. Soviet. Math. **38**(5), 2171–2185 (1987)
17. Prakasa Rao B.L.S.: Remarks on Cramer-Rao type integral inequalities for random censored data. Analysis of censored data **27**, 163–175 (1995)
18. Improved Cramer-Rao inequalities for randomly censored data: Prakasa Rao B.L.S. J. Iranian Stat. Soc. **17**, 17–26 (2018)
19. Zheng, G., Gastwirth, J.: On the Fisher information in randomly censored data. J. Statist. Probab. Lett. **52**, 421–426 (2001)
20. Zheng, G., Gastwirth, J.: On the Fisher information in ordered randomly censored data with application to characterization problems. Stat. Sin. **13**, 507–517 (2003)
21. Abdushukurov, A.A. Erisbaev, S.A.: Fisher information decomposition in terms of hazard rate functions under random censoring from the right. Electron. J. Sci. Educ. Karakalpakstan. **3-4**, 36–43 (2020)
22. Abdushukurov, A.A. Erisbaev, S.A.: Fisher information and Cramer-Rao inequality for Competing Risks Model. Electron. Bull. Inst. Math. **4**5, 50–58. (2022)(in Russian)
23. Walker, S.G.: A self-improvement to the Cauchy-Swartz inequality. Statist. Probab. Lett. **122**, 86–90 (2017)
24. Nurmukhamedova N.S., Yusupov J.R.: Fisher information in models of right random censoring. Modern problems of applied mathematics and information technology AL-KHORASMY-2016, pp. 222–226 (2016)

Further Improvement of the Basic Lemma of Critical Bienaymé-Galton-Watson Branching Systems and Its Applications

Azam A. Imomov[1,2](✉) and Sarvar B. Iskandarov[2]

[1] Karshi State University, Karshi, Uzbekistan
imomov_azam@mail.ru
[2] Urgench State University, Urgench, Uzbekistan
sarvar.i@urdu.uz

Abstract. The paper considers the Bienaymé-Galton-Watson Branching Systems, in which the mean per-capita offspring number is equal to one and the variance is infinite. This system is called a critical one. We state and prove an alternative variant of Basic Lemma of the theory of critical Bienaymé-Galton-Watson system. The proved lemma essentially improves the corresponding result of the previous works of the authors. This assertion plays a key role in formulating the local limit theorem with explicit terms in the asymptotic expansion of local probabilities on positive trajectories of the system considered.

Keywords: Branching system · Generating functions · Markov chain · Slow variation · Basic Lemma · Transition probabilities · Invariant measures · Limit theorems

1 Background, Assumptions and Purpose

Models of stochastic branching systems describe the evolution of the population size in the reproductive individuals system. These models most clearly illustrate numerous stochastic phenomena occurring both in nature and in human activity. The simplest and most famous Branching model is the discrete-time Bienaymé-Galton-Watson (BGW) branching system, in which the sequence of generation numbers defines the homogeneous-discrete-time Markov chain, and the reproduction law of each individual is independent of time and other individuals. This model originally evolved as a family survival model in the second half of the 19th century, today has numerous generalizations and modifications. The integration of various scientific fields has made it possible to find new applications of the branching system models in many fields, such as graph theory, queuing theory, combinatorics, cell biology, molecular biology, etc. Depending on the context, the branching system of one or another model is used to describe an evolution mechanism of individuals. Branching systems, as population growth models have an obvious influence on the development of the population dynamics theory; see [1,2,6,10,16].

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 228–240, 2024.
https://doi.org/10.1007/978-3-031-65385-8_17

Let $\mathbb{N} = \{1, 2, \ldots\}$ be the set of natural numbers and $\mathbb{N}_0 = \{0\} \cup \mathbb{N}$. We consider the BGW Branching System with branching rates $\{p_j, j \in \mathbb{N}_0\}$. Denoting by Z_n the population size in the system at the time $n \in \mathbb{N}_0$, we have a reducible, homogeneous-discrete-time Markov chain with a state space consisting of two classes: $\mathcal{S}_0 = \{0\} \cup \mathcal{S}$, where $\mathcal{S} \subset \mathbb{N}$, therein the state $\{0\}$ is absorbing, and $\mathcal{S}$ is the class of possible essential communicating states. The population size of the system can be consistently described by the following recursive relations:

$$Z_{n+1} = \xi_{n1} + \xi_{n2} + \cdots + \xi_{nZ_n} \tag{1.1}$$

for all $n \in \mathbb{N}$, where ξ_{nk} are independent and identically distributed random variables with the common distribution

$$\mathbb{P}\left\{\xi_{nk} = j\right\} = p_j \qquad \text{for all} \quad j \in \mathcal{S}_0$$

and they are interpreted as the number of descendants of the kth individual in the nth generation.

Put into consideration Markov chain n-step transition probabilities

$$P_{ij}(n) := \mathbb{P}\left\{Z_{n+k} = j \;\middle|\; Z_k = i\right\} \qquad \text{for all} \quad i, j \in \mathcal{S}_0.$$

These probabilities are completely determined by the branching rates $\{p_j\}$, since, denoting $\mathsf{p}_j(n) := P_{1j}(n)$, we observe that the probability Generating Function (GF)

$$\sum_{j \in \mathcal{S}_0} P_{ij}(n) s^j = \left[f_n(s)\right]^i \tag{1.2}$$

for any $i \in \mathcal{S}$ and $s \in [0, 1)$, where $f_n(s) = \sum_{j \in \mathcal{S}_0} \mathsf{p}_j(n) s^j$ is the n-fold iteration of the GF

$$f(s) := \sum_{j \in \mathcal{S}_0} p_j s^j.$$

Let the series $\sum_{j \in \mathcal{S}} j p_j$ converges. Then

$$m := \sum_{j \in \mathcal{S}} j p_j = f'(1-)$$

is the mean per-capita offspring number, the value of which regulates the classification of $\mathcal{S}$. In fact, using (1.2), it can be observed that $\mathbb{E}\left[Z_n \;\middle|\; Z_0 = 1\right] = m^n$, i.e., the mathematical expectation of Z_n asymptotically behaves differently depending on the value of the parameter m. So, the chain $\{Z_n\}$ is classified as sub-critical, critical and supercritical if $m < 1$, $m = 1$ and $m > 1$ respectively. For all cases $f_n(0) = \mathsf{p}_0(n)$ is a vanishing probability of the system initiated by one individual, and it is monotone and $\lim_{n \to \infty} \mathsf{p}_0(n) = q$, where q is an extinction probability of the system. This probability is a smallest nonnegative root of the fixed-point equation $f(s) = s$ on the domain of $\{s : s \in [0, 1]\}$. Moreover $f_n(s) \to q$ as $n \to \infty$ uniformly in $s \in [0, 1)$; see [2, Ch.I, §§1–5].

In this paper we consider the critical case with the offspring GF having a form of

$$f(s) = s + (1-s)^{1+\nu}\mathcal{L}\left(\frac{1}{1-s}\right), \qquad [f_\nu]$$

for $s \in [0,1)$, where $0 < \nu < 1$ and $\mathcal{L}(*)$ is slowly varying (SV) function at infinity in the sense of Karamata. By the criticality of our system, the assumption $[f_\nu]$ implies that the second $2b := f''(1-) = \infty$. If $0 < b < \infty$ then $\nu = 1$ and $\mathcal{L}(t) \to b$ as $t \to \infty$.

The critical case considered by authors of [7,8,12–14].

In what follows we will use the condition $[f_\nu]$ in the following form:

$$f(s) - s = (1-s)\Lambda(1-s), \qquad [f_\Lambda]$$

where

$$\Lambda(y) := \frac{f(1-y)-(1-y)}{y} = y^{\nu}\mathcal{L}\left(\frac{1}{y}\right) \qquad \text{for} \quad y \in (0,1].$$

In our previous work [5] it was proved that

$$U_n(s) := \frac{f_n(s) - f_n(0)}{f_{n+1}(0) - f_n(0)} \qquad [\mathcal{S}_U]$$

approaches the limit function $U(s)$ as $n \to \infty$ for $s \in [0,1]$, such that $U(s)$ is the GF of the invariant measure for the critical BGW system $\{Z_n\}$.

Theorem 1 **([5]).** *Let*

$$\mathcal{V}(s) := \frac{1}{\nu\Lambda(1-s)} \qquad \textit{and} \qquad J(s) := \frac{1-f'(s)}{\Lambda(1-s)} - 1.$$

If condition $[f_\Lambda]$ is satisfied, then $U_n(s) \to U(s)$ as $n \to \infty$, where

(i) *the GF $U(s)$ has the following form:*

$$U(s) = \mathcal{V}(s) - \mathcal{V}(0), \tag{1.3}$$

(ii) *the derivative $U'(s)$ has the following expression:*

$$U'(s) = J(s)\frac{\mathcal{V}(s)}{1-s}. \tag{1.4}$$

The proof of the Theorem 1 result is based on Slack's [17] arguments, who defined the prelimit function in the form of

$$\widehat{U}_n(s) := \frac{f_n(s) - f_n(0)}{f_n(0) - f_{n-1}(0)}.$$

We slightly altered the Slack's definition to $[\mathcal{S}_U]$ in [5].

Remark 1. The function $U(s)$ admits a power series expansion

$$U(s)=\sum_{j\in\mathcal{S}}u_js^j,$$

where $u_j=\sum_{k\in\mathcal{S}}u_kP_{kj}(1)$ and $\sum_{k\in\mathcal{S}}u_kp_0^k=1$; see [17]. Then relation (1.4) implies that

$$u_1=U'(0)=\frac{J(0)}{\nu p_0}=\frac{1-p_0-p_1}{\nu p_0^2}.$$

We now recall the following statement related to the prelimit function $U_n(s)$ and the GF of the limiting invariant measure and clearly showing an explicit asymptotic expression of the function

$$R_n(s):=1-f_n(s).$$

Lemma 1 (Basic Lemma *[9]***).** *If the condition* $[f_\Lambda]$ *is satisfied then*

$$R_n(s)=\frac{\mathcal{N}(n)}{(\nu n)^{1/\nu}}\cdot\left[1-\frac{U_n(s)}{\nu n}\right], \tag{1.5}$$

where the function $\mathcal{N}(x)$ *is SV at infinity and*

$$\mathcal{N}(n)\cdot\mathcal{L}^{1/\nu}\left(\frac{(\nu n)^{1/\nu}}{\mathcal{N}(n)}\right)\longrightarrow 1 \qquad as \quad n\to\infty, \tag{1.6}$$

and the function $U_n(s)$ *has the following properties:*

(i) $U_n(s)\to U(s)$ *as* $n\to\infty$*, where* $U(s)$ *has the form of* (1.3)*;*
(ii) $\lim_{s\uparrow 1}U_n(s)=\nu n$ *for each fixed* $n\in\mathbb{N}$*;*
(iii) $U_n(0)=0$ *for each fixed* $n\in\mathbb{N}$*.*

Further, along with our main condition $[f_\Lambda]$, we make also some extra assumption for the function $\mathcal{L}(\cdot)$ as follows. Since $\mathcal{L}(\cdot)$ is SV, then by its definition $\mathcal{L}(\lambda x)/\mathcal{L}(x)\to 1$ as $x\to\infty$ for each $\lambda>0$. Then

$$\omega_\lambda(x):=\frac{\mathcal{L}(\lambda x)}{\mathcal{L}(x)}-1$$

decreases to zero as $x\to\infty$. If a decreasing rate of $\omega_\lambda(x)$ is known, then the function $\mathcal{L}(\cdot)$ is called SV with remainder $\omega_\lambda(x)$ at infinity; see [3, p. 185].

In the paper we will use the Landau symbols o, $\mathcal{O}$, $\mathcal{O}^*$ for comparison of two functions $f(\cdot)$ and $h(\cdot)$ on the point x_0 (finite or infinite):

$$f(x)=o\bigl(h(x)\bigr) \qquad\Longleftrightarrow\qquad \frac{|f(x)|}{|h(x)|}\to 0,$$

$$f(x)=\mathcal{O}\bigl(h(x)\bigr) \qquad\Longleftrightarrow\qquad \frac{|f(x)|}{|h(x)|}\le A,$$

$$f(x)=\mathcal{O}^{*}\big(h(x)\big) \qquad \Longleftrightarrow \qquad \frac{|f(x)|}{|h(x)|}\to A,$$

as $x\to x_0$, where $A\neq 0$. And also $f(x)\sim h(x)$ means that $f(x)/h(x)\to 1$.

The following statement is an improved analogue of the Basic Lemma 1.

Lemma 2 **([?]).** *Let the condition* $[f_\Lambda]$ *is satisfied and* $\omega_\lambda(x)=o\big(\mathcal{L}(x)/x^{\nu}\big)$. *Then*

$$\frac{1}{\Lambda\big(R_n(s)\big)}-\frac{1}{\Lambda(1-s)}=\nu n+\frac{1+\nu}{2}\cdot\ln\big[\Lambda(1-s)\nu n+1\big]+\rho_n(s), \tag{1.7}$$

where $\rho_n(s)=o\big(\ln n\big)+\sigma_n(s)$ *and, the function* $\sigma_n(s)$ *is bounded uniformly in* $s\in[0,1)$ *for each* $n\in\mathbb{N}$ *and converges to the limit function* $\sigma(s)$ *as* $n\to\infty$ *which is bounded for all* $s\in[0,1)$.

The first direct result from Lemma 1 and Lemma 2 are assuredly, the expression for the survival probability $Q_n:=\mathbb{P}\{Z_n>0\}=R_n(0)$ at the moment n of the BGW system initiated by a single founder-individual.

Remark 2. By writing formula (1.5) as

$$U_n(s)=\left[1-\frac{R_n(s)}{Q_n}\right]\nu n, \tag{1.8}$$

we state that Lemma 1 reports an asymptotic relation between the probabilities $\big\{\mathsf{p}_j(n), j\in\mathcal{S}\big\}$ and the invariant measure $\{u_j, j\in\mathcal{S}\}$ generated by the limit function $U(s)$.

Remark 3. The advantage of Lemma 2 is that it more exactly improves an analogous statement established in the paper [11, Theorem 1], in which the offspring law variance was assumed to be finite and later it was refined under the third finite moment assumption in [6, p. 20]. In both mentioned works, $\nu=1$ and $\Lambda(y)\equiv y$, and at the same time $f''(1-)n/2$ appeared instead of first term νn, furthermore, the subsequent tail terms were found on the right-hand side of (1.7).

Our purpose is as follows. We first state and prove an improved and strengthened form of the Basic Lemma 1. To do this, we essentially use the asymptotic formula (1.7). As a result, we find a normalizing constant $C(n)$ such that $C(n)f_n(s)$ approaches $U(s)$ as $n\to\infty$. In addition, we will estimate the speed rate of this approximation. For our purpose, we adopt the decreasing rate of the remainder term of the SV-function $\mathcal{L}(\cdot)$ to be

$$\omega(x):=\omega_\lambda(x)=\mathcal{O}\left(\frac{\mathcal{L}(x)}{x^{\nu}}\right) \qquad \text{as} \quad x\to\infty, \tag{$\mathcal{L}_\omega$}$$

that is more exact decreasing speed rate condition, than it was assumed in the Lemma 2. Our results facilitate to refine classical limit theorems.

The rest of this paper is organized as follows. Section 2 contains main results. Section 3 provides the proof of main results.

2 Main Results

In this section we present our main results.

Lemma 3. *If conditions* $[f_\Lambda]$ *and* $[\mathcal{L}_\omega]$ *are satisfied, then assertion of Lemma 2 remains true.*

We bypass the proof of this lemma since it duplicates the rationale for the proof of Lemma 2, demonstrated in detail in the work [?].

Let

$$\mathcal{M}_n(s) := 1 - \frac{\Lambda\big(R_n(s)\big)}{\Lambda\left(Q_n\right)}.$$

We state our first result, which improves the Basic Lemma 1 as follows.

Lemma 4. *If conditions* $[f_\Lambda]$ *and* $[\mathcal{L}_\omega]$ *are satisfied, then*

$$R_n(s) = Q_n \cdot \Big[1 - \Lambda\left(Q_n\right) U_n(s)\Big], \tag{2.1}$$

where the function $U_n(s)$ *has the following properties:*

(i) $U_n(s) \to U(s)$ *as* $n \to \infty$*, where* $U(s)$ *has the form of* (1.3)*;*
(ii) $\lim_{s\uparrow 1} U_n(s) = 1\big/\Lambda\left(Q_n\right)$ *for each fixed* $n \in \mathbb{N}$*;*
(iii) $U_n(0) = 0$ *for each fixed* $n \in \mathbb{N}$*;*
(iv) $n\mathcal{M}_n(s) \to U(s)$ *and*

$$\nu\Lambda\left(Q_n\right)U_n(s) = \mathcal{M}_n(s) + \mathcal{O}^*\left(\ln n\big/n\right) \qquad \textit{as} \quad n \to \infty. \tag{2.2}$$

Now consider probabilities

$$\mathsf{p}_j^{\mathcal{S}}(n) := \mathbb{P}\left\{Z_n = j \,\middle|\, j \neq 0, Z_0 = 1\right\}$$

and define the GF

$$\mathcal{V}_n(s) = \sum_{j\in\mathcal{S}} \mathsf{p}_j^{\mathcal{S}}(n) s^k.$$

This generates a BGW branching system $\left\{Z_n^{\mathcal{S}}\right\}$ with positive trajectories. The system $\left\{Z_n^{\mathcal{S}}\right\}$ is an irreducible, homogeneous-discrete-time Markov chain with a state space $\mathcal{S}$.

Using Lemma 4, we establish the following theorem.

Theorem 2. *If conditions* $[f_\Lambda]$ *and* $[\mathcal{L}_\omega]$ *are satisfied, then*

$$\frac{\nu n}{Q_n}\mathcal{V}_n(s) = \mathcal{U}_n(s)\left(1 - \frac{1+\nu}{2\nu}\frac{\ln n}{n}\big(1+o(1)\big)\right) \qquad \textit{as} \quad n \to \infty \tag{2.3}$$

uniformly in $s \in [0, 1)$ *and*

$$\mathcal{U}_n(s) = U(s)\left(1 - \frac{1+\nu}{2}\Lambda(1-s)\frac{\ln \nu_n(s)}{\nu_n(s)}\big(1+o(1)\big)\right) \tag{2.4}$$

as $n \to \infty$*, where* $\nu_n(s) = \Lambda(1-s)\nu n + 1$*.*

The next result is a local limit theorem for the branching system $\{Z_n^{\mathcal{S}}\}$, follows from Theorem 2.

Theorem 3. *If conditions $[f_{\Lambda}]$ and $[\mathcal{L}_\omega]$ are satisfied, then*

$$\frac{(\nu n)^{(1+\nu)/\nu}}{\mathcal{N}_\nu(n)} p_j^{\mathcal{S}}(n) = u_j \cdot \left(1 + \mathcal{O}^* \left(\frac{\ln n}{n}\right)\right) \qquad as \quad n \to \infty,$$

where $\mathcal{N}_\nu(n)$ is SV at infinity, such that

$$\mathcal{N}_\nu(n) \mathcal{L}^{1/\nu} \left(\frac{(\nu n)^{1/\nu}}{\mathcal{N}_\nu(n)}\right) \longrightarrow 1 \qquad as \quad n \to \infty.$$

Remark 4. If conditions $[f_{\Lambda}]$ and $[\mathcal{L}_\omega]$ are satisfied, then

$$\sum_{j=1}^{n} u_j = \frac{1}{\nu \Gamma(1+\nu)} n^\nu \mathcal{L}_\nu(n),$$

where $\mathcal{L}_\nu(n)\mathcal{L}(n) \to 1$ as $n \to \infty$. This statement appears due to the Hardy-Littlewood Tauberian theorem for power series; see [4, Ch.XIII.5].

3 Preliminaries

We will need the following auxiliary statements on properties of SV-functions with the remainder, which are especially important in our purpose.

Lemma 5. *Let $K(y)$ be a positive function in $y \in (0, \infty)$ and $K(y) \downarrow 0$ as $y \downarrow 0$ and let*

$$\phi(y) := y - yK(y).$$

If conditions $[f_{\Lambda}]$ and $[\mathcal{L}_\omega]$ hold, then

$$\mathcal{L}\left(\frac{1}{\phi(y)}\right) = \mathcal{L}\left(\frac{1}{y}\right)\left(1 + K(y)\omega\left(\frac{1}{y}\right)\right) \qquad as \quad y \downarrow 0. \tag{3.1}$$

Proof. In our conditions, $\mathcal{L}(x)$ is differentiable and $\mathcal{L}(x) = x^\nu \Lambda(1/x)$. Substituting $y = 1/x$ in the function $\phi(y)$, we have

$$\mathcal{L}\left(\frac{1}{\phi(y)}\right) = \mathcal{L}\left(\frac{1}{y - yK(y)}\right) = \mathcal{L}\left(\frac{x}{1 - k(x)}\right), \tag{3.2}$$

where $k(x) = K(1/x) > 0$. It is easy to see that $x < [x/(1 - k(x))]$. We can write now the mean value theorem in the following form:

$$\mathcal{L}\left(\frac{x}{1 - k(x)}\right) - \mathcal{L}(x) = \mathcal{L}'(\xi) \frac{xk(x)}{1 - k(x)}, \tag{3.3}$$

where $\xi = \xi(x)$ is a mean value, such that $x < \xi < \big[x/\big(1 - k(x)\big)\big]$. Then it can be written as follows:

$$\xi(x) = \frac{x}{1 - k(x)} - \theta \frac{xk(x)}{1 - k(x)} = \frac{1 - \theta k(x)}{1 - k(x)} x \tag{3.4}$$

for some $\theta \in (0, 1)$. At the same time, by Lamperti's arguments [3, p. 401], the function $\mathcal{L}(x)$ can be represented as

$$\mathcal{L}(x) = p_0 \exp \int_1^x \frac{\varepsilon(t)}{t}\, dt, \tag{3.5}$$

where $\varepsilon(x) = \mathcal{O}\left(\omega(x)\right)$. This representation appeared in the work [9]. Next, in virtue of the assumption $[\mathcal{L}_\omega]$, we have $\varepsilon(x) = \mathcal{O}\left(\mathcal{L}(x)/x^\nu\right)$ as $x \to \infty$. Then the integral representation (3.5) implies that

$$\mathcal{L}'(x) = \mathcal{L}(x)\frac{\varepsilon(x)}{x} = \mathcal{O}\left(\frac{\mathcal{L}^2(x)}{x^{1+\nu}}\right) \qquad \text{as} \quad x \to \infty. \tag{3.6}$$

In our assumption, $k(x) \downarrow 0$ as $x \to \infty$ and hence $\xi(x) \sim x$. Therefore $\mathcal{L}'(\xi) \sim \mathcal{L}'(x)$. Then combination of relations (3.2)–(3.6) leads to the fact that

$$\begin{aligned} \mathcal{L}\left(\frac{1}{\phi(y)}\right) - \mathcal{L}\left(\frac{1}{y}\right) &= \mathcal{L}'(\xi)\frac{xk(x)}{1 - k(x)} \\ &= \frac{xk(x)}{1 - k(x)}\mathcal{O}\left(\frac{\mathcal{L}^2(x)}{x^{1+\nu}}\right)\big(1 + o(1)\big) \qquad \text{as} \quad x \to \infty \\ &= K(y)\mathcal{O}\left(y^\nu \mathcal{L}^2\left(\frac{1}{y}\right)\right)\big(1 + o(1)\big) \qquad \text{as} \quad y \downarrow 0. \end{aligned}$$

The last equality implies that

$$\mathcal{L}\left(\frac{1}{\phi(y)}\right) - \mathcal{L}\left(\frac{1}{y}\right) = K(y)\mathcal{L}\left(\frac{1}{y}\right)\mathcal{O}\left(y^\nu \mathcal{L}\left(\frac{1}{y}\right)\right)\big(1 + o(1)\big) \tag{3.7}$$

as $y \downarrow 0$. Now, from (3.7) and $[\mathcal{L}_\omega]$ the assertion (3.1) readily follows.

The Lemma 5 is proved. □

Lemma 6. *Let $L(x)$ be the SV-function with remainder at infinity. If the SV-remainder term of this function is $\omega(x) = \mathcal{O}\left(L(x)/x^\sigma\right)$ for some $\sigma > 0$, then*

$$C_L := \lim_{x\to\infty} L(x) < \infty$$

and

$$L(x) = C_L + \mathcal{O}^*\left(1/x^\sigma\right) \qquad \textit{as} \quad x \to \infty. \tag{3.8}$$

Conversely, if the function $L(x)$ is in the asymptotic expansion form (3.8) for some $\sigma > 0$, then the remainder term $\omega(x)$ has a decreasing rate in the order of $\mathcal{O}^*\left(1/x^\sigma\right)$ as $x \to \infty$.

Proof. The well-known representation theorem states that

$$L(x) = \exp\left\{\eta(x) + \int_b^x \frac{\varepsilon(u)}{u} du\right\}, \tag{3.9}$$

where b is in the domain of $L(x)$ and $\eta(x)$ is a bounded measurable function on $[b, \infty)$ such that $\eta(x) \to C_\eta$ as $x \to \infty$, $|C_\eta| < \infty$; see [15, Ch I, §1]. Under the conditions of the lemma, the integral expression (3.5) is valid and $\varepsilon(x) = \mathcal{O}\left(L(x)/x^\sigma\right)$. Since $\sigma > 0$, due to the properties of SV-functions we observe that the improper integral $\int_b^\infty [\varepsilon(u)/u]\, du$ converges. Thus, (3.9) entails that

$$L(x) \longrightarrow \exp\left\{C_\eta + \int_b^\infty \frac{\varepsilon(u)}{u} du\right\} =: C_L < \infty \qquad \text{as} \quad x \to \infty. \tag{3.10}$$

Therefore we have

$$\omega(x) = \mathcal{O}^*\left(1/x^\sigma\right) \qquad \text{as} \quad x \to \infty.$$

Now we prove formula (3.8). To do this, we first note that

$$\eta(x) = C_\eta + \mathcal{O}^*\left(1/x^\sigma\right) \qquad \text{as} \quad x \to \infty,$$

which follows from the arguments of [3, Ch.3.12.1] in combination with the statement (3.10). Then we write relation (3.9) in the form

$$L(x) = \exp\Big\{C_\eta + \mathcal{O}^*\left(1/x^\sigma\right)\Big\}L_0(x) \qquad \text{as} \quad x \to \infty, \tag{3.11}$$

where $L_0(x)$ is the normalised SV-function, such that

$$L_0(\infty) = \int_b^\infty \frac{\varepsilon(u)}{u} du =: C_0 < \infty.$$

Therefore, we have

$$C_0 - L_0(x) = C_0\left(1 - \frac{L_0(x)}{C_0}\right) = C_0\left(1 - \exp\left(-\int_x^\infty \frac{\varepsilon(u)}{u} du\right)\right).$$

The integral in the last line tends to zero as the tail of a convergent integral. Then, taking into account that $1 - e^{-u} \sim u$ as $u \to 0$, we obtain the following relations:

$$C_0 - L_0(x) = \mathcal{O}^*\left(\int_x^\infty \frac{1}{u^{1+\sigma}} du\right) = \mathcal{O}^*\left(1/x^\sigma\right) \qquad \text{as} \quad x \to \infty. \tag{3.12}$$

Combining relation (3.12) with formula (3.11) gives

$$\begin{aligned} L(x) &= \exp\Big\{C_\eta + \mathcal{O}^*\left(1/x^\sigma\right)\Big\}\Big(C_0 + \mathcal{O}^*\left(1/x^\sigma\right)\Big) \\ &= C_0 \exp\big\{C_\eta\big\}\Big(1 + \mathcal{O}^*\left(1/x^\sigma\right)\Big). \end{aligned}$$

Hence, denoting $C_L := C_0 \exp\bigl\{C_\eta\bigr\}$, we come to the relation (3.8).

The converse part of the theorem follows from the relation (3.8). In fact,

$$\frac{L(\lambda x)}{L(x)} = \frac{C_L + \mathcal{O}^*\left(1/x^\sigma\right)}{C_L + \mathcal{O}^*\left(1/x^\sigma\right)} = 1 + \mathcal{O}^*\left(1/x^\sigma\right) \qquad \text{as} \quad x \to \infty$$

for each $\lambda > 0$, which means that $L(x)$ is the SV-function with remainder

$$\omega(x) = \mathcal{O}^*\left(1/x^\sigma\right).$$

The Lemma 6 is proved. □

4 Proof of Results

Proof ***(Proof of Lemma*** **4*).*** Combining $[f_\Lambda]$ and $[\mathcal{S}_U]$ we write

$$U_n(s) = \frac{Q_n - R_n(s)}{Q_n \Lambda\bigl(Q_n\bigr)} = \frac{1}{\Lambda\bigl(Q_n\bigr)}\left[1 - \frac{R_n(s)}{Q_n}\right]. \tag{4.1}$$

Therefore, formula (2.1) is immediate. Now we verify the rest part of the lemma concerning the properties of the prelimit function $U_n(s)$. Initially, Theorem 1 states that $U_n(s) \to U(s)$ as $n \to \infty$. The immediate facts are that

$$U_n(1-) = \frac{1}{\Lambda\left(Q_n\right)} \qquad \text{for each fixed} \quad n \in \mathbb{N}$$

and $U_n(0) = 0$ for each fixed $n \in \mathbb{N}$, since $R_n(1-) = 0$ and $R_n(0) = Q_n$.

Now, recalling that $\Lambda(y) = y^\nu \mathcal{L}\left(1/y\right)$ and Lemma 1 we write

$$\mathcal{M}_n(s) = 1 - \left(\frac{R_n(s)}{Q_n}\right)^\nu \frac{\mathcal{L}\left(1/R_n(s)\right)}{\mathcal{L}\left(1/Q_n\right)}. \tag{4.2}$$

Here we use the Lemma 5 for the asymptotic estimate of last ratio in right-hand side of (4.2). We compile the function to be estimated with the conditions of the lemma as follows. First, we write out from (4.1) that

$$R_n(s) = Q_n - Q_n \Lambda\bigl(Q_n\bigr) U_n(s). \tag{4.3}$$

Since $U_n(s) \to U(s)$ as $n \to \infty$ for all $s \in [0, 1)$, statement (3.1) is justified for $y = Q_n$, $K(y) = \mathcal{O}^*\left(\Lambda(y)\right)$ and $\omega\left(1/y\right) = \mathcal{O}^*\left(y^\nu\right)$ as $y \downarrow 0$. In the last step, to estimate the remainder $\omega\left(1/y\right)$, we relied on Lemma 6. At the same time, Lemma 3 implies

$$\Lambda\bigl(R_n(s)\bigr) = \frac{\Lambda(1-s)}{\nu_n(s)}\left(1 - \frac{1+\nu}{2}\Lambda(1-s)\frac{\ln \nu_n(s)}{\nu_n(s)}\bigl(1 + o(1)\bigr)\right) \tag{4.4}$$

as $n \to \infty$, where $\nu_n(s) = \Lambda(1-s)\nu n + 1$.

Therefore

$$\mathcal{L}\left(\frac{1}{R_n(s)}\right) = \mathcal{L}\left(\frac{1}{Q_n}\right)\left(1 + \mathcal{O}^*\left(\frac{1}{n^2}\right)\right) \qquad \text{as} \quad n \to \infty. \tag{4.5}$$

Then from (1.5) and (4.5) it follows

$$\begin{aligned}\mathcal{M}_n(s) &= 1 - \left[1 - \frac{U_n(s)}{\nu n}\right]^{\nu}\left(1 + \mathcal{O}^*\left(\frac{1}{n^2}\right)\right)\\ &= \frac{U_n(s)}{n} + \mathcal{O}^*\left(\frac{1}{n^2}\right) \qquad \text{as} \quad n \to \infty. \end{aligned}\tag{4.6}$$

Hence $n\mathcal{M}_n(s) \to U(s)$ as $n \to \infty$ and

$$U_n(s) = n\mathcal{M}_n(s) + \mathcal{O}^*\left(\frac{1}{n}\right) \qquad \text{as} \quad n \to \infty.$$

Since $\Lambda(1) = \mathcal{L}(1) = p_0$, relation (4.4) entails that

$$\Lambda(Q_n) = \frac{1}{\nu n}\left(1 - \frac{1+\nu}{2\nu}\frac{\ln n}{n} + o\left(\frac{\ln n}{n}\right)\right) \qquad \text{as} \quad n \to \infty. \tag{4.7}$$

Then considering (4.7) as $s = 0$, we have

$$\begin{aligned}\Lambda(Q_n)U_n(s) &= \Lambda(Q_n)n\mathcal{M}_n(s) + \mathcal{O}^*\left(\frac{\Lambda(Q_n)}{n}\right)\\ &= \frac{1}{\nu}\mathcal{M}_n(s)\left(1 - \frac{1+\nu}{2\nu}\frac{\ln n}{n} + o\left(\frac{\ln n}{n}\right)\right)\end{aligned}\tag{4.8}$$

as $n \to \infty$ and thus formula (2.2) is immediate. □

Proof ***(Proof of Theorem*** **2*).*** We can write directly that

$$U_n(s) = \frac{1}{\Lambda(Q_n)}\frac{1}{Q_n}\mathcal{V}_n(s). \tag{4.9}$$

Denoting $\mathcal{U}_n(s) := n\mathcal{M}_n(s)$, from (4.8) and (4.9) we have

$$\begin{aligned}\frac{\nu n}{Q_n}\mathcal{V}_n(s) &= \nu n\Lambda(Q_n)U_n(s)\\ &= \mathcal{U}_n(s)\left(1 - \frac{1+\nu}{2\nu}\frac{\ln n}{n} + o\left(\frac{\ln n}{n}\right)\right) \qquad \text{as} \quad n \to \infty.\end{aligned}$$

Relation (2.3) is proved.

Next, we proceed to the proof of (2.4). It is known that the invariance property of the limiting measure $\{u_j, j \in \mathcal{S}\}$ is expressed by the Abel equation

$U\big(f(s)\big) = U(s) + 1$ for the function $U(s)$. Then repeatedly use of this equation, with considering of relation (1.3), yields

$$\begin{cases} \dfrac{1}{\varLambda\big(R_n(s)\big)} - \dfrac{1}{\varLambda(1-s)} = \nu n \\[2ex] \text{and} \\[2ex] \dfrac{1}{\varLambda\big(Q_n\big)} - \dfrac{1}{p_0} = \nu n. \end{cases}$$

Term-by-term subtraction of these equalities produces

$$\frac{1}{\varLambda\big(R_n(s)\big)} - \frac{1}{\varLambda\big(Q_n\big)} = \nu U(s),$$

therefor

$$\nu \varLambda\big(R_n(s)\big) U(s) = 1 - \frac{\varLambda\big(R_n(s)\big)}{\varLambda\left(Q_n\right)} = \mathcal{M}_n(s).$$

Thus we have

$$\mathcal{U}_n(s) = \nu n \varLambda\big(R_n(s)\big) U(s). \tag{4.10}$$

Now a combination of relations (4.4) and (4.10) entails

$$\mathcal{U}_n(s) = U(s) \frac{\nu_n(s) - 1}{\nu_n(s)} \left(1 - \frac{1+\nu}{2} \varLambda(1-s) \frac{\ln \nu_n(s)}{\nu_n(s)} \big(1 + o(1)\big) \right)$$

as $n \to \infty$ and formula (2.4) is immediate.

The theorem is proved completely. □

Proof ***(Proof of Theorem** 3**).*** This statement follows from Theorem 2, according to the continuity theorem for power series. □

5 Conclusion

The paper discusses the asymptotic properties of expansion of GF of the law of evolution of the critical BGW branching system. This statement is called the Basic Lemma. Most of the fundamental results was established on the basis of this lemma. As the theory of branching systems developed, the formulation of this lemma was improved and strengthened, bypassing the finiteness conditions of high-order moments of the offspring law.

This paper presents an improvement to the Basic Lemma 4 on conditions $[f_\varLambda]$ and $[\mathcal{L}_\omega]$. A direct consequence of this lemma is that it contributes to establishing the local limit Theorem 2 in a more refined form.

Moreover, one can find an asymptotic expansion of the function $R'_n(s)$ which is of special interest. Since $\mathsf{p}_1(n) = f'_n(0)$, this expansion involves finding an asymptote for the local probabilities $\mathsf{p}_1(n)$ and therefor we can state the Monotone Ratio Convergence theorem analogue, asserting the fact that the ratios $\mathsf{p}_j(n)/\mathsf{p}_1(n)$ monotonically converge assuming that $p_1 > 0$, i.e.

$$\frac{\mathsf{p}_j(n)}{\mathsf{p}_1(n)} \uparrow \pi_j < \infty \qquad as \quad n \to \infty$$

for all $j \in \mathcal{S}$, where the numbers $\{\pi_j\}$ satisfy the conditions

$$\pi_j = \sum_{k \in \mathcal{S}} \pi_k P_{kj}(1) \qquad and \qquad \sum_{j \in \mathcal{S}} \pi_j < \infty.$$

References

1. Asmussen, S., Hering, H., et al.: Branching Processes, vol. 3. Springer, Heidelberg (1983). https://doi.org/10.1007/978-3-642-65371-1
2. Athreya, K.B., Ney, P.E., Ney, P.: Branching Processes. Courier Corporation, Chelmsford (2004)
3. Bingham, N.H., Goldie, C.M., Teugels, J.L.: Regular Variation, no. 27. Cambridge University Press, Cambridge (1989)
4. Feller, W., et al.: An introduction to probability theory and its applications (1971)
5. Feller, W., et al.: Further remarks on the explicit generating function expression of the invariant measure of critical Galton-Watson branching systems. J. Siber. Fed. Univ.: Math. Phys. (2024, to appear)
6. Harris, T.E., et al.: The Theory of Branching Processes, vol. 6. Springer, Berlin (1963)
7. Imomov, A.A.: On conditioned limit structure of the Markov branching process without finite second moment. Malaysian J. of Math. Sci. **11**(3), 393–422 (2017)
8. Imomov, A.A.: On a limit structure of the Galton-Watson branching processes with regularly varying generating functions. Probab. Math. Stat. **39**(1), 61–73 (2019)
9. Imomov, A.A., Tukhtaev, E.E.: On asymptotic structure of critical Galton-Watson branching processes allowing immigration with infinite variance. Stoch. Model. **39**(1), 118–140 (2023)
10. Jagers, P., et al.: Branching processes with biological applications (1975)
11. Kesten, H., Ney, P., Spitzer, F.: The Galton-Watson process with mean one and finite variance. Theory Probab. Appl. **11**(4), 513–540 (1966)
12. Pakes, A.: Some new limit theorems for the critical branching process allowing immigration. Stoch. Process. Appl. **4**(2), 175–185 (1976)
13. Pakes, A.G.: Revisiting conditional limit theorems for the mortal simple branching process. Bernoulli, pp. 969–998 (1999)
14. Seneta, E.: The Galton-Watson process with mean one. J. Appl. Probab. **4**(3), 489–495 (1967)
15. Seneta, E.: Regularly Varying Functions, vol. 508. Springer, Heidelberg (2006). https://doi.org/10.1007/BFb0079658
16. Sevastyanov, B.A.: Branching Processes. Nauka, Moscow (1971)
17. Slack, R.: A branching process with mean one and possibly infinite variance. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **9**(2), 139–145 (1968)

On Estimation of Structural Parameters in Q-Process

Azam A. Imomov[1,2](✉), Zuhriddin A. Nazarov[2], and Svetlana Moiseeva[3]

[1] Karshi State University, Karshi, Uzbekistan
imomov_azam@mail.ru
[2] Romanovskiy Institute of Mathematics, Tashkent, Uzbekistan
[3] National Research Tomsk State University State University, Tomsk, Russia
smoiseeva@mail.ru

Abstract. In this paper, we examine the population growth system called Q-processes. This is defined by the Galton–Watson Branching system conditioned on non-extinction of its trajectory in the remote future. This work is devoted to a statistical investigation of the properties of Q-processes. The purpose of our paper is to estimate the main parameter – an average number of offspring of one particle and structural parameter. Thus, we find an unbiased estimators for these parameters. We also prove limit theorems for the above unbiased estimates.

Keywords: Branching system · Q-process · Markov chain · Generating function · Extinction time · Transition probabilities · Positive recurrent · Transient · Unbiased estimator · Schröder case · Schröder equation · Kronecker delta · Invariant measure · Stationary measure · Characteristic function

1 Introduction

The study of branching processes has a long history, which, as might be expected, is closely interwoven with a number of applications in the physical and biological sciences. The original problem, which was introduced by Francis Galton in 1873 and first successfully attacked by the Reverend Henry Watson in that year, was in fact concerned with the extinction of family names in the British peerage. For a most enjoyable historical introduction we refer the reader to D. Kendall and for a complete early bibliography to T.E. Harris; see [1,2,6,14].

Branching processes are widely used in mathematical modeling. For example, they may be applied for modeling of virtual machines' (VM) life cycles in a cloud node by using queueing theory methods [4]. After a VM appears in the node, it starts switching randomly between different states generating a stochastic process. So, each VM appearance born a new branch in the global process. Other approaches for modeling the same processes may be found in [16,24,25].

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space on which an array of non-negative integer-valued random variables

$$\left\{ \xi_n^{(i)} : \; n \in \mathsf{N}_0 \quad \text{and} \quad i \in \mathsf{N} \right\}$$

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 241–254, 2024.
https://doi.org/10.1007/978-3-031-65385-8_18

is given, where $\left\{\xi_n^{(i)}\right\}$ are independent and identically distributed with a common probability function $\{p_k,\ k \in \mathsf{N}_0\}$, $\mathsf{N}_0 = \{0\} \cup \mathsf{N}$ and $\mathsf{N} = \{1, 2, \dots\}$.

Definition 1. *The Galton–Watson Branching (GWB) system is a homogeneous-discrete-time Markov chain $\{Z(n),\ n \in \mathsf{N}_0\}$ defined inductively by $Z(0) = 1$ and*

$$Z(n+1) = \begin{cases} \sum\limits_{i=1}^{Z(n)} \xi_{n+1}^{(i)}, & \textit{if} \quad Z(n) > 0, \\ 0, & \textit{if} \quad Z(n) = 0, \end{cases} \tag{1}$$

where $\xi_{n+1}^{(i)}$ is the number of descendants of the n-th particle in the i-th generation and $\mathsf{P}\left\{\xi_n^{(i)} = k\right\} = p_k$.

The variable $Z(n)$ denotes the population size at the moment n in the system. The evolution of the system occurs according to the following mechanism. Each individual lives a unit length lifetime and then gives $k \in \mathsf{N}_0$ descendants with probability p_k. This system is a reducible, homogeneous-discrete-time Markov chain with a state space consisting of two classes: $\mathcal{S}_0 = \{0\} \cup \mathcal{S}$, where $\{0\}$ is absorbing state, and $\mathcal{S} \subset \mathsf{N}$ is the class of possible essential communicating states. To avoid trivialities, we assume that $p_0 > 0$, $p_j \neq 1$ for any $j \in \mathcal{S}_0$. This implies $p_0 + p_1 > 0$ called the Schröder case in which the particle can either die or leave behind 1 offspring. In Böttcher case $(p_0 + p_1 = 0)$, the particle does not die and at the end of its life leaves behind at least 2 generations.

We suppose that $p_0 + p_1 < 1$ and

$$m := \sum_{k \in \mathcal{S}} k p_k < \infty.$$

Definition 2. *The GWB system with offspring mean m is called*

$$\begin{cases} \textit{subcritical}, & \textit{if} \quad m < 1, \\ \textit{critical}, & \textit{if} \quad m = 1, \\ \textit{supercritical}, & \textit{if} \quad m > 1. \end{cases} \tag{2}$$

The GWB system is a Markov chain $\{Z(n),\ n \in \mathsf{N}_0\}$ on the non-negative integers. Its transition function is defined in terms of a given probability function $\{p_k,\ k \in \mathsf{N}_0\}$, $p_k \geq 0$, $\sum_{k \in \mathsf{N}_0} p_k = 1$, by

$$\mathsf{P}\left\{Z(n+1) = j \;\middle|\; Z(n) = i\right\} = \begin{cases} p_j^{*i}, & \text{if} \quad i \geq 0,\ j \geq 0, \\ \delta_{0j}, & \text{if} \quad i = 0,\ j \geq 0, \end{cases} \tag{3}$$

where δ_{ij} is the Kronecker delta and

$$p_j^{*i} := \sum_{j_1+j_2+\cdots+j_i=j} p_{j_1} p_{j_2} \ldots p_{j_i}$$

is the j-th term of the i-th fold convolution of the sequence $\{p_k,\ k \in \mathsf{N}_0\}$.

Considering transition probabilities

$$P_{ij}(n) := \mathsf{P}\left\{Z(n+k) = j \;\middle|\; Z(k) = i\right\} \qquad \textsf{for any} \quad k \in \mathsf{N}_0$$

we observe that the corresponding probability generating function (GF)

$$\mathsf{E}_i s^{Z(n)} := \sum_{k \in \mathcal{S}_0} P_{ij}(n) s^k = \left[f_n(s)\right]^i \qquad \textsf{for any} \quad i \in \mathcal{S}, \tag{4}$$

where $\mathsf{E}_i\left[*\right] = \mathsf{E}_i\left[*\middle| Z(0) = i\right]$, $\mathsf{E}\left[*\right] := \mathsf{E}_1\left[*\right]$ and

$$f_n(s) := \mathsf{E} s^{Z(n)} = \sum_{k \in \mathcal{S}_0} \mathsf{p}_k(n) s^k,$$

therein $\mathsf{p}_k(n) := P_{1k}(n)$ and, $f_n(s)$ is n-fold iteration of the offspring GF $f(s) := \sum_{k \in \mathcal{S}_0} p_k s^k$. Needless to say that $f_n(0) = \mathsf{p}_0(n)$ is a vanishing probability of the system initiated by single individual. Note that $\{\mathsf{p}_0(n)\}$ is monotone and tends to q as $n \to \infty$, which called an extinction probability of the system, i.e. $\lim_{n\to\infty} \mathsf{p}_0(n) = q$. The extinction probability is the smallest non-negative root of the equation $s = f(s)$ on the set $0 < s \leq 1$:

$$q := \mathsf{P}\left\{\lim_{n\to\infty} Z(n) = 0\right\} = 1 - \mathsf{P}\left\{\lim_{n\to\infty} Z(n) = \infty\right\}.$$

It is known that the extinction probability in the subcritical and critical cases $q = 1$ and in the supercritical case $q < 1$; see [1,2,5], [11,13,20,22].

2 The Q-Process

Among the random trajectories of branching systems, there are those that continue a long time. In the case of the GWB model, the class of such trajectories forms another stochastic model called Q-process; see [2,12]. In the case of continuous-time Markov branching systems, an analogous model called the *Markov Q-process*, was first introduced in [10]. Some properties of Q-processes have been studied in [8]. A continuous analogue as a Markov Q-process has been studied [7,9,10,12].

We have previously conditioned the GWB system $Z(n)$ on the event

$$\{n < \mathcal{H} < \infty\},$$

where $\mathcal{H}$ is the extinction time, i.e.

$$\mathcal{H} := \min\{n \in \mathsf{N} : Z(n) = 0\}.$$

It is meaningful, more generally, to condition on $\{n+k<\mathcal{H}<\infty\}$, $k\geq 0$, namely, the event that the process is not extinct at time $n+k$ but does eventually die out. We remark again that when $m\leq 1$ this is the same as conditioning on $\{n+k<\mathcal{H}\}$.

An asymptote of $\mathsf{P}\{\mathcal{H}=n\}$ has been studied in [15,23]. The event $\{n<\mathcal{H}<\infty\}$ represents a condition of $\{Z(n)\neq 0\}$ at the moment n and

$$\{Z(n+k)=0\} \qquad \text{for some} \quad k\in\mathsf{N}.$$

By the extinction theorem $\mathsf{P}_i\{\mathcal{H}<\infty\}=q^i$. Therefore in non-supercritical case

$$\mathsf{P}_i\{n<\mathcal{H}<\infty\}\equiv\mathsf{P}_i\{\mathcal{H}>n\}\to 0.$$

Hence, $Z(n)\longrightarrow 0$ with probability one, so in these cases the process will eventually die out. In [2] can serve as a general source of reference for the above and other classical facts of the theory GWB systems. Let $\mathsf{P}_i\{*\}:=\mathsf{P}\left\{*\,\middle|\,Z(0)=i\right\}$ and we also consider a conditional distribution

$$\mathsf{P}_i^{\mathcal{H}(n)}\{*\}:=\mathsf{P}_i\left\{*\,\middle|\,n<\mathcal{H}<\infty\right\}.$$

The classical limit theorems state that if $q>0$ then under certain moment assumptions the limit $P_{ij}^*(n):=\mathsf{P}_i^{\mathcal{H}(n)}\{Z(n)=j\}$ exists always; see [2, p. 16]. In particular, [21] has proved that if $m\neq 1$ then the set $a_j:=\lim_{n\to\infty}P_{1j}^*(n)$ represents a probability distribution and for $\beta:=f'(q)$ limiting GF $\pi(s)=\sum_{j\in\mathcal{S}}a_js^j$ satisfies to Schröder equation:

$$1-\pi(\varphi(s))=\beta\left[1-\pi(s)\right], \tag{5}$$

where $\pi(s)$ is a probabilistic GF satisfying the above functional equation and

$$\varphi_k(s):=\frac{f_k(qs)}{q} \qquad \text{and} \qquad \varphi(s):=\varphi_1(s) \qquad \text{for any} \quad k\in\mathsf{N}_0. \tag{6}$$

The Eq. (5) determines an invariant property of numbers $\{a_j\}$ with respect to the transition functions $\left\{P_{1j}^*(n)\right\}$ and, the set $\{a_j\}$ is called R-invariant measure with parameter $\mathsf{R}=1/\beta$; see [?]. Investigations have shown that the limiting behaviors of the GWB system trajectory are very sensitive to a change in the classical condition $\{\mathcal{H}>n\}$ to the non-degeneracy condition of the system in the remote future $\{\mathcal{H}=\infty\}$. Apparently, this condition was first used in [5]. In the critical case we know the Yaglom theorem about a convergence of conditional distribution of $2Z(n)/f''(1)n$ given that $\{\mathcal{H}>n\}$ to the standard exponential law. Subsequently, in [17], later in [18,19], [?] and [20], the properties of GWB system were investigated under the condition $\{\mathcal{H}=\infty\}$.

We define conditioned probability measure

$$\mathsf{P}_i^{\mathcal{H}(n+k)}\{*\}:=\mathsf{P}_i\left\{*\,\middle|\,n+k<\mathcal{H}<\infty\right\} \qquad \text{for any} \quad k\in\mathsf{N}.$$

In [5] it was observed that the limit

$$\mathcal{Q}_{ij}(n):=\lim_{k\to\infty}\mathsf{P}_i^{\mathcal{H}(n+k)}\{Z(n)=j\}=\mathsf{P}_i\left\{Z(n)=j\,\middle|\,\mathcal{H}=\infty\right\}$$

always exists. In [2, p. 58] proved, that

$$\lim_{k\to\infty} \mathsf{P}_i^{\mathcal{H}(n+k)}\{Z(n)=j\} = \frac{jq^{j-i}}{i\beta^n} P_{ij}(n). \tag{7}$$

Observe that $\sum_{j\in\mathsf{N}} \mathcal{Q}_{ij}(n) = 1$ for each $i \in \mathsf{N}$. Thus, the probability measure $\mathcal{Q}_{ij}(n)$ can determine a new population growth system with the state space $\mathcal{E} \subset \mathsf{N}$ which we denote by $\{W(n),\ n \in \mathsf{N}_0\}$. This is a discrete-homogeneous-time irreducible Markov chain and called *the Q-process*. Undoubtedly $W(0) \overset{d}{=} Z(0)$ and transition probabilities

$$\mathcal{Q}_{ij}(n) = \mathsf{P}\left\{W(n)=j \;\middle|\; W(0)=i\right\},$$

so that the Q-process can be interpreted as a "long-living" GWB system.

Put into consideration a GF

$$w_n^{(i)}(s) := \sum_{j\in\mathcal{E}} \mathcal{Q}_{ij}(n) s^j. \tag{8}$$

Then from (4) and (7) we obtain

$$w_n^{(i)}(s) = \left[\frac{f_n(qs)}{q}\right]^{i-1} \cdot w_n(s), \tag{9}$$

where the GF $w_n(s) := w_n^{(1)}(s) = \mathsf{E}\left[s^{W(n)} \;\middle|\; W(0)=1\right]$ has a form of

$$w_n(s) = s\frac{f_n'(qs)}{\beta^n} \qquad \text{for all} \quad n \in \mathsf{N}. \tag{10}$$

Using iterations for $f(s)$ in (9) leads to the following functional equation:

$$w_{n+1}^{(i)}(s) = \frac{w(s)}{f_q(s)} w_n^{(i)}\bigl(f_q(s)\bigr), \tag{11}$$

where $w(s) := w_1(s)$. Thus, Q-process is completely defined by setting the GF

$$w(s) = s\frac{f'(qs)}{\beta}. \tag{12}$$

We shall refer to $\{W(n),\ n \in \mathsf{N}_0\}$ as the Q-process associated with $\{Z(n),\ n \in \mathsf{N}_0\}$. It was introduced by F. Spitzer (unpublished) and in [17]. It can be roughly thought of as $Z(n)$ process conditioned on not being extinct in the distant future and on being extinct in the even more distant future. Note that when the original process is aperiodic and irreducible the Q-process is aperiodic and irreducible in the same sense.

Theorem 1. *[2, p. 59.]*

(i) *If $m > 1$ then the Q-process is positive recurrent.*

(ii) *If $m = 1$ then the Q-process is transient.*

(iii) *If $m < 1$ then the Q-process is positive recurrent if and only if*

$$\sum_{k\in\mathsf{N}} (k \log k)\, p_k < \infty.$$

(iv) *In the positive recurrent cases the stationary measure for Q is*

$$\mu_j = j q^{j-1} \nu_j \qquad \textsf{\textit{for}} \quad j \geq 1, \tag{13}$$

where

$$\mu_j := \lim_{n\to\infty} \mathcal{Q}_{ij}(n) = \lim_{n\to\infty} \frac{j q^{j-i}}{i\beta^n} P_{ij}(n) =: j q^{j-1}\nu_j \qquad \textsf{\textit{and}} \qquad \sum_{j\in\mathsf{N}} j q^{j-1}\nu_j = 1.$$

An evolution of the Q-process is in essentially regulated by the structural parameter $\beta > 0$. We can easily see that by Theorem 1 $\mathcal{E}$ is positive recurrent if $\beta < 1$ and $\mathcal{E}$ is transient if $\beta = 1$. On the other hand, it is easy to be convinced that positive recurrent case $\beta < 1$ of Q-process is in a definition character of the non-critical case $m \neq 1$ of the initial GWB system. Note that $\beta \leq 1$ and nothing but. As in the GWB system, the case $m = 1$ plays a special role in the Q-process. Due to its transience, $W(n) \to \infty$ with probability 1, but $W(n)/n$ will converge to a non-degenerate limit law.

3 Moments

Let $\alpha_n := w_n'(1-) < \infty$ and $\alpha := \alpha_1 < \infty$. $w'(1-) < \infty$ and $w''(1-) < \infty$ it is equivalent to that $f''(1-) < \infty$ and $f'''(1-) < \infty$, respectively. The moments of the Q-process, when they exist, can be expressed in terms of the derivatives of $f(s)$ at $s = 1$. It is easy to check for $f(s) = \mathsf{E}s^{\xi}$ that

$$\mathsf{E}\xi = f'(1), \;\; \mathsf{E}\xi\,(\xi - 1) = f''(1) \quad \textsf{and} \quad \textsf{var}\xi = f''(1) + f'(1)\left[1 - f'(1)\right]. \tag{14}$$

Then differentiating (12) on the point $s = 1$ we obtain $\mathsf{E}W(1) = 1 + B_q\bigl(1 - \beta\bigr)$, where

$$B_q := \frac{b_q}{\beta\,(1 - \beta)},$$

and $b_q := q f''(q)$. It follows from (9) and (10) that

$$\mathsf{E}_i W(n) := \mathsf{E}\left[W(n) \;\middle|\; W(0) = i\right] = (i - 1)\,\beta^n + \mathsf{E}W(n), \tag{15}$$

where

$$\mathsf{E}W(n) = \begin{cases} 1 + n b_1, & \textsf{if} \quad \beta = 1, \\ \\ 1 + B_q\bigl(1 - \beta^n\bigr), & \textsf{if} \quad \beta < 1. \end{cases} \tag{16}$$

(15) and (16) formulas show that the following asymptotic relation holds

$$\mathsf{E}_i W(n) \sim \mathsf{E}W(n) \sim \begin{cases} n b_1, & \textsf{if} \quad \beta = 1, \\ \\ 1 + B_q, & \textsf{if} \quad \beta < 1 \end{cases} \tag{17}$$

as $n \to \infty$. From (6) and (10), we have

$$w_n(s) = s \prod_{k=0}^{n-1} \frac{f'\left(f_k(qs)\right)}{\beta} = s \prod_{k=0}^{n-1} \frac{f'\left(q\varphi_k(s)\right)}{\beta} =: s \prod_{k=0}^{n-1} \mathsf{G}\left(\varphi_k(s)\right), \tag{18}$$

where

$$\mathsf{G}(x) = \frac{f'(qx)}{\beta}. \tag{19}$$

Let $\sigma_n^2 := \mathsf{var}W(n)$ and $\sigma^2 := \sigma_1^2$, according to (14) and (18), we have

$$\sigma^2 = \begin{cases} b_q\left(1 - b_q\right) + c_q, & \text{if} \quad \beta = 1, \\ \frac{b_q}{\beta}\left(1 - \frac{b_q}{\beta}\right) + \frac{c_q}{\beta}, & \text{if} \quad \beta < 1, \end{cases} \tag{20}$$

and

$$\sigma_n^2 = \begin{cases} \frac{b_q^2}{2} n^2 - \left(b_q - \frac{3}{2} b_q^2 + c_q\right) n, & \text{if} \quad \beta = 1, \\ \left(1 - \beta^n\right) B_q + \frac{1 - \beta^{2n}}{1 + \beta} C_q + \left(2 - \beta^n - \frac{\beta^{2n+1} - 2\beta^{2n} + 3}{1 + \beta}\right) B_q^2, & \text{if} \quad \beta < 1, \end{cases}$$

where

$$C_q = \frac{c_q}{\beta\left(1 - \beta\right)} \qquad \text{and} \qquad c_q = q^2 f'''(q).$$

We have the following asymptotic relation:

$$\sigma_n^2 \sim \begin{cases} \frac{b_q^2}{2} n^2, & \text{if} \quad \beta = 1, \\ B_q + \frac{1}{1+\beta} C_q - \frac{1-2\beta}{1+\beta} B_q^2, & \text{if} \quad \beta < 1. \end{cases} \tag{21}$$

Also, higher moments can be derived similarly.

4 Main Results

This paper is devoted to a statistical investigation of the properties of Q-processes. Following the classical methods of statistical estimation (see [3]), we introduce the estimator function for the parameters α and β as follows:

$$\widehat{\alpha_n} = W(n+1) - \left[W(n) - 1\right]\beta \qquad \text{and} \qquad \widehat{\beta_n} = \frac{W(n+1) - \alpha}{W(n) - 1}.$$

According to the total probability formula, we have

$$\begin{aligned}
\mathsf{E}\widehat{\alpha_n} &= \mathsf{E}\left[W(n+1) - [W(n)-1]\,\beta\right] = \sum_{k\in\mathsf{N}} \mathsf{E}\left[W(n+1) - [W(n)-1]\,\beta,\, W(n)=k\right] \\
&= \sum_{k\in\mathsf{N}} \mathcal{Q}_{1k}(n)\mathsf{E}\left[W(n+1) - [W(n)-1]\,\beta \,\middle|\, W(n)=k\right] \\
&= \sum_{k\in\mathsf{N}} \mathcal{Q}_{1k}(n)\mathsf{E}\left[W(n+1) \,\middle|\, W(n)=k\right] - \beta\sum_{k\in\mathsf{N}} \mathcal{Q}_{1k}(n)\mathsf{E}\left[W(n)-1 \,\middle|\, W(n)=k\right] \\
&= \sum_{k\in\mathsf{N}} \mathcal{Q}_{1k}(n)\mathsf{E}\left[1+\zeta_1+\zeta_2+\cdots+\zeta_{k-1}+\eta\right] - \beta\sum_{k\in\mathsf{N}} \mathcal{Q}_{1k}(n)(k-1) \\
&= \sum_{k\in\mathsf{N}} \mathcal{Q}_{1k}(n)\left[1+(k-1)\beta+(\alpha-1)\right] - \beta\sum_{k\in\mathsf{N}} k\mathcal{Q}_{1k}(n) + \beta = \alpha \qquad (22)
\end{aligned}$$

and if α exists

$$\begin{aligned}
\mathsf{E}\widehat{\beta_n} &= \mathsf{E}\left[\frac{W(n+1)-\alpha}{W(n)-1}\right] = \sum_{k\in\mathsf{N}} \mathsf{E}\left[\frac{W(n+1)-\alpha}{W(n)-1},\, W(n)=k\right] \\
&= \sum_{k\in\mathsf{N}} \mathcal{Q}_{1k}(n)\mathsf{E}\left[\frac{W(n+1)-\alpha}{W(n)-1} \,\middle|\, W(n)=k\right] \\
&= \sum_{k\in\mathsf{N}} \frac{\mathcal{Q}_{1k}(n)}{k-1}\mathsf{E}\left[W(n+1)-\alpha \,\middle|\, W(n)=k\right] \\
&= \sum_{k\in\mathsf{N}} \frac{\mathcal{Q}_{1k}(n)}{k-1}\mathsf{E}\left[1+\zeta_1+\zeta_2+\cdots+\zeta_{k-1}+\eta-\alpha\right] \\
&= \sum_{k\in\mathsf{N}} \frac{\mathcal{Q}_{1k}(n)}{k-1}\left[1+(k-1)\beta+(\alpha-1)-\alpha\right] = \beta, \qquad (23)
\end{aligned}$$

where to calculate $\mathsf{var}\left[W(1)\middle| W(0)=k\right]$, from (9) and (10) we write

$$\begin{aligned}
w^{(k)}(s) := w_1^{(k)}(s) &= \mathsf{E}\left[x^{W(1)}\middle| W(0)=k\right] \\
&= \varphi^{k-1}(s)\cdot w(s) = s\varphi^{k-1}(s)\mathsf{G}(s). \qquad (24)
\end{aligned}$$

It can be seen that the relation

$$W(1) = 1+\zeta_1+\zeta_2+\cdots+\zeta_{k-1}+\eta$$

corresponds to the latter, where ζ_k independent and the variable η does not depend from all ζ_k. We have

$$f_{\zeta_1}(s) := \mathsf{E}s^{\zeta_1} = \varphi(s) \qquad \text{and} \qquad f_\eta(s) := \mathsf{E}s^{\eta} = \mathsf{G}(s).$$

Easy to find

$$\mathsf{var}\left[W(1) \,\middle|\, W(0)=k\right] = (k-1)\mathsf{var}\zeta_1 + \mathsf{var}\eta, \qquad (25)$$

where

$$\mathsf{var}\zeta_1 = f''_{\zeta_1}(1-) + f'_{\zeta_1}(1-) - \left[f'_{\zeta_1}(1-)\right]^2 = \beta(\alpha-\beta), \tag{26}$$

$$\mathsf{var}\eta = f''_{\eta}(1-) + f'_{\eta}(1-) - \left[f'_{\eta}(1-)\right]^2 = w''(1-) - \alpha(\alpha-1) \tag{27}$$

and from (25) and (27), we have

$$\mathsf{var}\left[W(1) \,\middle|\, W(0)=k\right] = (k-1)\mathsf{V}_1 + \mathsf{V}_2,$$

$$\mathsf{V}_1 = \beta(\alpha-\beta) \quad \mathsf{and} \quad \mathsf{V}_2 = w''(1-) - \alpha(\alpha-1). \tag{28}$$

Thus, the estimates $\widehat{\alpha_n}$ and $\widehat{\beta_n}$ are unbiased estimators for parameter α and β, respectively.

Throughout the paper we will use famous Landau symbols o, $\mathcal{O}$ and $\mathcal{O}^*$ to describe kinds of bounds on asymptotic varying rates of positive functions $f(x)$ and $g(x)$. So, $f = o(g)$ means that $\lim_x f(x)/g(x) = 0$, and we write $f = \mathcal{O}(g)$ if $\limsup_x f(x)/g(x) < \infty$ and also we write $f = \mathcal{O}^*(g)$ if the ratio $f(x)/g(x)$ has a positive explicit limit. i.e. $\lim_x f(x)/g(x) = C < \infty$. Moreover, $f(x) \sim g(x)$ means that $\lim_x f(x)/g(x) = 1$.

The following theorems characterize the proposed estimates.

Theorem 2. *Let $w''(1-) < \infty$.*

- *If $\beta < 1$, then*

$$\mathsf{Var}\widehat{\alpha_n} \sim \mathsf{V}_1 B_q + \mathsf{V}_2 \qquad \mathsf{as} \quad n \to \infty, \tag{29}$$

  where V_1 and V_2 are defined in (28)
- *If $\beta = 1$, then*

$$\mathsf{Var}\widehat{\alpha_n} = \mathcal{O}^*(n) \qquad \mathsf{as} \quad n \to \infty. \tag{30}$$

Theorem 3. *Let $w''(1-) < \infty$ and α exists.*

- *If $\beta < 1$, then*

$$\mathsf{Var}\widehat{\beta_n} = \mathsf{K} \cdot \left[\mathsf{V}_1 + \mathsf{V}_2 \cdot \int_0^1 \frac{\pi(x)}{x} dx\right] (1 + o(1)) \qquad \mathsf{as} \quad n \to \infty, \tag{31}$$

- *If $\beta = 1$, then*

$$\frac{n}{2}\mathsf{Var}\widehat{\beta_n} \sim 1 + \varepsilon_n \qquad \mathsf{as} \quad n \to \infty, \tag{32}$$

where K – positive constant and

$$\varepsilon_n := \frac{2\mathsf{V}_2}{(\alpha-1)^2} \cdot \frac{\ln n}{n}.$$

5 Proof of Theorems

Proof of Theorem 2.

For the estimate $\widehat{\alpha_n}$, according to the total probability formula, we find its variance:

$$\begin{aligned}\mathsf{Var}\widehat{\alpha_n} &= \mathsf{Var}\left[W(n+1)-[W(n)-1]\beta\right] = \mathsf{E}\left[\widehat{\alpha_n}-\mathsf{E}\widehat{\alpha_n}\right]^2 = \mathsf{E}\left[\widehat{\alpha_n}-\alpha\right]^2 \\ &= \sum_{k\in\mathsf{N}}\mathcal{Q}_{1k}(n)\cdot\mathsf{E}\left[(W(n+1)-[W(n)-1]\beta-\alpha)^2 \,\middle|\, W(n)=k\right].\end{aligned} \tag{33}$$

Further, due to the homogeneity of the Q-process, we obtain

$$\begin{aligned}\mathsf{Var}\widehat{\alpha_n} &= \sum_{k\in\mathsf{N}}\mathcal{Q}_{1k}(n)\cdot\mathsf{E}_k\left[W(1)-[W(0)-1]\beta-\alpha\right]^2 \\ &= \sum_{k\in\mathsf{N}}\mathcal{Q}_{1k}(n)\cdot\mathsf{E}_k\left[W(1)-\alpha\right]^2 = \sum_{k\in\mathsf{N}}\mathcal{Q}_{1k}(n)\cdot\mathsf{E}_k\left[W(1)-\mathsf{E}_1W(1)\right]^2 \\ &= \sum_{k\in\mathsf{N}}\mathcal{Q}_{1k}(n)\cdot\mathsf{Var}_kW(1) = \sum_{t\in\mathsf{N}_0}\mathcal{Q}_{1t+1}(n)\cdot\mathsf{Var}_{t+1}W(1).\end{aligned} \tag{34}$$

where $t=k-1$,

$$\mathsf{Var}_iW(n) = \mathsf{Var}\left[W(n) \,\middle|\, W(0)=i\right] \quad \text{and} \quad \mathsf{Var}_1W(n) := \mathsf{Var}W(n). \tag{35}$$

From (25)–(27), we have

$$\mathsf{Var}\widehat{\alpha_n} = \mathsf{V}_2 + \mathsf{V}_1\cdot\mathcal{I}_1(t), \tag{36}$$

where

$$\mathcal{I}_1(t) := \sum_{t\in\mathsf{N}_0} t\mathcal{Q}_{1t+1}(n).$$

It is easy to see that by the definition of the Q-process and in turn, from the form (8) and (16), we have

$$\mathcal{I}_1(t) = \begin{cases} nb_1, & \text{if} \quad \beta=1, \\ \\ (1-\beta^n)B_q, & \text{if} \quad \beta<1. \end{cases} \tag{37}$$

The last relation proves the theorem. Theorem 2 is proved.

Proof of Theorem 3.

$$\begin{aligned}\mathsf{Var}\widehat{\beta_n} &= \mathsf{Var}\left[\frac{W(n+1)-\alpha}{W(n)-1}\right] = \mathsf{E}\left[\widehat{\beta_n} - \mathsf{E}\widehat{\beta_n}\right]^2 = \mathsf{E}\left[\widehat{\beta_n} - \beta\right]^2 \\ &= \mathsf{Var}\left[\frac{W(n+1)-\alpha}{W(n)-1} - \beta\right] = \sum_{k\in\mathsf{N}} \mathcal{Q}_{1k}(n)\cdot \mathsf{E}_k\left[\left(\frac{W(n+1)-\alpha}{W(n)-1} - \beta\right)^2\right] \\ &= \sum_{k\in\mathsf{N}} \mathcal{Q}_{1k}(n)\cdot \mathsf{E}_k\left[\left(\frac{W(n+1)-\alpha-[W(n)-1]\,\beta}{W(n)-1}\right)^2\right] \\ &= \sum_{k\in\mathsf{N}} \frac{\mathcal{Q}_{1k}(n)}{(k-1)^2}\cdot \mathsf{E}_k\left[W(1) - \mathsf{E}_1 W(1)\right]^2 \\ &= \sum_{t\in\mathsf{N}_0} \frac{\mathcal{Q}_{1t+1}(n)}{t^2}\cdot \mathsf{Var}_{t+1}W(1). \end{aligned} \tag{38}$$

From (25)–(27), we have

$$\mathsf{Var}\widehat{\beta_n} = \mathsf{V}_1\cdot\varSigma_1 + \mathsf{V}_2\cdot\varSigma_2, \tag{39}$$

where

$$\varSigma_1 := \sum_{t\in\mathsf{N}_0}\frac{\mathcal{Q}_{1t+1}(n)}{t} \qquad \text{and} \qquad \varSigma_2 := \sum_{t\in\mathsf{N}_0}\frac{\mathcal{Q}_{1t+1}(n)}{t^2}. \tag{40}$$

It is easy to see that by the definition of the Q-process and in turn, from the form (9), we have

$$\varSigma_1 = \int_0^1 \frac{w_n(x)}{x}dx = \frac{1}{\beta^n}\left(1 - f_q^{(n)}(0)\right),$$

where $f_q^{(n)}(s) := f_n(qs)/q$. It is almost obvious that the GF $f_q^{(n)}(s)$ is equal to the n-fold iteration $f_q(s) = f(qs)/q$. The latter generates a subcritical GWB system, i.e. $f'(1) = \beta < 1$. In this case, it is known that $1 - f_q^{(n)}(0) \sim \mathsf{K}\beta^n$ as $n\to\infty$, where K is a positive constant; see [22, p. 56]. Hence

$$\varSigma_1 \to \mathsf{K} \qquad \text{as} \quad n\to\infty. \tag{41}$$

Then we have

$$\varSigma_2 = \sum_{t\in\mathsf{N}_0}\frac{\mathcal{Q}_{1t+1}(n)}{t}\cdot\int_0^1 x^{k-1}dx = \int_0^1 \frac{1}{x}\left[\sum_{k\in\mathsf{N}_0}\frac{\mathcal{Q}_{1t+1}(n)}{t}x^k\right]dx.$$

Not difficult to get using (6)

$$\sum_{k\in\mathsf{N}_0}\frac{\mathcal{Q}_{1t+1}(n)}{t}x^k = \frac{f_q^{(n)}(x) - f_q^{(n)}(0)}{\beta^n}.$$

From the last two equalities we have

$$\varSigma_2 = \int\limits_0^1 \frac{1}{x} \cdot \frac{f_q^{(n)}(x) - f_q^{(n)}(0)}{\beta^n} dx = \frac{1 - f_q^{(n)}(0)}{\beta^n} \cdot \int\limits_0^1 \frac{f_q^{(n)}(x) - f_q^{(n)}(0)}{x\left(1 - f_q^{(n)}(0)\right)} dx. \tag{42}$$

In the monograph of [2, p. 16] it is proved that the integrand

$$\left[f_q^{(n)}(s) - f_q^{(n)}(0)\right] \Big/ \left[1 - f_q^{(n)}(0)\right]$$

on the right side of equality (28) converges to the GF $\pi(x)$ satisfying equation (5). According to the latter and, again, the relation $1 - f_q^{(n)}(0) \sim \mathsf{K}\beta^n$, from (28) we obtain

$$\varSigma_2 \sim \mathsf{K} \cdot \int\limits_0^1 \frac{\pi(x)}{x} dx \qquad \text{as} \quad n \to \infty. \tag{43}$$

Statement (31) will now be obtained from relations (39)–(43).

Let us now prove relation (32).

In this case, $q = 1$ and, arguing similarly to the case $\beta < 1$, we find

$$\sigma_\beta^2 = (\alpha - 1) \cdot \int\limits_0^1 \frac{w_n(x)}{x} dx + \mathsf{V}_2 \cdot \int\limits_0^1 \frac{f_q^{(n)}(x) - f_q^{(n)}(0)}{x} dx. \tag{44}$$

Let us initially estimate the first integral. According to (9), we have.

$$\int\limits_0^1 \frac{w_n(x)}{x} dx = \int\limits_0^1 \left(f_q^{(n)}(x)\right)' dx \sim \frac{2}{(\alpha - 1)n} \qquad \text{as} \quad n \to \infty. \tag{45}$$

This takes into account the fact that $1 - f_q^{(n)}(0) \sim 2/(\alpha - 1)n$; see [2, p. 19]. We write the second integral in the form

$$\int\limits_0^1 \frac{f_q^{(n)}(x) - f_q^{(n)}(0)}{x} dx = \left(1 - f_q^{(n)}(0)\right) \int\limits_0^1 \frac{f_q^{(n)}(x) - f_q^{(n)}(0)}{x\left(1 - f_q^{(n)}(0)\right)} dx. \tag{46}$$

In the monograph of [3, p. 10] it is proved that under our conditions and notation,

$$\int\limits_0^1 \frac{f_q^{(n)}(x) - f_q^{(n)}(0)}{x\left(1 - f_q^{(n)}(0)\right)} dx \sim \frac{2\ln n}{(\alpha - 1)n} \qquad \text{as} \quad n \to \infty. \tag{47}$$

Considering relations (45)–(47) together, taking into account $1 - f_q^{(n)}(0) \sim 2/(\alpha - 1)n$, we complete the proof of assertion (32). Theorem 3 is proved.

Remark 1. In Theorem 3, the variable K is the Kolmogorov constant, which is equal to the product of

$$\prod_{n\in\mathsf{N}_0}\frac{q-f\left(f_n(0)\right)}{\beta\left(q-f_n(0)\right)},$$

see [22, p. 56]. Furthermore, note that the right side of (16) remains bounded, i.e.

$$\sigma_{\beta}^{2}=\mathcal{O}(1)\qquad\text{as}\quad n\to\infty.$$

6 Conclusion

As noted above, we call Q-processes the class of trajectories determined by the immortal Galton-Watson branching system in the remote future. In this paper, we present the numerical characteristics of the Q-process. We also provided the statistical estimators of the structural parameters of Q-processes and showed that these estimates are unbiased. In addition, limit theorems for the variances of the proposed estimators are proved.

In our future work, we intend to propose analogous estimators in the case of continuous-time Markov Q-processes.

References

1. Asmussen, S., Hering, H., et al.: Branching processes, vol. 3. Springer (1983)
2. Athreya, K.B., Ney, P.E., Ney, P.: Branching processes. Courier Corporation (1972)
3. Badalbaev, I.S., Mukhitdinov, A.A.: Statistical methods multitype branching processes (1990)
4. Fedorova, E., Lapatin, I., Lizyura, O., Moiseev, A., Nazarov, A., Paul, S.: Mathematical modeling of virtual machine life cycle using branching renewal process, pp. 29–39 (2022)
5. Harris, T.E.: Some mathematical models for branching processes. In: Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, vol. 2, pp. 305–329. University of California press (1951)
6. Harris, T.E., et al.: The Theory of Branching Processes, vol. 6. Springer, Berlin (1963)
7. Imomov, A.A.: A differential analog of the main lemma of the theory of Markov branching processes and its applications. Ukr. Math. J. **57**(2), 307–315 (2005)
8. Imomov, A.A.: On conditioned limit structure of the Markov branching process without finite second moment. Malaysian J. Math. Sci. **11**(3), 393–422 (2017)
9. Imomov, A.A., Nazarov, Z.A.: On structural parameter estimation of the Markov q-process. Bull. Inst. Math. **5**(1), 44–55 (2022)
10. Imomov, A.: On Markov continuous time analogue of q-processes. J. Theory Probability Math. Stat. **84**, 57–64 (2012)
11. Imomov, A.A., Nazarov, Z.A.: Limit theorems for the positive recurrent q-process. In: International Conference on Information Technologies and Mathematical Modelling, pp. 1–15. Springer (2022)
12. Imomov, A.A.: On long-term behavior of continuous-time Markov branching processes allowing immigration. J. Siberian Federal Univ. Math. Phys. **7**(4), 443–454 (2014)

13. Jagers, P., et al.: Branching processes with biological applications (1975)
14. Kendall, D.G.: Branching processes since 1873. J. Lond. Math. Soc. **1**(1), 385–406 (1966)
15. Kesten, H., Ney, P., Spitzer, F.: The Galton-Watson process with mean one and finite variance. Theory Probability Appl. **11**(4), 513–540 (1966)
16. Kumar, R., Soodan, B.S., Kuaban, G.S., Czekalski, P., Sharma, S.: Performance analysis of a cloud computing system using queuing model with correlated task reneging. J. Phys. Conf. Ser. **2091**, 012003. IOP Publishing (2021)
17. Lamperti, J., Ney, P.: Conditioned branching processes and their limiting diffusions. Theory Probability Appl. **13**(1), 128–139 (1968)
18. Pakes, A.: Some new limit theorems for the critical branching process allowing immigration. Stochastic Processes Appl. **4**(2), 175–185 (1975)
19. Pakes, A.G.: Some limit theorems for the total progeny of a branching process. Adv. Appl. Probab. **3**(1), 176–192 (1971)
20. Pakes, A.G.: Critical Markov branching process limit theorems allowing infinite variance. Adv. Appl. Probab. **42**(2), 460–488 (2010)
21. Seneta, E.: Functional equations and the Galton-Watson process. Adv. Appl. Probab. **1**(1), 1–42 (1969)
22. Sevastyanov, B.A.: Branching Processes. Nauka, Moscow (1971)
23. Slack, R.: A branching process with mean one and possibly infinite variance. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **9**(2), 139–145 (1968)
24. Vetha, S., Devi, K.V.: Dynamic resource allocation in cloud using queueing model. J. Ind. Pollut. Control **33**(2), 1547–1554 (2017)
25. Vilaplana, J., Solsona, F., Teixidó, I., Mateo, J., Abella, F., Rius, J.: A queuing theory model for cloud computing. J. Supercomput. **69**, 492–507 (2014)

Workshop on Retrial Queues

Retrial Queueing System of $MAP/PH/N$ Type with a Finite Buffer and Group Service. The Process Describing the System Dynamics

Alexander Dudin(✉) and Olga Dudina

Department of Applied Mathematics and Computer Science, Belarusian State University, 220030 Minsk, Belarus
dudin@bsu.by, dudina@bsu.by

Abstract. A queuing system with many identical servers, a finite buffer, and order retrials is under study. The order's arrival is described by the Markov arrival process. Service is offered for orders in groups. The size of the group is bounded from above by the capacity of the buffer and from below by a fixed threshold. A group's service time follows a phase-type distribution with the irreducible representation determined by group size. A classical retrial strategy is assumed. Retrying and waiting orders can renege from the service after a random time interval. The duration of this interval has an exponential distribution. We built a multidimensional continuous-time Markov chain that includes the number of retrying orders, the number of busy servers, the state of the underlying arrivals process, and an auxiliary multidimensional Markov chain that defines the number of servers providing service at all possible service phases. The infinitesimal generator of the constructed chain is written down and, and the explicit expressions for the matrix blocks of the generator are presented.

Keywords: Multi-server retrial queue · group service · Markov arrival process · phase-type distribution

1 Introduction

The group servicing of orders is the major distinctive aspect of the queueing model discussed in this study. Such a kind of service is typical for a variety of real-world systems, including various manufacturing, transportation, and telecommunication systems, in which, from an economic point of view, it is reasonable to provide not an individual but a group service. Concrete examples of such systems given in [1] are as follows:

(a) Processing jobs, such as flushing with the same coolant, coating bricks with precious metals (dipped in liquid concentrate), sandblasting, and heat treatment, can all be done in groups.

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 257–271, 2024.
https://doi.org/10.1007/978-3-031-65385-8_19

(b) Certain hazardous petrochemical and petroleum wastes may require a specific treatment procedure, such as a thermal treatment method that uses high temperatures to break down the hazardous compounds into simpler, less poisonous forms. These might be dealt with in groups.
(c) On machine vision systems, tasks arriving for processing may all have the same features; therefore, all jobs may be placed on a similar tray or belt for the camera to take the pictures and deliver the information.
(d) Jobs requiring processes in which the beam from one laser is divided into numerous beams, one for each task present, can be processed in groups using industrial lasers.

These examples may be supplemented by the various systems of goods and food delivery, see, e.g., [2], in which the delivered items are preliminary packed into containers or pallets of a finite capacity, and various transportation systems where a server (bus, minivan, car, plane, ferry, etc.) should have enough passengers to justify the operational costs (fuel, salary for the staff, taxes and fees, etc.).

Due to their wide applicability, investigation of queueing systems with a group (bulk, batch, etc.) order service started back in the 1950s; see, e.g., [3–9]. The state of the art in queueing system analysis with a group service is represented, for example, in articles [10–18]. The essential disadvantage of many known results consists of the imposed assumption that the arrivals of orders occur according to the stationary Poisson process. This assumption simplifies the study of a queueing system significantly. However, it rarely holds true in a variety of real-world systems where flows are bursty and may change the arrival rate due to certain reasons, e.g., the time of a day or a week, season, weather conditions, etc. When the real flow has a more or less essential positive correlation (the lag-1 coefficient of correlation is more than 0.1), this assumption may result in significant inaccuracies in evaluating the primary performance measures of an actual queueing system and wrong managerial decisions.

A more adequate model of real flows, the so-called versatile arrival flow, was offered by M. Neuts, see [19]. In [20], this flow was named by V. Ramaswami as N flow. A detailed study of the $N/G/1$ queue is implemented there. In [21, 22], this flow was renamed by D. Lucantoni into the $BMAP$ (Batch Markov arrival process). The $BMAP$ assumes that the orders may arrive in batches. The Markov arrival process (MAP) is the particular case of the $BMAP$ in which orders can arrive only one by one. Useful information about the $BMAP$ and MAP can be found in [21–27]. The challenge of fitting real-world flows by the MAP is extensively discussed in the literature; see, for example, [28–30].

Queueing systems with the MAP flow of orders and service provisioning in groups have been analysed, e.g., in articles by S. Chakravarthy and coauthors; see, e.g., [1,2,16–18,31–41]. The papers by other authors on this topic are as follows: [13,42–51].

It is worth noting that almost all these papers deal with single-server queueing models. Queues with many servers, a group service, input buffers, and the MAP were considered in [34,35], where systems with an infinite and a finite capacity of the buffer were considered, correspondingly, in [36], where the system similar to [35] but with a varying number of servers is under study. In paper [37], a two-server queue was considered. The distribution of a group service time is assumed to be exponential. Recently, an essentially more generic multi-server model with the PH distribution of group service times dependent on group size was investigated in [52].

Besides the group service and MAP flow, one more essential distinguishing feature of the queueing system considered here is the following assumption. The arriving orders that met all servers busy and the buffer full were not lost but made repeated attempts to enter the service later on. Queueing models that account for the possibility of repeated attempts are called retrial queues. In the literature, it is said that retrying orders stay in the orbit. In contrast to a buffer, which usually has a definite physical meaning, e.g., some area in computer memory, and the orders (requests, messages, packets, etc.) are temporarily stored in this memory, the orbit is a virtual place.

The main difference between queueing systems with a buffer and an orbit is the following. In systems with a buffer, the server permanently knows the state of the buffer, and vice versa. The server, which completes the service, immediately begins a new service if the buffer is not empty. The service time is usually not dependent on the number of orders in a buffer. In systems with an orbit and without a buffer, the server always remains non-busy during a certain time. This time finishes with the arrival of a new (primary) order from outside or by the retrial of one of the orders staying in the orbit. The likelihood that the idle time will be terminated by the retrial is, in general, dependent on the current number of orders in the orbit. As a result, the stochastic process that describes the system's behavior is state-inhomogeneous. This explains why the analysis of the systems with retrials is significantly more difficult than the study of the respective queueing systems with buffers. This explains why the literature in retrial queues is much poorer. We can mention only two monographs [53,54] in English in this field.

Multi-server queues of $BMAP/PH/N$ type with orders retrial and individual service were investigated, e.g., in [55–60]. It was assumed in these papers that the total order retrial rate depends on the number of orbiting orders. This assumption allows us to consider the systems with an even more general, than the classical, retrial strategy. The classical strategy assumes that each order residing in the orbit generates retrials with a constant rate, say, α independently of other orders. Hence, if the current orbiting order number is, say, i, then the total retrial rate is equal to $i\alpha$.

As far as we know, the retrial queues with many servers and group service of the orders were considered only in papers [17,38]. In both of these papers, it was assumed that the retrial rate is permanent, irrespective to the number of orbiting orders. Such a suggestion essentially reduces the complexity of the system's analysis. However, it is not realistic in many real-world systems where the orders make the retrials independently of each other. The model under consideration in this study is the first in the literature of a multi-server retrial queue with group servicing of orders and a total retrial rate depending on the number of orders in the orbit. Additional advantages of this model compared to those existing in the literature are: (i) we suppose a phase-type distribution of a group processing time that is essentially more general than the exponential distribution in the vast majority of publications; (ii) we allow dependence of a group service time on the size of a group; and (iii) we suppose that the orders staying in the buffer and in the orbit have a limited impatience time and can renege from the system without service receiving. It is critical to take order impatience into account when modeling real-world systems, see [46,61]. But this complicates the analysis of a system.

The following is the structure of the results presentation. In Sect. 2, a mathematical model is formulated. Section 3 introduces a multidimensional Markov process that describes the system's behavior. Its generator is derived there. Further investigation of the model, including the proof of the ergodicity conditions of this process in cases when the orders are patient and impatient during the stay in the orbit, calculation of the stationary distribution and the system's major performance metrics, illustration of the feasibility of the presented results, and impact of the capacity of the buffer and the minimal size of a group on the performance characteristics of the system, is presented in [62].

2 Mathematical Model

We consider a queueing system having N identical servers and a finite input buffer of size K. The scheme of its system operation is depicted in Fig. 1.

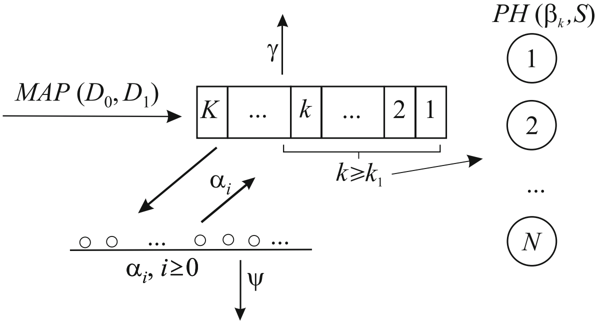


Fig. 1. The scheme of the system operation

Orders arrival is defined by the MAP. This arrival process is determined by the continuous-time Markov chain ν_t, $t \geq 0$, having a finite state space $\{1, 2, ..., W\}$ and an irreducible generator denoted by $D(1)$. This generator is represented as the sum of the non-negative matrix D_1 and matrix D_0. The entries of the matrix D_1 determine the intensities of transitions of the chain ν_t

that are accompanied by the arrival of an order. The intensity of the corresponding transition of the chain ν_t without the arrival of an order is determined by the non-diagonal elements of the matrix D_0, and the intensity of the process ν_t exit from the corresponding state is determined by the modules of the negative diagonal elements.

The average order rate λ is calculated as $\lambda = \boldsymbol{\theta} D_1 \mathbf{e}$ where $\boldsymbol{\theta}$ is an invariant probability row vector of the Markov chain ν_t. This vector is found from the system of the linear algebraic equations $\boldsymbol{\theta} D(1) = \mathbf{0},\ \boldsymbol{\theta}\mathbf{e} = 1$. Here, $\mathbf{0}$ is a row vector of acceptable size made up of zeros, and $\mathbf{e}$ is a column vector of proper size made up of ones.

Any of the N servers may provide service to a group of orders consisting of at least k_1 orders. Thus, the parameter k_1, $1 \leq k_1 \leq K$, determines the minimal size of the group to which service can be provided. The maximal size of the group taken for service coincides with the buffer capacity K.

If the number of orders in the buffer is less than or equal to $k_1 - 2$ when the order arrives, the incoming order is buffered and awaiting service. If the number of orders in the buffer is $k_1 - 1$ and a server is available, the complete group of size k_1 is selected for servicing. If all servers are busy, the order is added to the buffer if it is not already full.

If the buffer is full, the order is placed in a virtual place called orbit. The capacity of the orbit is infinite. An order residing in the orbit (orbiting order) repeats attempts to enter the buffer, regardless of other orders in the orbit, at random time intervals. If at an arbitrary moment the number of orbiting orders is i, $i > 0$, then the inter-retrial time from the orbit is exponential with the rate $\alpha_i = i\alpha,\ i \geq 0$, where α is an individual retrial rate of an order. Note that all the presented results are extendable to a more general case when the dependence of α_i on i is arbitrary with the limiting condition $\lim\limits_{i\to\infty} \alpha_i = \infty$. An attempt is considered successful if it finds a free space in the buffer. The retrying order immediately moves from the orbit to the buffer.

If at the end of a group servicing by some server the number of orders in the buffer is not less than k_1, then all orders from the buffer are taken for servicing by this server immediately.

We suggest that the group service time has a phase-type (PH) distribution, specified by a Markov chain m_t, $t \geq 0$, with the set $\{1, 2, \ldots, M\}$ of the transient states and an absorbing state $M + 1$. The irreducible representation of the PH distribution of service of a group consisting of k orders is the pair of a stochastic row vector and sub-generator $(\boldsymbol{\beta}_k, S)$, $k = \overline{k_1, K}$. The average service time for a size k group is $b_1^{(k)} = \boldsymbol{\beta}_k(-S)^{-1}\mathbf{e}$. It is worth noting that by assuming that the starting probability vector of service time relies on the size of the group, we account for the service process's dependency on the size of the group.

The orders staying in the buffer are impatient and can renege from the system without service. Each order reneges after a time interval length of which has the exponential distribution with the parameter $\gamma \geq 0$. Orders staying in the orbit are also impatient and leave the system without service after an interval that has the exponential distribution with the parameter ψ, $\psi \geq 0$.

3 The Markov Process Describing Behavior of the System and Its Generator

The state of the considered system at an arbitrary time moment $t,\ t\geq 0$, is completely defined if the states of the following processes are known: (i) the number i_t of orders in the orbit, $i_t\geq 0$; (ii) the number k_t of orders in the buffer, $k_t=\overline{0,K}$; (iii) the number n_t of busy servers, $n_t=\overline{0,N}$; (iv) the state of the underlying process ν_t of the MAP, $\nu_t=\overline{1,W}$; (v) the number $m_t^{(l)}$ of servers on the l-th phase of service, $m_t^{(l)}=\overline{0,n_t}$, $l=\overline{1,M}$, $\sum\limits_{l=1}^{M}m_t^{(l)}=n_t$.

Thus, analysis of the considered queueing system reduces to consideration of the $M+4$-dimensional is described Markov chain

$$\xi_t=\{i_t,\,k_t,\,n_t,\,\nu_t,\,m_t^{(1)},\dots,\,m_t^{(M)}\},\ t\geq 0.$$

Evidently, this chain is a regular and irreducible.

Let us renumerate the states of the Markov chain ξ_t in the component's $(i_t,\ k_t,\ n_t,\ \nu_t)$ direct lexicographical order and the components $(m_t^{(1)},\dots,\,m_t^{(M)})$ reverse lexicographical order. We refer to the set of states of the Markov chain having the value i of the first component as level $i,\ i\geq 0$. The set of states of the Markov chain having the values (i,k) of the first and second components is called macrostate $(i,k),\ i\geq 0,\ k=\overline{0,K}$.

Theorem 1. *The generator Q of the Markov chain $\xi_t,\ t\geq 0,$ has the block tridiagonal structure shown below:*

$$Q=\begin{pmatrix}Q_{0,0} & Q_{0,1} & O & O & O & \dots\\ Q_{1,0} & Q_{1,1} & Q_{1,2} & O & O & \dots\\ O & Q_{2,1} & Q_{2,2} & Q_{2,3} & O & \dots\\ O & O & Q_{3,2} & Q_{3,3} & Q_{3,4} & \dots\\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots\end{pmatrix}$$

where the non-zero matrices $Q_{i,j},\ |i-j|\leq 1,$ combined by the transition intensities from level i to level j are specified as follows:

1. *the diagonal blocks* $Q_{i,i}$, $i \geq 0$, *have the form* $Q_{i,i} = (Q_{i,i})_{k,k'}$, $k, k' = \overline{0, K}$, *where the non-zero blocks* $(Q_{i,i})_{k,k'}$ *are given as*

$$\begin{gathered}
(Q_{i,i})_{k,k} = \mathrm{diag}\{D_0, D_0 \oplus (A_n + \varDelta_n),\, n = \overline{1,N}\} \\
-(k\gamma + (\alpha+\psi)i) I_{W \sum\limits_{n=0}^{N} T_n} + \mathrm{diag}^-\{I_W \otimes L_n,\, n = \overline{1,N}\} \\
+\delta_{k_1,1}\mathrm{diag}^+\{D_1 \otimes P_n(\boldsymbol{\beta}_1),\, n = \overline{0,N-1}\},\ k = \overline{0,k_1-1}, \\
(Q_{i,i})_{k,k} = D_0 \oplus (A_N + \varDelta_N) \\
-(k\gamma + (\alpha+\psi)i) I_{WT_N} + \delta_{k,K}\alpha i I_{WT_N},\ k = \overline{k_1, K}, \\
(Q_{i,i})_{k,k+1} = \mathrm{diag}\{D_1 \otimes I_{T_n},\, n = \overline{0,N}\},\ k = \overline{0,k_1-2}, \\
(Q_{i,i})_{k_1-1,k_1} = \begin{pmatrix} O_{W \sum\limits_{n=0}^{N-1} T_n \times WT_N} \\ \otimes I_{T_N} \end{pmatrix}, \\
(Q_{i,i})_{k,k+1} = D_1 \otimes I_{T_N},\ k = \overline{k_1, K-1}, \\
(Q_{i,i})_{k,k-1} = k\gamma I_{W \sum\limits_{n=0}^{N} T_n},\ k = \overline{1, k_1-1}, k_1 \neq 2, \\
(Q_{i,i})_{1,0} = \begin{pmatrix} O_{WT_N \times W \sum\limits_{n=0}^{N-1} T_n} & \gamma I_{WT_N} + I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_1) \end{pmatrix},\ \textit{if}\ k_1 = 1, \\
(Q_{i,i})_{1,0} = \gamma I_{W \sum\limits_{n=0}^{N} T_n} + \mathrm{diag}^+\{D_1 \otimes P_n(\boldsymbol{\beta}_2),\, n = \overline{0,N-1}\},\ \textit{if}\ k_1 = 2, \\
(Q_{i,i})_{k_1,k_1-1} = k_1\gamma \begin{pmatrix} O_{WT_N \times W \sum\limits_{n=0}^{N-1} T_n} & I_{WT_N} \end{pmatrix},\ k_1 \neq 1, \\
(Q_{i,i})_{k,k-1} = k\gamma I_{WT_N},\ k = \overline{k_1+1, K}, \\
(Q_{i,i})_{k_1-1,0} = \mathrm{diag}^+\{D_1 \otimes P_n(\boldsymbol{\beta}_{k_1}),\, n = \overline{0,N-1}\},\ k_1 \neq 1, \\
(Q_{i,i})_{k,0} = \begin{pmatrix} O_{WT_N \times W \sum\limits_{n=0}^{N-1} T_n} & I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_k) \end{pmatrix}, \\
k = \overline{k_1, K},\ \textit{if}\ k_1 \neq 1,\ \textit{and}\ k = \overline{k_1+1, K},\ \textit{if}\ k_1 = 1;
\end{gathered} \tag{1}$$

2. *the updiagonal blocks* $Q_{i,i+1}$, $i \geq 0$, *have the form:*

$$Q_{i,i+1} = \begin{pmatrix} O_{W\bar{T} \times W\bar{T}} & O_{W\bar{T} \times WT_N} \\ WT_N \times W\bar{T} & D_1 \otimes I_{T_N} \end{pmatrix}; \tag{2}$$

3. *the subdiagonal blocks* $Q_{i,i-1}, i \geq 1$, *have the form* $Q_{i,i-1} = (Q_{i,i-1})_{k,k'}$, $k, k' = \overline{0, K}$, *where the non-zero blocks* $(Q_{i,i-1})_{k,k'}$ *are given as*

$$\begin{gathered}
(Q_{i,i-1})_{k,k} = \psi i I_{W \sum\limits_{n=0}^{N} T_n},\ k = \overline{0, k_1-1}, \\
(Q_{i,i-1})_{k,k} = \psi i I_{WT_N},\ k = \overline{k_1, K}, \\
(Q_{i,i-1})_{k,k+1} = \alpha i I_{W \sum\limits_{n=0}^{N} T_n},\ k = \overline{0, k_1-2}, \\
(Q_{i,i-1})_{k_1-1,0} = \alpha i \mathrm{diag}^+\{I_W \otimes P_n(\boldsymbol{\beta}_{k_1}),\, n = \overline{0,N-1}\}, \\
(Q_{i,i-1})_{k_1-1,k_1} = \alpha i \begin{pmatrix} O_{W \sum\limits_{n=0}^{N-1} T_n \times WT_N} \\ WT_N \end{pmatrix}, \\
(Q_{i,i-1})_{k,k+1} = \alpha i I_{WT_N},\ k = \overline{k_1, K-1}.
\end{gathered} \tag{3}$$

Here the following denotations are used:

the symbols $\otimes$ *and* $\oplus$ *stand for the Kronecker product and sum of matrices, respectively; see, for example, [64];*

I *represents the identity matrix, and* O *represents the zero matrix, whose dimension is provided by a subscript if appropriate;*

$\delta_{i,j}$ *is the Kronecker's symbol, i.e.,* $\delta_{i,j} = \begin{cases} 1, \ i = j; \\ 0, \ i \neq j. \end{cases}$

$\mathrm{diag}\{d_1, d_2, \ldots, d_n\}$, $\mathrm{diag}^+\{d_1, d_2, \ldots, d_n\}$, *and* $\mathrm{diag}^-\{d_1, d_2, \ldots, d_n\}$ *are the diagonal matrix, updiagonal matrix, and subdiagonal matrix with the diagonal, updiagonal, and subdiagonal elements* $d_1, d_2, \ldots, d_n$*, respectively;*

$\mathbf{m}_t = \{m_t^{(1)}, \ \ldots, m_t^{(M)}\}$;

the numbers T_n *that are given by the formula*

$$T_n = \binom{n+M-1}{n} = \frac{(n+M-1)!}{n!(M-1)!}, \ n = \overline{1,N}, \ T_0 = 1,$$

specify the cardinality of the state space of the process $\mathbf{m}_t$ *when* n *servers are busy;*

$\bar{T} = k_1 \sum\limits_{n=0}^{N} T_n + (K - k_1 - 1)T_N$.

We also use the following matrices characterizing various transitions of the vector random process $\mathbf{m}_t$:

The matrix L_n *represents the process transition intensities* $\mathbf{m}_t$ *at the time when service in one of* n *busy servers is finished,* $n = \overline{1,N}$;

The matrix A_n *comprises the transition intensities of the process* $\mathbf{m}_t$ *at the time of the change in service phase in one of* n *busy servers,* $n = \overline{1,N}$;

The matrix $P_n(\boldsymbol{\beta}_k)$ *specifies the process* $\mathbf{m}_t$ *transition probabilities at the point when the group of* k *orders begins service in the presence of* n *busy servers,* $n = \overline{0,N-1}$;

The diagonal entries of the diagonal matrix Δ_n *define the rates of the departure of the process* $\mathbf{m}_t$ *from the corresponding states. The matrices* Δ_n *are defined by the formula*

$$\Delta_n = -\mathrm{diag}\{A_n\mathbf{e} + L_n\mathbf{e}\}, \ n = \overline{1,N}.$$

The detailed description of the matrices $P_n(\boldsymbol{\beta}_k)$ $n = \overline{0,N-1}$, $k = \overline{k_1,K}$, L_n, A_n, Δ_n, $n = \overline{1,N}$, *and the recursive algorithms for their computation are given in [65].*

Proof. The theorem is proved by studying the intensities of all conceivable transitions of the Markov chain ξ_t during an infinitesimal time period. Since during such an period orders enter and leave the orbit one at a time, the matrices $Q_{i,j}$, $i,j \geq 0$, are zero matrices for all i,j such that $|i-j| > 1$. The blocks $Q_{i,j}$, $|i-j| \leq 1$, are built from the matrices $(Q_{i,j})_{k,k'}$ containing the transition rates of the Markov chain ξ_t from the macrostate (i,k) to the macrostate (j,k'), $k,k' = \overline{0,K}$.

Let us explain the form of all these blocks.

1. The matrices $Q_{i,i}$, $i \geq 0$, have the non-zero diagonal blocks $(Q_{i,i})_{k,k}$, $k = \overline{0,K}$, subdiagonal blocks $(Q_{i,i})_{k,k-1}$, $k = \overline{1,K}$, and updiagonal blocks $(Q_{i,i})_{k,k+1}$, $k = \overline{0,K-1}$, and also the blocks $(Q_{i,i})_{k,0}$, $k = \overline{k_1,K}$. This is explained by the fact that during an interval of infinitesimal length, orders can arrive to the buffer one-by-one, renege the system one at a time due to impatience, and move to service in groups of size k, where $k = \overline{k_1,K}$.

The diagonal elements of the diagonal blocks $(Q_{i,i})_{k,k}$, $k = \overline{0,K}$, of the $Q_{i,i}$ matrices are negative. Their modules determine the intensity of departure of the Markov chain ξ_t from the respective state. The Markov chain ξ_t can exit from its current state in the following cases:

a) The underlying process ν_t of order arrival leaves the current state. The corresponding transition intensities are determined up to sign by the diagonal entries of the matrix $D_0 \otimes I_{\sum\limits_{n=0}^{N} T_n}$ for $k = \overline{0,k_1-1}$, and the matrix $D_0 \otimes I_{T_N}$ for $k = \overline{k_1,K}$.

b) One of the busy servers' service processes changes its phase. In this case, the transition rates are determined by the diagonal entries of the matrix $\text{diag}\{O_{W\times W}, I_W \otimes \varDelta_n, n = \overline{1,N}\}$, if $k = \overline{0,k_1-1}$, and matrix $I_W \otimes \varDelta_N$, if $k = \overline{k_1,K}$.

c) An order from the buffer reneges from the system. The corresponding rates are given by the matrices $k\gamma I_{W\sum\limits_{n=0}^{N} T_n}$, $k = \overline{0,k_1-1}$, and $k\gamma I_{WT_N}$, $k = \overline{k_1,K}$.

d) An order from the orbit makes a successful attempt to enter the buffer. The matrices $\alpha i I_{W\sum\limits_{n=0}^{N} T_n}$, $k = \overline{0,k_1-1}$, and $\alpha i I_{WT_N}$, $k = \overline{k_1,K-1}$, set the corresponding intensities. Note that if the buffer is full, the order cannot make a successful attempt, which explains the summand $\delta_{k,K}\alpha i I_{WT_N}$ in the block specified by formula (1).

f) The order leaves orbit due to impatience. The matrices $\psi i I_{W\sum\limits_{n=0}^{N} T_n}$, $k = \overline{0,k_1-1}$, and $\psi i I_{WT_N}$, $k = \overline{k_1,K}$, set the corresponding intensities.

The non-diagonal entries of the matrices $(Q_{i,i})_{k,k}$, $k = \overline{0,K}$, of the matrices $Q_{i,i}$ determine the transition rates of the Markov chain ξ_t without changing the values of the components i and k. These transitions are defined by:

a) the non-diagonal entries of the matrix $D_0 \otimes I_{\sum\limits_{n=0}^{N} T_n}$, if $k = \overline{0,k_1-1}$, or $D_0 \otimes I_{T_N}$, if $k = \overline{k_1,K}$ when the underlying process ν_t makes a jump without an order generation;

b) the entries of the matrix $\text{diag}^{-}\{I_W \otimes L_n, n = \overline{1,N}\}$ when the process $\mathbf{m}_t$ makes a transition implying the finish of the service, but a new service does not begin, since the number k of the orders in the buffer is such that $k < k_1$;

c) the entries of the matrix $\text{diag}\{O_{W\times W}, I_W \otimes A_n, n = \overline{1,N}\}$, if $k = \overline{0,k_1-1}$, and matrix $I_W \otimes A_N$, if $k = \overline{k_1,K}$, when the process $\mathbf{m}_t$ makes a jump that does not lead to service termination;

d) if $k_1 = 1$ and there is a free server, the entries of the matrix $\mathrm{diag}^+\{D_1 \otimes P_n(\boldsymbol{\beta}_1),\ n = \overline{0, N-1}\}$ define the transition rates when a new order arrives while the buffer is empty. In this case, this incoming order is sent for service.

Next, we will comment the expressions for the blocks $(Q_{i,i})_{k,k+1}$, $k = \overline{0, K-1}$, which contain the rates of Markov chain ξ_t transitions from the macrostate (i, k) to the macrostate $(i, k+1)$. Obviously, an increase in the number of orders in the buffer by one may occur when a new order arrives in the system. The transition rates of the process ν_t at the moment of an order arrival are determined by the entries of the matrix D_1; therefore, the blocks $(Q_{i,i})_{k,k+1}$ are given by: the matrix $\mathrm{diag}\{D_1 \otimes I_{T_n},\ n = \overline{0, N}\}$, when $k = \overline{0, k_1 - 2}$, the matrix $\begin{pmatrix} O_{W\sum\limits_{n=0}^{N-1} T_n \times WT_N} \\ D_1 \otimes I_{T_N} \end{pmatrix}$ when $k = k_1 - 1$ (in this case the order adds to the buffer only is all servers are busy) and the matrix $D_1 \otimes I_{T_N}$ for all other k.

The blocks $(Q_{i,i})_{k,k-1}$, $k = \overline{1, K}$, contain the transition rates of the Markov chain ξ_t occurring when the number of orders in buffer decreases by one. This can happen only when some order reneges due to impatience.

Thus, the matrices $(Q_{i,i})_{k,k-1}$ are given by the matrix $k\gamma I_{W\sum\limits_{n=0}^{N} T_n}$, if $k = \overline{1, k_1 - 1}$, $k_1 \neq 2$, the matrix $k_1\gamma \begin{pmatrix} O_{WT_N \times W\sum\limits_{n=0}^{N-1} T_n} & I_{WT_N} \end{pmatrix}$ for $k = k_1$, $k_1 \neq 1$, and the matrix $k\gamma I_{WT_N}$, if $k = \overline{k_1 + 1, K}$.

Let us explain in more detail the form of blocks $(Q_{i,i})_{1,0}$ when $k_1 = 1$ and $k_1 = 2$. If $k_1 = 1$, then a released server always starts service if the buffer is not empty. The reduction of the number of orders in the buffer occurs if the service is finished or an order reneges. The rates of occurring these events are specified by the matrices $\begin{pmatrix} O_{WT_N \times W\sum\limits_{n=0}^{N-1} T_n} & I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_1) \end{pmatrix}$ and $\begin{pmatrix} O_{WT_N \times W\sum\limits_{n=0}^{N-1} T_n} & \gamma I_{WT_N} \end{pmatrix}$ respectively.

If $k_1 = 2$, then the loss of an order due to impatience can imply the decrease in the number of orders in the buffer from one to zero. The rates of this event occurrence are given by the matrix $\gamma I_{W\sum\limits_{n=0}^{N} T_n}$. Also, the scenario is possible when there is an idle server and one order stays in the buffer, a new order arrives, and two orders move to service. The corresponding rates are given by the components of the matrix $\mathrm{diag}^+\{D_1 \otimes P_n(\boldsymbol{\beta}_2),\ n = \overline{0, N-1}\}$.

Next, let's comment the expressions for the blocks $(Q_{i,i})_{k,0}$, specifying the transition rates of the process ξ_t from the macrostate (i, k) to the macrostate $(i, 0)$, which occurs when k orders are accepted for simultaneous service. The corresponding rates are given by the entries of the matrix

$$\begin{pmatrix} O_{WT_N \times W\sum\limits_{n=0}^{N-1} T_n} & I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_k) \end{pmatrix}$$

for $k = \overline{k_1, K}$, if $k_1 \neq 1$, and for $k = \overline{k_1 + 1, K}$, if $k_1 = 1$. Also, we need to specially take into account the scenario when there is an idle server and $k_1 - 1$, $k_1 \neq 1$, orders are already in the buffer at the moment a new order arrives. In this case, a group of k_1 orders starts receiving service. The intensities of this event occurrence are given by the entries of the matrix $\mathrm{diag}^{+}\{D_1 \otimes P_n(\boldsymbol{\beta}_{k_1}),\ n = \overline{0, N-1}\}$.

As a result of the presented considerations, we obtain the expressions for the blocks $Q_{i,i}$, $i \geq 0$, presented above.

2. The updiagonal blocks $Q_{i,i+1}$, $i \geq 0$, contain the transition rates of the Markov chain ξ_t occurring when the number of orbiting orders increases. This can only occur when a new order enters the system when the buffer is full. Therefore, these blocks are specified by a matrix of form (2).

3. The subdiagonal blocks $Q_{i,i-1}$, $i \geq 1$, contain the rates of the Markov chain ξ_t transition when the number of orders in the orbit decreases by one. This can occur when an order from the orbit makes a successful attempt to enter the buffer, so $Q_{i,i-1}$ blocks have the non-zero updiagonal blocks $(Q_{i,i-1})_{k,k+1}$, $k = \overline{0, K-1}$, which are specified by the matrix $\alpha i I_{W \sum\limits_{n=0}^{N} T_n}$, if $k = \overline{0, k_1 - 2}$, and matrix $\alpha i \begin{pmatrix} O_{W \sum\limits_{n=0}^{N-1} T_n \times W T_N} \\ I_{W T_N} \end{pmatrix}$ for $k = k_1 - 1$ and matrix $\alpha i I_{W T_N}$, if $k = \overline{k_1, K-1}$. The blocks $Q_{i,i-1}$ also have a non-zero block $(Q_{i,i-1})_{k_1-1,0}$, given by formula (3), which specifies the transition intensity of the Markov chain in the case when an order from the orbit enters the buffer when it already contains $k_1 - 1$, and then a group of k_1 goes for service. Additionally, a decrease in the number of orders in the orbit may occur due to an order leaving the orbit due to impatience. This departure does not in any way affect the number of occupied devices, as well as the process of arrival and service; in view of this, the intensities of the corresponding transitions are specified by the diagonal blocks $(Q_{i,i-1})_{k,k}$, which have the form $(Q_{i,i-1})_{k,k} = \psi i I_{W \sum\limits_{n=0}^{N} T_n}$ in the case of $k < k_1 - 1$ and $(Q_{i,i-1})_{k,k} = \psi i I_{W T_N}$ otherwise.

The theorem has been proven.

4 Conclusion

In this paper, we implemented a detailed description of the multi-server retrial queue with the *MAP* arrival process, finite buffer, group service of orders, *PH* distribution of the group service time with the irreducible representation depending on the size of a group, dependence of the retrial rate on the number of orbiting orders, and orders impatience during the stay in the orbit and in the buffer. The dynamics of the system is determined by the suitable constructed multidimensional Markov chain. The explicit form of the block-structured generator of this Markov chain is obtained. The presented results give an opportunity to implement an analysis of the steady-state behavior of the constructed Markov chain and queueing system. Detailed analysis of this queueing system is presented in [62].

References

1. Chakravarthy, S.R.: A finite capacity $GI/PH/1$ queue with group services. Nav. Res. Logist. (NRL) **39**, 345–357 (1992)
2. Dudin, S.A., Dudin, A.N., Dudina, O.S., Chakravarthy, S.R.: Analysis of a tandem queuing system with blocking and group service in the second node. Int. J. Syst. Sci. Oper. Logist. **10**, 2235270 (2023)
3. Bailey, N.T.: On queueing processes with bulk service. J. R. Stat. Soc. Ser. B (Methodol.) **16**, 80–87 (1954)
4. Downton, F.: Waiting time in bulk service queues. J. R. Stat. Soc. Ser. B (Methodol.) **17**, 256–261 (1955)
5. Miller, R.G. Jr.: A contribution to the theory of bulk queues. J. R. Stat. Soc. Ser. B Stat. (Methodol.) **21** 320–337 (1959)
6. Keilson, J.: The general bulk queue as a Hilbert problem. J. R. Stat. Soc. Ser. B Stat. (Methodol.) **24**(2) 344–358 (1962)
7. Neuts, M.F.: A general class of bulk queues with Poisson input. Ann. Math. Stat. **38**, 759–770 (1967)
8. Deb, R., Serfozo, R.: Optimal control of batch service queues. Adv. Appl. Probab. **5**, 340–361 (1973)
9. Chaudhry, M.L., Templeton, J.G.C.: A First Course in Bulk Queues. Wiley, New York (1983)
10. Sasikala, S., Indhira, K.: Bulk service queueing models-a survey. Int. J. Pure Appl. Math. **106**(6), 43–56 (2016)
11. Niranjan, S.P., Indhira, K.: A review on classical bulk arrival and batch service queueing models. Int. J. Pure Appl. Math. **106**(8), 45–51 (2016)
12. Brugno, A., D Apice, C., Dudin, A., Manzo, R.: Analysis of an $MAP/PH/1$ queue with flexible group service. Int. J. Appl. Math. Comput. Sci. **27** 119–131 (2017)
13. Pradhan, S., Gupta, U.C.: Analysis of an infinite-buffer batch-size-dependent service queue with Markovian arrival process. Ann. Oper. Res. **277**, 161–196 (2019)
14. Nakamura, A., Phung-Duc, T.: Equilibrium analysis for batch service queueing systems with strategic choice of batch size. Mathematics **11**, 3956 (2023)
15. Claeys, D., Steyaert, B., Walraevens, J., Laevens, K., Bruneel, H.: Analysis of a versatile batch-service queueing model with correlation in the arrival process. Perform. Eval. **70**, 300–316 (2013)
16. Chakravarthy, S.R.: Analysis of a queueing model with MAP arrivals and heterogeneous phase-type group services. Mathematics. **10**, 3575 (2022)
17. Chakravarthy, S.R., Dudin, A.N.: A multi-server retrial queue with BMAP arrivals and group services. Queueing Syst. **42**, 5–31 (2002)
18. Banerjee, A., Gupta, U.C., Chakravarthy, S.R.: Analysis of a finite-buffer bulk-service queue under Markovian arrival process with batch-size-dependent service. Comput. Oper. Res. **60**, 138–149 (2015)
19. Neuts, M.F.: A versatile Markovian point process. J. Appl. Probab. **16**(4), 764–779 (1979)
20. Ramaswami, V.: The $N/G/1$ queue and its detailed analysis. Adv. Appl. Probab. **12**(1), 222–261 (1980)

21. Lucantoni, D.M.: New results on the single server queue with a batch Markovian arrival process. Commun. Stat. Stoch. Model **7**, 1–46 (1991)
22. 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
23. Chakravarthy, S.R.: The batch Markovian arrival process: a review and future work. Adv. Probab. Theory Stoch. Process. **1**, 21–49 (2001)
24. Chakravarthy, S.R.: Introduction to Matrix-Analytic Methods in Queues 1: Analytical and Simulation Approach-Basics. ISTE Ltd., London. Wiley, New York (2022)
25. Chakravarthy, S.R.: Introduction to Matrix-Analytic Methods in Queues 2: Analytical and Simulation Approach-Queues and Simulation, ISTE Ltd.: London. Wiley, New York (2022)
26. Dudin, A.N., Klimenok, V.I., Vishnevsky, V.M.: The Theory of Queuing Systems with Correlated Flows. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-32072-0
27. Vishnevskii, V.M., Dudin, A.N.: Queueing systems with correlated arrival flows and their applications to modeling telecommunication networks. Autom. Remote Control. **78**, 1361–1403 (2017)
28. Buchholz, P., Kriege, J., Felko, I.: Input Modeling with Phase-Type Distributions and Markov Models: Theory and Applications. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-319-06674-5
29. Kriege, J., Buchholz, P.: PH and MAP fitting with aggregated traffic traces. In: Fischbach, K., Krieger, U.R. (eds.) MMB&DFT 2014. LNCS, vol. 8376, pp. 1–15. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-05359-2_1
30. Okamura, H., Dohi, T.: mapfit: an R-based tool for PH/MAP parameter estimation. In: Campos, J., Haverkort, B.R. (eds.) QEST 2015. LNCS, vol. 9259, pp. 105–112. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22264-6_7
31. Chakravarthy, S.R.: Analysis of a finite $MAP/G/1$ queue with group services. Queuing Syst. Theory Appl. **13**, 385–407 (1993)
32. Chakravarthy, S.R.: Two finite queues in series with nonrenewal input and group services. In: Seventh International Symposium on Applied Stochastic Models and Data Analysis, pp. 78–87 (1995)
33. Chakravarthy, S.R., Shruti, G., Rumyantsev, A.: Analysis of a queueing model with batch Markovian arrival process and general distribution for group clearance. Methodol. Comput. Appl. Probab. **23**, 1551–1579 (2021)
34. Chakravarthy, S., Alfa, A.S.: A multiserver queue with Markovian arrivals and group services with thresholds. Nav. Res. Logist. (NRL) **40**, 811–827 (1993)
35. Chakravarthy, S.R.: Analysis of a multi-server queue with batch Markovian arrivals and group services. Eng. Simul. **18**, 51–66 (2000)
36. Chakravarthy, S.R., Dudin, A.N.: A batch Markovian queue with a variable number of servers and group services. In: Matrix-Analytic Methods: Theory and Applications, pp. 63–88. World Scientific Publishing Co. (2002)
37. Chakravarthy, S., Alfa, A.S.: A finite capacity queue with Markovian arrivals and two servers with group services. J. Appl. Math. Stoch. Anal. **7**, 161–178 (1994)
38. Chakravarthy, S.R., Dudin, A.: Analysis of a retrial queuing model with MAP arrivals and two types of customers. Math. Comput. Modell. **37**(3–4), 343–363 (2003)
39. Dudin, A.N., Chakravarthy, S.R.: Multi-threshold control of the $BMAP/SM/1/K$ queue with group services. J. Appl. Math. Stoch. Anal. **16**(4), 327–347 (2003)

40. Bini, D.A., Chakravarthy, S.R., Meini, B.: Control of the $BMAP/PH/1/K$ queue with group services. In: Advances in Algorithmic Methods for Stochastic Models, pp. 57–72. Notable Publications Inc., New Jersey (2000)
41. Dudin, A., Chakravarthy, S.: Optimal hysteretic control for the $BMAP/G/1$ system with single and group service modes. Ann. Oper. Res. **112**, 153–169 (2002)
42. Dudin, A., Manzo, R., Piscopo, R.: Single server retrial queue with adaptive group admission of customers. Comput. Oper. Res. **61**, 89–99 (2015)
43. Brugno, A., Dudin, A.N., Manzo, R.: Retrial queue with discipline of adaptive permanent pooling. Appl. Math. Model. **50**, 1–16 (2017)
44. Brugno, A., D Apice, C., Dudin, A., Manzo, R.: Analysis of an $MAP/PH/1$ queue with flexible group service. Int. J. Appl. Math. Comput. Sci. **27**, 119–131 (2017)
45. Brugno, A., Dudin, A.N., Manzo, R.: Analysis of a strategy of adaptive group admission of customers to single server retrial system. J. Ambient. Intell. Humaniz. Comput. **9**, 123–135 (2018)
46. D Arienzo, M.P., Dudin, A.N., Dudin, S.A., Manzo, R.: Analysis of a retrial queue with group service of impatient customers. J. Ambient. Intell. Humaniz. Comput. **11** 2591–2599 (2020)
47. Singh, G., Gupta, U.C., Chaudhry, M.L.: Computational analysis of bulk service queue with Markovian arrival process: $MAP/R(a,b)/1$ queue. Opsearch **50**, 582–603 (2013)
48. Avram, F., Gomez-Corral, A.: On bulk-service $MAP/P^{L,N}/1/N$ G-Queues with repeated attempts. Ann. Oper. Res. **141**, 109–137 (2006)
49. Banik, A.D.: Queueing analysis and optimal control of $BMAP/G(a,b)/1/N$ and $BMAP/MSP(a,b)/1/N$ systems. Comput. Ind. Eng. **57**, 748–761 (2009)
50. Banik, A.D.: Single server queues with a batch Markovian arrival process and bulk renewal or non-renewal service. J. Syst. Sci. Syst. Eng. **24**, 337–363 (2015)
51. Gupta, U.C., Laxmi, P.V.: Analysis of the $MAP/G^{a,b}/1/N$ queue. Queueing Syst. **38**, 109–124 (2001)
52. Dudin, S., Dudina, O.: Analysis of a multi-server queue with group service and service time dependent on the size of a group as a model of a delivery system. Mathematics **11**(4587), 1–20 (2023)
53. Falin, G., Templeton, J.G.: Retrial Queues, vol. 75. CRC Press, Boca Raton (1997)
54. 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
55. He, Q.M., Li, H., Zhao, Y.Q.: Ergodicity of the $BMAP/PH/s/s+K$ retrial queue with PH-retrial times. Queueing Syst. **35**(1), 323–347 (2000)
56. Breuer, L., Dudin, A., Klimenok, V.: A retrial $BMAP/PH/N$ system. Queueing Syst. **40**(4), 433–457 (2002)
57. Breuer, L., Klimenok, V., Birukov, A., Dudin, A., Krieger, U.R.: Modeling the access to a wireless network at hot spots. Eur. Trans. Telecommun. **16**(4), 309–316 (2005)
58. Dudin, S., Dudina, O.: Retrial multi-server queuing system with PHF service time distribution as a model of a channel with unreliable transmission of information. Appl. Math. Model. **65**, 676–695 (2019)
59. Klimenok, V.I., Orlovsky, D.S., Dudin, A.N.: A $BMAP/PH/N$ system with impatient repeated calls. Asia-Pac. J. Oper. Res. **24**(03), 293–312 (2007)
60. Kim, C.S., Klimenok, V., Mushko, V., Dudin, A.: The $BMAP/PH/N$ retrial queueing system operating in Markovian random environment. Comput. Oper. Res. **37**(7), 1228–1237 (2010)

61. Swensen, A.R.: Remaining loads in a $PH/M/c$ queue with impatient customers. Methodol. Comput. Appl. Probab. **25**, 25 (2023)
62. Dudina O., Dudin A.: Retrial queueing system of $MAP/PH/N$ type with a finite buffer and group service. Stationary analysis of the system. In: Nazarov, A., et al. (eds.) ITMM 2023/WRQ 2023. CCIS, vol. 2163, pp. 257–271. Springer, Cham (2024)
63. Neuts, M.F.: Matrix-Geometric Solutions in Stochastic Models. The Johns Hopkins University Press, Baltimore (1981)
64. Graham, A.: Kronecker Products and Matrix Calculus with Applications. Ellis Horwood, Cichester (1981)
65. Kim, C., Dudin, A., Dudin, S., Dudina, O.: Mathematical model of operation of a cell of a mobile communication network with adaptive modulation schemes and handover of mobile users. IEEE Access **9**, 106933–106946 (2021)

Retrial Queueing System of $MAP/PH/N$ Type with a Finite Buffer and Group Service. Stationary Analysis of the System

Olga Dudina and Alexander Dudin(✉)

Department of Applied Mathematics and Computer Science,
Belarusian State University, 220030 Minsk, Belarus
dudina@bsu.by, dudin@bsu.by

Abstract. In this paper, we is consider a multi-server queueing system with a finite buffer and customer (order) retrials. Arrival flow is defined by the Markov Arrival Process. Service is provided to groups of orders. The size of the group is bounded from below by a fixed threshold. The service time of a group has a phase-type distribution with irreducible representation depending on the size of a group. The linear dependence of the total retrial rate on the number of retrying orders is assumed. Retrying and waiting orders are assumed impatient. The behavior of the system is described by the multidimensional continuous-time Markov chain, including, as the components, the number of retrying orders, the number of orders waiting in the buffer, the number of busy servers, the state of the underlying process of arrivals, and the auxiliary multidimensional Markov chain, which defines the number of servers providing service at all possible phases of service. Ergodicity conditions of the analysed chain in cases of patient and impatient orders in the orbit are obtained. Algorithms for the computation of the stationary distribution and the main performance measures of the system are briefly outlined. The feasibility of the presented algorithms is confirmed by presenting the numerical results. The impact of the buffer capacity and the minimum group size on the main performance characteristics of the system is highlighted. An example of solving an optimization problem is presented.

Keywords: Multi-server retrial queue · group service · MAP · phase-type distribution · stationary distribution · ergodicity

1 Introduction

In a recent paper [1], a novel multi-server queueing system with group service and a finite buffer is considered. The arrival flow is defined by the Markov Arrival Process (MAP). The size of a serviced group is random and is bounded from below and above by the fixed thresholds. The service time of the group has a

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 272–287, 2024.
https://doi.org/10.1007/978-3-031-65385-8_20

phase-type distribution with parameters depending on the group size. Waiting orders are impatient. Service of a group having a size below the fixed low threshold is possible when some order intends to abandon service and depart from the system.

In [2], a more complicated model that assumes that in the case of a full buffer the orders are not lost but will make repeated attempts is formulated and analyzed. A multidimensional continuous-time Markov chain (MC) including the number of retrying orders, the number of customers in the buffer, the number of busy servers, the state of the underlying process of arrivals, and multidimensional MC defining the number of servers providing service at all possible phases of service is constructed. The infinitesimal generator of the chain is obtained, and the explicit form of the blocks of the generator is presented. Here, we implement the stationary analysis of this MC.

The structure of the results presentation is as follows. In Sect. 2, the mathematical model is very briefly formulated, and the necessary notation is given. Ergodicity conditions for the multidimensional Markov process describing the behavior of the system are derived in Sect. 3 for the cases of the patient and impatient orders in the orbit. Problems of computation of the stationary distribution of the process and the main performance measures of the system are discussed in Sect. 4. An illustration of the feasibility of the presented results and the impact of buffer capacity and the minimum group size is presented in Sect. 5. A small example of the application of the obtained results for optimization of the operation of the system is presented there. Section 6 contains some concluding remarks.

2 Brief Description of Mathematical Model and Notation

We consider a multi-server queueing system with N independent identical servers and a finite buffer of size K.

Orders enter the system in the MAP flow defined by an irreducible MC with continuous time ν_t, $t \geq 0$, having a finite state space $\{1, 2, ..., W\}$, and matrices D_0 and D_1. The matrix $D(1) = D_0 + D_1$ is the generator of the MC ν_t.

The average order rate λ is given by $\lambda = \boldsymbol{\theta} D_1 \mathbf{e}$ where the row vector $\boldsymbol{\theta}$ is the only solution to the system $\boldsymbol{\theta} D(1) = \mathbf{0},\ \boldsymbol{\theta}\mathbf{e} = 1$.

Each of the N servers can serve a group of orders consisting of at least k_1 orders. Thus, the parameter k_1, $1 \leq k_1 \leq K$, determines the minimum size of the group that can be taken for service. If an arriving order finds less than $k_1 - 1$, orders in the buffer, it waits until the accumulation of k_1 orders in the buffer. When this happens, all orders from the buffer start service as one group by one available server. If no server is available, the order joins a buffer and waits for picking up from the buffer when some server will become available. The server, which finishes the service of a group, immediately starts the service of all orders from the buffer if the number of these orders is not less than k_1.

If the buffer is full at an order arrival epoch, the order goes to an orbit of unlimited size, from where it repeats attempts to enter the buffer. If at an arbitrary moment the number of orders in the orbit is i, $i > 0$, then the total intensity of retrials is equal to αi, $\alpha > 0$. An attempt is successful if there is at least one free space in the buffer.

The service time of the group has a phase type distribution (PH), specified by a MC m_t, $t \geq 0$, with the state space $\{1, 2, \ldots, M\}$ of the transient states and a unique absorbing state $M+1$. The irreducible representation of MC m_t, $t \geq 0$, is given as $(\boldsymbol{\beta}_k, S)$, $k = \overline{k_1, K}$, where k is the number of orders taken for service.

Orders staying in the buffer and in the orbit may become impatient and leave the system without service after a random interval having an exponential distribution with the rate $\gamma \geq 0$ or ψ, $\psi \geq 0$, correspondingly.

In [2], behavior of the system is described by a continuous-time, regular irreducible MC

$$\xi_t = \{i_t,\, k_t,\, n_t,\, \nu_t,\, m_t^{(1)}, \ldots\, m_t^{(M)}\},\ t \geq 0,$$

where i_t, $i_t \geq 0$, is the number of orders in the orbit, k_t, $k_t = \overline{0, K}$, is the number of orders in the buffer, n_t, $n_t = \overline{0, N}$, is the number of busy servers, ν_t, $\nu_t = \overline{1, W}$, is the state of the underlying process MAP, $m_t^{(l)}$ is the number of servers on the l-th service phase, $m_t^{(l)} = \overline{0, n_t}$, $l = \overline{1, M}$, $\sum\limits_{l=1}^{M} m_t^{(l)} = n_t$, at time t, $t \geq 0$.

The states of the MC ξ_t are enumerated in the direct lexicographical order of the components $(i_t, k_t,\ n_t, \nu_t)$ and reverse lexicographical order of the components $(m_t^{(1)}, \ldots, m_t^{(M)})$. The set of states of the chain having the values (i, k) of the first and second components of the MC is called macrostate (i, k), $i \geq 0$, $k = \overline{0, K}$. The set of macrostates (i, k) for all $k = \overline{0, K}$ is called the level i of MC, $i \geq 0$.

The generator Q of the MC ξ_t consists of the matrices $Q_{i,j}$, $i, j \geq 0$, containing the intensities of transitions from level i to level j. In [2], it is shown that the generator Q has the block tridiagonal structure, and explicit expressions for all blocks $Q_{i,j}$, $i, j \geq 0$, $|i - j| \leq 1$, are presented.

In particular, these expressions include the following matrices:

The matrix L_n defines the transition intensities of the vector random process

$$\mathbf{m}_t = \{m_t^{(1)},\ \ldots, m_t^{(M)}\}$$

at the moment when service in one of n busy servers is completed, $n = \overline{1, N}$.

The matrix A_n contains the transition intensities of the process $\mathbf{m}_t$ at the moment of the change in the phase of service in one of n busy servers, $n = \overline{1, N}$.

The matrix $P_n(\boldsymbol{\beta}_i)$ defines the transition probabilities of the process $\mathbf{m}_t$ at the moment when the group of i orders starts service in the presence of n busy servers, $n = \overline{0, N-1}$.

The diagonal elements of the diagonal matrix $\varDelta_n$ determine the rates of the exit of the process $\mathbf{m}_t$ from the corresponding states. The matrices $\varDelta_n$ are computed by the formula

$$\varDelta_n = -\mathrm{diag}\{A_n\mathbf{e} + L_n\mathbf{e}\}.$$

The detailed description of the matrices $P_n(\boldsymbol{\beta}_k)$ $n = \overline{0, N-1}$, $k = \overline{k_1, K}$, L_n, A_n, $\Delta^{(n)}$, $n = \overline{1, N}$, and algorithms for their calculation are presented in [3].

The numbers T_n specify the dimension of the state space of the process $\mathbf{m}_t$ when n servers are busy. They are calculated as

$$T_n = \frac{(n+M-1)!}{n!(M-1)!},\ n = \overline{1, N}.$$

For convenience, we set $T_0 = 1$.

Let us briefly present the results of the analysis of the steady-state behavior of the MC with the generator Q.

3 Ergodicity Condition

It is easy to see that this MC belongs to the class of level-dependent quasi-birth-and-death processes ($LDQBDP$). Therefore, the famous results by M. Neuts, see [4], obtained for level-independent quasi-birth-and-death processes are not applicable here.

The analysis of the steady-state behavior of the MC includes two mandatory steps. Step 1 consists of the derivation of conditions for the existence of the stationary distribution of the chain. Step 2 consists of the computation of this distribution under the assumption that it exists.

In the majority of existing works, computation of the stationary distribution is implemented via its approximation by the solution of a finite system of linear algebraic equations, which is obtained by means of truncation of the infinite system of linear algebraic equations for the stationary probabilities of the MC which has an infinite state space. This truncation may be rough or soft. The soft truncation is based on the assumption that after some level the considered MC behaves as a $LDQBDP$. This allows to use the results by M. Neuts in [4] to compute the stationary probabilities of the MC. Usually, researchers who use the soft truncation refer, for justification, to the paper [5].

Relating to step 1, the situation is more difficult. Some researchers derive an ergodicity condition for the softly truncated MC and consider it as the ergodicity condition for the original MC. But this is completely wrong because the ergodicity condition is defined by the behavior of MCs at high levels. However, this behavior is completely different for the original level-dependent quasi-birth-and-death process and for the truncated MC.

Some other researchers try to refer to the results in [6] for level-dependent quasi-birth-and-death processes. However, the ergodicity condition given in [6] is non-constructive. This condition is given in terms of some matrices, say $F_k,\ k \geq 0$, which are computed from an infinite system of recursive equations. But a solution exists (even in the simplest case of the level independent process) only if the chain is ergodic. While to check the ergodicity, these matrices must be computed first.

In such a situation, we will use the fact that the constructed MC ξ_t belongs to the class of asymptotically quasi-Toeplitz MCs ($AQTMC$s).

According to the definition given in [7], the MC having a generator Q with blocks $Q_{i,j}$ belongs to the class of $AQTMC$ if the following conditions are satisfied: there exist the limits

$$Y_0 = \lim_{i\to\infty} R_i^{-1}Q_{i,i-1},\ Y_1 = \lim_{i\to\infty} R_i^{-1}Q_{i,i} + I,\ Y_2 = \lim_{i\to\infty} R_i^{-1}Q_{i,i+1},$$

and the matrix $\sum\limits_{l=0}^{2} Y_l$ is stochastic. Here, R_i is a diagonal matrix with positive diagonal elements defined as the moduli of the corresponding diagonal elements of matrix $Q_{i,i}$.

Let us show that these limits for the considered MC ξ_t indeed exist and present their explicit form.

It can be verified that in our case the matrix R_i has the form

$$R_i = \mathrm{diag}\{R_i^{(k)},\ k = \overline{0,K}\},$$

where the diagonal blocks $R_i^{(k)}$, $k = \overline{0,K}$, are given by formula

$$R_i^{(k)} = \begin{cases} \mathrm{diag}\{\hat{D}_0, \hat{D}_0 \oplus \varDelta_n,\ n = \overline{1,N}\} + (k\gamma + (\alpha+\psi)i)I_{W\sum\limits_{n=0}^{N} T_n}, & k = \overline{0,k_1-1}; \\ \hat{D}_0 \oplus \varDelta_N + (k\gamma + (\alpha+\psi)i)I_{WT_N}, & k = \overline{k_1,K-1}; \\ \hat{D}_0 \oplus \varDelta_N + (K\gamma + \psi i)I_{WT_N}, & k = K. \end{cases}$$

Here $\hat{D}_0$ is a diagonal matrix whose diagonal elements are the diagonal elements of matrix D_0,

Let's derive expressions for the limiting matrices Y_0, Y_1, and Y_2. It is easy to see that it is necessary to consider two cases. If the individual rate ψ of orders departing from the orbit due to impatience is positive, then it is easy to verify that the limiting matrices have the form $Y_0 = I$, and $Y_1 = Y_2 = O$.

The case $\psi = 0$ is more complicated. In this case, it can be verified that the matrices Y_0, Y_1, and Y_2 exist and have the following form:

1) the matrix Y_0 contains the non-zero blocks $(Y_0)_{k,k'}$, $k, k' = \overline{0,K}$, which are given by

$$(Y_0)_{k,k+1} = I_{W\sum\limits_{n=0}^{N} T_n},\ k = \overline{0,k_1-2},$$

$$(Y_0)_{k_1-1,0} = \mathrm{diag}^{+}\{I_W \otimes P_n(\boldsymbol{\beta}_{k_1}),\ n = \overline{0,N-1}\},$$

$$(Y_0)_{k_1-1,k_1} = \begin{pmatrix} O_{W\sum\limits_{n=0}^{N-1} T_n \times WT_N} \\ I_{WT_N} \end{pmatrix},$$

$$(Y_0)_{k,k+1} = I_{WT_N},\ k = \overline{k_1,K-1}.$$

2) the matrix Y_1 is defined by the formula

$$Y_1 = \begin{pmatrix} O & O \dots O & O & O \\ \vdots & \vdots \ddots \vdots & \vdots & \vdots \\ O & O \dots O & O & O \\ R^{-1}\varPhi & O \dots O & R^{-1}K\gamma I_{WT_N} & R^{-1}[D_0 \oplus (A_N + \varDelta_N) - K\gamma I_{WT_N}] + I \end{pmatrix}$$

where

$$R = \hat{D}_0 \oplus \Delta_N + K\gamma I_{WT_N},$$

$$\Phi = \begin{pmatrix} O_{WT_N \times W\sum\limits_{n=0}^{N-1} T_n} & I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_K) \end{pmatrix}.$$

3) the matrix Y_2 has the form:

$$Y_2 = \begin{pmatrix} O_{W\bar{T}\times W\bar{T}} & O_{W\bar{T}\times WT_N} \\ O_{WT_N\times W\bar{T}} & R^{-1}(D_1 \otimes I_{T_N}) \end{pmatrix}.$$

It is easy to verify that the sum of matrices $\sum\limits_{l=0}^{2} Y_l$ is a stochastic matrix. Then, in accordance with [7], the MC $\xi_t,\ t \geq 0$, is $AQTMC$, and the sufficient ergodicity condition for it can be written in the following form: the MC $\xi_t,\ t \geq 0$, is ergodic if the following inequality holds true

$$\mathbf{y}Y_0\mathbf{e} > \mathbf{y}Y_2\mathbf{e} \tag{4}$$

where the vector $\mathbf{y}$ is the unique solution to the following system

$$\mathbf{y}(Y_0 + Y_1 + Y_2) = \mathbf{y},\ \mathbf{y}\mathbf{e} = 1. \tag{5}$$

Note that the condition

$$\mathbf{y}Y_0\mathbf{e} < \mathbf{y}Y_2\mathbf{e},$$

is sufficient for the $AQTMC$ to be non-ergodic.

In the case of $\psi > 0$, it is easy to verify that inequality (4) takes the form $1 > 0$, which means that the MC under study is ergodic for any value of the system parameters.

Next, consider the case $\psi = 0$.

Let us represent the vector $\mathbf{y}$ in block form as $(\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_K)$, and, substituting it into system (5), we obtain the following system of equations for unknown subvectors $\mathbf{y}_k,\ k = \overline{0, K}$:

$$\begin{cases} \mathbf{y}_{k_1-1}\mathrm{diag}^+\{I_W \otimes P_n(\boldsymbol{\beta}_{k_1}),\ n = \overline{0, N-1}\} + \mathbf{y}_K R^{-1}\Phi = \mathbf{y}_0, \\ \mathbf{y}_0 = \mathbf{y}_1 = \cdots = \mathbf{y}_{k_1-1}, \\ \mathbf{y}_{k_1-1}\begin{pmatrix} O_{W\sum\limits_{n=0}^{N-1} T_n \times WT_N} \\ I_{WT_N} \end{pmatrix} = \mathbf{y}_{k_1}, \\ \mathbf{y}_{k_1} = \mathbf{y}_{k_1+1} = \cdots = \mathbf{y}_{K-2}, \\ \mathbf{y}_{K-2} + \mathbf{y}_K R^{-1} K\gamma I_{WT_N} = \mathbf{y}_{K-1}, \\ \mathbf{y}_{K-1} + \mathbf{y}_K(R^{-1}[D(1) \otimes I_{T_N} + I_W \otimes (A_N + \Delta_N) - K\gamma I_{WT_N}] + I) = \mathbf{y}_K. \end{cases} \tag{6}$$

Taking into account that $\mathbf{y}_0 = \mathbf{y}_{k_1-1}$, and using the explicit form of the matrices appearing in the first equation of system (6), this equation can be rewritten as

$$\mathbf{y}_K R^{-1}\begin{pmatrix} O_{WT_N \times W\sum\limits_{n=0}^{N-1} T_n} & I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_K) \end{pmatrix}$$
$$= \mathbf{y}_{k_1-1}(I - \mathrm{diag}^+\{I_W \otimes P_n(\boldsymbol{\beta}_{k_1}),\ n = \overline{0, N-1}\}),$$

whence it follows that the vector $\mathbf{y}_{k_1-1}$, and therefore all vectors $\mathbf{y}_k, k = \overline{0, k_1-2}$, have the form

$$\mathbf{y}_k = (\mathbf{0}_W, \mathbf{0}_{WT_1}, \mathbf{0}_{WT_2}, \dots, \mathbf{0}_{WT_{N-1}}, \mathbf{y}_K R^{-1}(I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_K))),\ k = \overline{0, k_1-1},$$

or

$$\mathbf{y}_k = (\mathbf{0}_{W \sum\limits_{n=0}^{N-1} T_n}, \mathbf{y}_K R^{-1}(I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_K))),\ k = \overline{0, k_1-1}.$$

Consequently, from system (6), we obtain that

$$\mathbf{y}_k = \mathbf{y}_K R^{-1}(I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_K)),\ k = \overline{k_1, K-2}.$$

By substituting the expression obtained above for the vector $\mathbf{y}_{K-2}$ into the penultimate equation of system (6), we obtain equation:

$$\mathbf{y}_{K-1} = \mathbf{y}_K R^{-1}(I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_K) + K\gamma I_{WT_N}). \tag{7}$$

Next, we substitute the vector $\mathbf{y}_{K-1}$ of form (7) into the last equation of system (6):

$$\mathbf{y}_K R^{-1}(D(1) \otimes I_{T_N} + I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_K) + I_W \otimes (A_N + \Delta_N)) = \mathbf{0}.$$

By direct substitution, we can verify that the vector $\mathbf{y}_K R^{-1}$ can be represented in the form

$$\mathbf{y}_K R^{-1} = c(\boldsymbol{\theta} \otimes \boldsymbol{\varphi})$$

where $\boldsymbol{\theta}$ is the vector of the stationary probabilities of the $MC\ \nu_t$, the vector $\boldsymbol{\varphi}$ is the unique solution to the following system of linear algebraic equations

$$\begin{cases} \boldsymbol{\varphi}(L_N P_{N-1}(\boldsymbol{\beta}_K) + A_N + \Delta_N) = \mathbf{0}, \\ \boldsymbol{\varphi}\mathbf{e} = 1, \end{cases} \tag{8}$$

and c is some non-zero constant.

Finally, the solution to system (6) can be written as:

$$\begin{cases} \mathbf{y}_k = (\mathbf{0}_{W \sum\limits_{n=0}^{N-1} T_n}, c(\boldsymbol{\theta} \otimes \boldsymbol{\varphi})(I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_K))),\ k = \overline{0, k_1-1}. \\ \mathbf{y}_k = c(\boldsymbol{\theta} \otimes \boldsymbol{\varphi})(I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_K)),\ k = \overline{k_1, K-2}, \\ \mathbf{y}_{K-1} = c(\boldsymbol{\theta} \otimes \boldsymbol{\varphi})(I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_K) + K\gamma I_{WT_N}), \\ \mathbf{y}_K R^{-1} = c(\boldsymbol{\theta} \otimes \boldsymbol{\varphi}). \end{cases} \tag{9}$$

Substituting the vector $\mathbf{y}$ in block form into inequality (4), after some algebraic transformations, we obtain the following inequality:

$$\sum_{k=0}^{K-1} \mathbf{y}_k \mathbf{e} > \mathbf{y}_K R^{-1}(D_1 \otimes I_{T_N})\mathbf{e}. \tag{10}$$

Using (9), we rewrite inequality (10) as:

$$K(\boldsymbol{\theta} \otimes \boldsymbol{\varphi})(I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_K) + K\gamma I_{WT_N})\mathbf{e} > (\boldsymbol{\theta} \otimes \boldsymbol{\varphi})(D_1 \otimes I_{T_N})\mathbf{e}.$$

Taking into account that $\boldsymbol{\theta} D_1 \mathbf{e} = \lambda$, and the vectors $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ and matrix $P_{N-1}(\boldsymbol{\beta}_K)$ are stochastic, we obtain the following inequality, equivalent to (4):

$$K[\boldsymbol{\varphi} L_N \mathbf{e} + \gamma] > \lambda. \tag{11}$$

Let us formalize the obtained result in the form of a theorem.

Theorem 1. *In the case of impatient orders staying in the orbit ($\psi > 0$), the MC ξ_t, $t \geq 0$, is ergodic for any system parameters.*

In the absence of impatience of orders staying in the orbit ($\psi = 0$), the sufficient condition for the ergodicity (existence of a stationary probability distribution) of the MC ξ_t, $t \geq 0$, is the fulfillment of inequality (11) where the vector $\boldsymbol{\varphi}$ is the only solution to the system of linear algebraic equations (8). The fulfillment of inequality (11) with the opposite sign provides a sufficient condition for the non-ergodicity of the chain under study.

Inequality (11) has the following probabilistic meaning: the MC ξ_t is ergodic if the average arrival rate of orders entering the system is less than the average rate of orders departing from the system under the condition that the system is overloaded, that is, when the number of orders in the system is very large.

If the ergodicity condition (11) is fulfilled, then the following stationary probabilities of the MC ξ_t exist:

$$\begin{gathered} \pi(i,k,n,\nu,m^{(1)},\dots,m^{(M)}) \\ = \lim_{t\to\infty} P\{i_t = i,\ k_t = k,\ n_t = n,\ \nu_t = \nu,\ m_t^{(1)} = m^{(1)},\ \dots, m_t^{(M)} = m^{(M)}\}, \\ i \geq 0,\ k = \overline{0,K},\ n = \overline{0,N},\ \nu = \overline{1,W},\ m^{(l)} = \overline{0,n},\ l = \overline{1,M},\ \sum_{l=1}^{M} m^{(l)} = n. \end{gathered}$$

Let's form the row vectors $\boldsymbol{\pi}(i,k)$, $i \geq 0$, $k = \overline{0,K}$, of the stationary probabilities of the states belonging to the macrostate (i,k), and the vectors $\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i,0), \boldsymbol{\pi}(i,1), \dots, \boldsymbol{\pi}(i,K))$ of the stationary probabilities of the states belonging to the level i, $i \geq 0$.

It is well known that the row vectors $\boldsymbol{\pi}_i$, $i \geq 0$, satisfy the following system of equations:

$$\begin{cases} (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_i, \dots)Q = \mathbf{0}, \\ (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_i, \dots)\mathbf{e} = 1. \end{cases} \tag{12}$$

The difficulty of solving the infinite system (12) was briefly discussed above. It can be solved using the numerically stable algorithms elaborated in [7–9]. The algorithms from [7,8] can be used for $AQTMC$ with a more general, upper-Hessenberg, structure of the generator. The algorithm from [9] is oriented to the case of the block tri-diagonal generator, which has MC ξ_t under study in this paper. Therefore, this algorithm was used for the preparation of the numerical illustrations presented below.

4 Performance Measures

Having computed the probability vectors $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}(i,k)$, $i \geq 0,\ k = \overline{0,K}$, it is possible to evaluate a variety of performance characteristics of the considered queueing system. Expressions for computing some of them are as follows.

The average number of orders in the orbit is calculated using the formula

$$L_{orbit} = \sum_{i=1}^{\infty} i\boldsymbol{\pi}\mathbf{e}.$$

The average number of orders in the buffer is calculated using the formula

$$L_{buffer} = \sum_{i=0}^{\infty}\sum_{k=1}^{K} k\boldsymbol{\pi}(i,k)\mathbf{e}.$$

The average number of busy servers is calculated by the formula

$$N_{serv} = \sum_{i=0}^{\infty}(\sum_{k=0}^{k_1-1}\sum_{n=1}^{N} n\boldsymbol{\pi}(i,k,n)\mathbf{e} + \sum_{k=k_1}^{K} N\boldsymbol{\pi}(i,k)\mathbf{e}).$$

Here vectors $\boldsymbol{\pi}(i,k,n),\ n = \overline{0,N}$, are defined by the partition

$$\boldsymbol{\pi}(i,k) = (\boldsymbol{\pi}(i,k,0), \boldsymbol{\pi}(i,k,1), \ldots, \boldsymbol{\pi}(i,k,N)).$$

The average intensity of the output flow of successfully served groups of orders is calculated by the formula

$$\mu_{out} = \sum_{i=0}^{\infty}(\sum_{k=0}^{k_1-1}\sum_{n=1}^{N} \boldsymbol{\pi}(i,k,n)(I_W \otimes L_n)\mathbf{e} + \sum_{k=k_1}^{K} \boldsymbol{\pi}(i,k)(I_W \otimes L_N)\mathbf{e}).$$

The probability that an arriving order will find the buffer full and go into orbit is found by the formula

$$P_{to-orbit} = \frac{1}{\lambda}\sum_{i=0}^{\infty} \boldsymbol{\pi}(i,K)(D_1 \otimes I_{T_N})\mathbf{e}.$$

The probability that an order will begin servicing immediately upon arrival is calculated by the formula

$$P_{to-serv} = \frac{1}{\lambda}\sum_{i=0}^{\infty}\sum_{n=0}^{N-1} \boldsymbol{\pi}(i,k_1-1,n)(D_1 \otimes I_{T_n})\mathbf{e}.$$

The rate of orders leaving the buffer for service is calculated by the formula

$$\mu_{to-serv} = \sum_{i=0}^{\infty}\Bigg[(k_1\Bigg(i\alpha\sum_{n=0}^{N-1} \boldsymbol{\pi}(i,k_1-1,n)\mathbf{e} + \sum_{n=0}^{N-1} \boldsymbol{\pi}(i,k_1-1,n)(D_1 \otimes I_{T_n})\mathbf{e}\Bigg) + \sum_{k=k_1}^{K} k\boldsymbol{\pi}(i,k)(I_W \otimes L_N)\mathbf{e}\Bigg].$$

The loss probability of an order from the buffer due to impatience is calculated using the formula

$$P^{imp-loss-buffer} = \frac{\gamma L_{buffer}}{\lambda}.$$

The loss probability of an order from the orbit due to impatience is calculated using the formula

$$P^{imp-loss-orbit} = \frac{\psi L_{orbit}}{\lambda}.$$

The loss probability of an arbitrary order is calculated using the formula

$$P^{loss} = P^{imp-loss-buffer} + P^{imp-loss-orbit} = 1 - \frac{\mu_{to-serv}}{\lambda}.$$

The average size of a group of orders taken for servicing is calculated by the formula

$$N_{batch} = \frac{\mu_{to-serv}}{\mu_{out}}.$$

The loss probability of an order from the buffer due to impatience while there is a free server is calculated by the formula

$$P^{imp-loss}_{idle-server} = \frac{1}{\lambda}\sum_{i=0}^{\infty}\sum_{k=1}^{k_1-1} k\gamma \sum_{n=0}^{N-1} \boldsymbol{\pi}(i,k,n)\mathbf{e}.$$

The loss probability of an order from the buffer due to impatience at a time when all servers are busy is calculated by the formula

$$P^{imp-loss}_{all-busy-servers} = \frac{1}{\lambda}\sum_{i=0}^{\infty}[\sum_{k=1}^{k_1-1} k\gamma\boldsymbol{\pi}(i,k,N)\mathbf{e} + \sum_{k=k_1}^{K} k\gamma\boldsymbol{\pi}(i,k)\mathbf{e}].$$

The probability that, at an arbitrary moment, there is at least one free server in the system is found by the formula

$$P_{idle-server} = \sum_{i=0}^{\infty}\sum_{k=0}^{k_1-1}\sum_{n=0}^{N-1} \boldsymbol{\pi}(i,k,n)\mathbf{e}.$$

The probability that at an arbitrary moment there are orders in the buffer while there is at least one free server is found by the formula

$$P^{customers}_{idle-server} = \sum_{i=0}^{\infty}\sum_{k=1}^{k_1-1}\sum_{n=0}^{N-1} \boldsymbol{\pi}(i,k,n)\mathbf{e}.$$

5 Numerical Examples

The goals of the examples are the demonstration of the feasibility of the presented results and the illustration of the form of dependence of the main performance measures of the system on the capacity K of the input buffer and the minimum size of the serviced group. The knowledge of these dependencies is useful for optimizing the values of K and k_1 in possible applications of the model.

In this example, we consider a model with $N = 5$ servers. Orders arrive into the system in the MAP that is defined by the matrices

$$D_0 = \begin{pmatrix} -1.8 & 0 \\ 0 & -0.6 \end{pmatrix}, \; D_1 = \begin{pmatrix} 1.74 & 0.06 \\ 0.012 & 0.588 \end{pmatrix}.$$

The MAP has the average arrival rate $\lambda = 0.8$, the coefficient of correlation of successive inter-arrival times $c_{cor} = 0.127928$, and the coefficient of variation $c_{var} = 1.37037$.

Let the individual retrial intensity α be equal to 0.2. The intensity of impatience from the buffer is $\gamma = 0.01$, the intensity of impatience from the orbit is $\psi = 0.02$.

The service time of groups is defined as follows. Let the mean service times of groups consisting of k orders in the modeled real system be $s_k, \; k = \overline{1, K}$. Let us assume that $s_1 < s_2 < \cdots < s_K$. To build the PH distributions of service times of the groups having such values of the mean service times, we choose distributions with representations $(\boldsymbol{\beta}_k, S)$, $k = \overline{1, K}$, of size two. The sub-generator S will be fixed in the form:

$$S = \begin{pmatrix} -\frac{1}{s_1} & 0 \\ 0 & -\frac{1}{s_K} \end{pmatrix}$$

while the vector $\boldsymbol{\beta}_k$ is given as

$$\boldsymbol{\beta}_k = (f_k, 1 - f_k)$$

where

$$f_k = \frac{s_K - s_k}{s_K - s_1}, \; k = \overline{1, K}.$$

In this numerical example, we assume that $s_1 = 20$ and $s_k = s_1 + 3(k-1)$, $k = \overline{2, K}$. Thus, the matrix S has the form $S = \begin{pmatrix} -\frac{1}{20} & 0 \\ 0 & -\frac{1}{137} \end{pmatrix}$.

Let us vary the buffer capacity K in the range from 2 to 40 and the parameter k_1 in the range from 1 to K with step 1.

Figures 1 and 2 illustrate the dependencies of the average number of orders in the orbit L_{orbit} and the average number of orders in the buffer L_{buffer} on the parameters K and k_1. It is natural that group service is reasonable when the mean service time per order in the group is less than the mean service time of one individual order. Our fixed choice of the parameters of the PH distributions of service times accounts for these considerations. Thus, it is clear that if the value of k_1 is small and service may be provided to small groups, then the advantages of group service are not used to a proper extent. As a consequence, the average number of orders in the orbit L_{orbit} is large for small k_1 and quickly decreases when k_1 increases. The behaviour of L_{buffer} is more complicated. From one hand, a large value of k_1 implies the better use of the advantages of a group service and decreases the number of orders in the buffer. From the other hand, a large value of k_1 may cause the long waiting for an order in the buffer until the group of the required size is accumulated when the servers are under-utilized. This makes valuable the results of our computations to understand the influence of k_1 on the value of L_{buffer} that can be different for the different loads of the system.

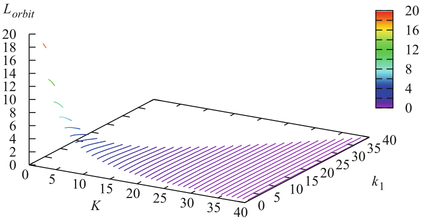


Fig. 1. The dependence of the average number of orders in the orbit L_{orbit} on the parameters K and k_1

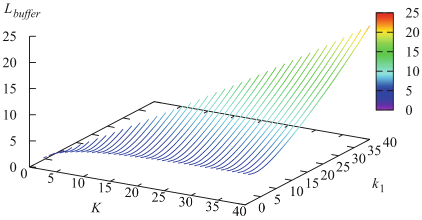


Fig. 2. The dependence of the average number of orders in the buffer L_{buffer} on the parameters K and k_1

Figures 3 and 4 illustrate the dependencies of the probability $P_{to-serv}$ that an order will begin servicing immediately upon arrival and the probability $P_{to-orbit}$ that an arrival order will find the buffer full and go into the orbit on the parameters K and k_1. The surface given in Fig. 4 strongly correlates with the form of dependence given in Fig. 1 because the average number of orders in the orbit strongly depends on the probability $P_{to-orbit}$ that an arrival order will find the buffer full and go into the orbit. Figure 3 is much more interesting. The probability $P_{to-serv}$ that an order will begin servicing immediately upon arrival is one of the most important characteristics of any retrial system. Figure 3 shows the complicated non-monotone dependence of this probability on the parameters K and k_1. This confirms that the problem of the optimal choice of these parameters is challenging, and the obtained results may be useful for its solution.

Figures 5 and 6 illustrate the dependencies of the probabilities $P^{imp-loss-buffer}$ of an order loss from the buffer and $P^{imp-loss-orbit}$ of an order loss from the orbit due to impatience on the parameters K and k_1. These probabilities are also very important from a practical point of view because they reflect the probability of the loss of a possible profit of the system gained via service provision. These probabilities are proportional to L_{buffer} and L_{orbit}, correspondingly.

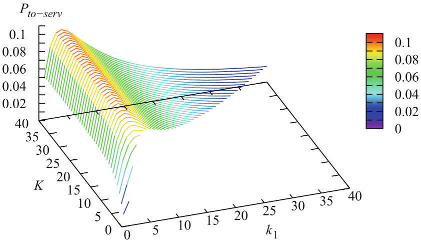


Fig. 3. The dependence of the probability $P_{to-serv}$ that an order will begin servicing immediately upon arrival on the parameters K and k_1

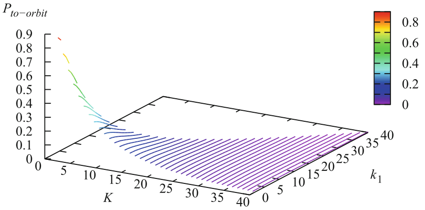


Fig. 4. The dependence of the probability $P_{to-orbit}$ that an arrival order will find the buffer full and go into the orbit on the parameters K and k_1

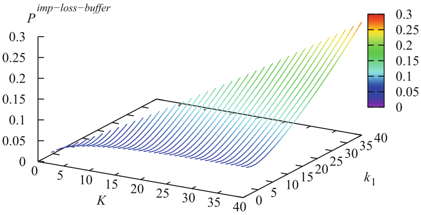


Fig. 5. The dependence of the loss probability $P^{imp-loss-buffer}$ of an order from the buffer due to impatience on the parameters K and k_1

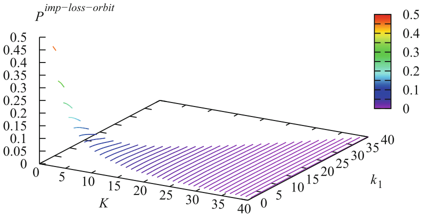


Fig. 6. The dependence of the loss probability $P^{imp-loss-orbit}$ of an order from the orbit due to impatience on the parameters K and k_1

Figure 7 illustrates the dependence of the integral loss probability P^{loss} of an arbitrary order on the parameters K and k_1. This dependence is non-monotone and cannot be exactly evaluated based on common-sense considerations. This confirms the value of the proposed research.

Having clarified the dependencies of the main performance measures of the system on the parameters K and k_1, it is reasonable to formulate and solve an optimization problem. There may be many different choices of criteria for the quality of the system's operation. Here, as an example, we assume that the quality of the system's operation is evaluated in terms of the following cost criterion:

$$E = E(K, k_1) = a\mu_{to-serv} - c_1\lambda P^{imp-loss-buffer} - c_2\lambda P^{imp-loss-orbit} - dK.$$

Here, a is the revenue of the system earned via the service of one order, c_1 and c_2 are the charges for the loss of an arbitrary order due to impatience from the buffer and from the orbit, respectively, and d is the cost for maintaining one place in a buffer per unit of time. Therefore, the criterion E determines the average profit obtained by the system per unit of time, and our managerial goal is to obtain such parameters as K and k_1 under which the system's profit is maximal.

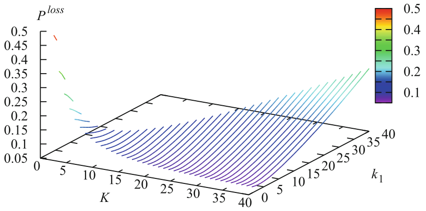


Fig. 7. The dependence of the loss probability P^{loss} of an arbitrary order on the parameters K and k_1

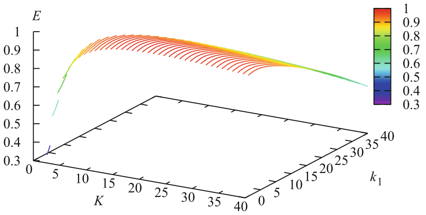


Fig. 8. The dependence of the cost criterion E on the parameters K and k_1 ($d = 0.0025$)

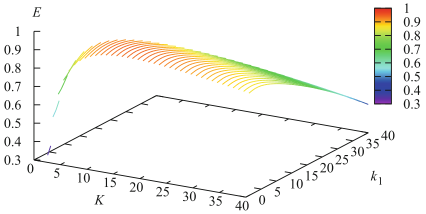


Fig. 9. The dependence of the cost criterion E on the parameters K and k_1 ($d = 0.005$)

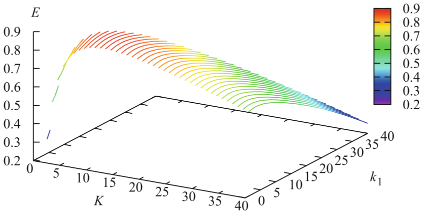


Fig. 10. The dependence of the cost criterion E on the parameters K and k_1 ($d = 0.01$)

In this numerical example, let us fix the following cost coefficients:

$$a = 1.5,\ c_1 = 1,\ c_2 = 0.7.$$

Figures 8, 9, 10, 11 and 12 show the dependence of the cost criterion E on the parameters K and k_1 under five different values of the cost coefficient $d = 0.0025, 0.005, 0.01, 0.02, 0.04$. The shapes presented in these figures show the high variability of the value of the cost criterion and the possibility of essentially increasing the profit of the system via the proper choice of the parameters K and k_1.

Table 1 contains the optimal values of the cost criteria E, the buffer size K and the parameter k_1 under different values of the cost d of maintenance of one size in the buffer.

More general criteria, accounting a cost of server maintenance, can be used as well.

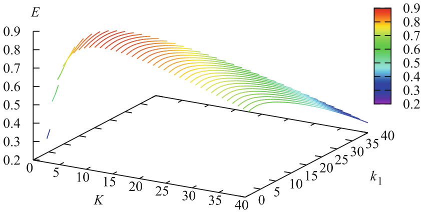


Fig. 11. The dependence of the cost criterion E on the parameters K and k_1 ($d = 0.02$)

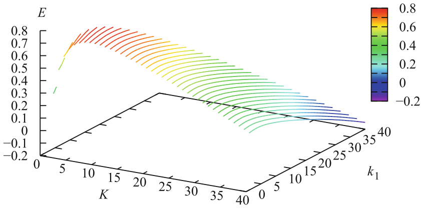


Fig. 12. The dependence of the cost criterion E on the parameters K and k_1 ($d = 0.04$)

Table 1. The optimal values E^*, K^*, k_1^* of the cost criterion E, the buffer size K and the parameter k_1 under different values of the cost coefficient d

	$d = 0.0025$	$d = 0.005$	$d = 0.01$	$d = 0.02$	$d = 0.04$
E^*	0.989587	0.943366	0.875663	0.790062	0.658792
K^*	22	16	11	7	6
k_1^*	4	4	6	7	6

6 Conclusion

In this paper, we continued the started in [2] analysis of a multi-server retrial queue with the MAP arrival process, finite buffer, group service of orders, phase-type distribution of group service time depending on the size of a group, arbitrary dependence of the total retrial rate on the number of orders in the orbit, and orders impatience during the stay in the orbit and in the buffer. In [2], the behavior of the system is described by the multidimensional Markov chain. The explicit form of the block-structured generator of this Markov chain is obtained. In this paper, we used these results to implement an analysis of the stationary behavior of the considered Markov chain and queueing system. We derived the constructive and simple conditions for ergodicity of the considered Markov chain and presented expressions for the computation of the main performance measures of the considered queueing system. The dependencies of these measures on the parameters of the system (capacity of the input buffer and minimal size of a serviced group) are numerically highlighted. The possibility of using the result for managerial goals is numerically illustrated.

References

1. Dudin, S., Dudina, O.: Analysis of a multi-server queue with group service and service time dependent on the size of a group as a model of a delivery system. Mathematics **11**(4587), 1–20 (2023)
2. Dudin, A., Dudina, O.: Retrial queueing system of $MAP/PH/N$ type with a finite buffer and group service. The process describing the system dynamics. In: Nazarov, A., et al. (eds.) ITMM 2023/WRQ 2023. CCIS, vol. 2163, pp. 257–271. Springer, Cham (2024)
3. Kim, C., Dudin, A., Dudin, S., Dudina, O.: Mathematical model of operation of a cell of a mobile communication network with adaptive modulation schemes and handover of mobile users. IEEE Access **9**, 106933–106946 (2021)
4. Neuts, M.F.: Matrix-Geometric Solutions in Stochastic Models. The Johns Hopkins University Press, Baltimore (1981)
5. Neuts, M.F., Rao, B.M.: Numerical investigation of a multiserver retrial model. Queueing Syst. **7**, 169–189 (1990)
6. Bright, L., Taylor, P.G.: Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. Stoch. Model. **11**(3), 497–525 (1995)
7. Klimenok, V.I., Dudin, A.N.: Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. Queueing Syst. **54**, 245–259 (2006)
8. Dudin, S., Dudin, A., Kostyukova, O., Dudina, O.: Effective algorithm for computation of the stationary distribution of multi-dimensional level-dependent Markov chain with upper block-Hessenberg structure of the generator. J. Comput. Appl. Math. **366**, 112425 (2020)
9. Dudin, S., Dudina, O.: Retrial multi-server queuing system with PHF service time distribution as a model of a channel with unreliable transmission of information. Appl. Math. Model. **65**, 676–695 (2019)

The Mean Regeneration Cycle Length in a Constant Retrial Rate System

Evsey Morozov[1,2,3] and Stepan Rogozin[1,2(✉)]

[1] Institute of Applied Mathematical Research, Karelian Research Centre RAS,
185910 Petrozavodsk, Russia
emorozov@karelia.ru

[2] Institute of Mathematics and Information Technologies,
Petrozavodsk State University, 185910 Petrozavodsk, Russia
ppexa@mail.ru

[3] Yaroslav-the-Wise Novgorod State University, Novgorod the Great, Russia

Abstract. In this research, we consider a $M/G/1$-type single-class retrial system with constant retrial rate. First we use the regenerative approach to reprove the stability condition of this system, and then we obtain the explicit expressions for the mean regeneration cycle length and the mean number of customers arrived within regeneration cycle in the stationary regime. Finally, the discrete-event simulation is applied to illustrate the obtained theoretical results.

Keywords: constant retrial rate · regeneration cycle · stability condition · simulation

1 Introduction

The retrial queues are important and convenient tool to model a wide spectrum of the stochastic systems, such as, for example, the wireless telecommunication systems. The analysis of some basic retrial models and comprehensive bibliography on the related research are presented, for instance, in the works [1–3, 7].

In the present research we use a regenerative approach to obtain the stability condition of a $M/G/1$-type retrial system with one class of customers, exponential retrial times and constant retrial rate. Then this condition is used to obtain the explicit expressions for the mean regeneration cycle length and the mean number of customers arrived within a regeneration cycle. Moreover, we compare these characteristics with the corresponding quantities known for the classic buffered $M/G/1$ system. Finally, we conduct a discrete-event simulation to demonstrate the convergence of the sample-mean of the regeneration cycle length and the sample-mean of the number of arrivals per cycle to the corresponding theoretical values. We note that Chap. 4 of the monograph [4] contains recursions allowing to calculate the Laplace-Stieltjes transform of the mean busy period in such a system. However the approach presented in this research, based

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 288–299, 2024.
https://doi.org/10.1007/978-3-031-65385-8_21

on the regenerative method, turns out to be much simpler, more intuitive and leads to the explicit expressions.

Summing up, the main contribution of this work is a regenerative analysis implying the explicit expressions for the mean regeneration cycle length (in continuous and discrete time) for the single-class retrial $M/G/1$-type system in the stationary regime. These expressions are found in the terms of given parameters and do not contain Laplace-Stieltjes transforms. As a by-product of the analysis, the stability condition of the system is also obtained. A basic observation of this analysis is that, within one regeneration cycle, the total idle time of the server can be expressed as the sum of the independent identically distributed (iid) idle periods distributed as the minimum of the remaining (exponential) retrial time and (exponential) interarrival time. We mention the fundamental works [13,14] in the theory of regenerative process, and the monographs [5,9,11] which contains the details of the regenerative approach and its applications in analysis of the modern queueing systems.

The paper is organized as follows. In Sect. 2, we describe the model in detail. In Sect. 3, the main stability result is proved using a regenerative approach. Then, assuming stability, the mean regeneration cycle length is obtained. Finally, in Sect. 4, we present some numerical results illustrating and verifying theoretical results.

2 Description of the Model

We consider $M/G/1$ retrial system with constant retrial rate. Input process is Poisson with rate λ. We denote t_n the arrival instant of the n-th customer and denote interarrival times as $\tau_n = t_{n+1} - t_n$. Because these times are iid we denote by τ the generic interarrival time and we note that $\lambda = 1/\mathsf{E}\tau$. We denote the iid service times as S_n with generic service time S and denote by $\mu = 1/\mathsf{E}S$. If a new customer finds server busy it joins an infinite capacity orbit. After the customer joins the orbit, after an exponential time with rate θ, he makes an attempt to capture the server. We emphasize that only one customer is allowed to make retrial attempts until he finds server idle, It is called the *constant retrial rate policy* [9].

To use the regenerative approach, we first describe the regenerative structure of the stochastic process describing the dynamics of the system. Denote by $Q(t_k) = Q_k$ the total number of customers in the system (that is in orbit and in server) just before arrival time t_k^- of customer k, $k \geq 1$. Then the regeneration instants $\{T_n\}$ of the process $\{Q(t), t \geq 0\}$ are recursively defined as

$$T_{n+1} = \inf_k (t_k > T_n : Q_k = 0), \ n \geq 0, \tag{1}$$

$(T_0 := 0)$, and the regeneration instants of the embedded process $\{Q_n\}$ are defined as

$$\hat{\theta}_{n+1} = \inf(k > \hat{\theta}_n : Q(t_k) = 0), \ n \geq 0 \ \ (\hat{\theta}_0 := 0). \tag{2}$$

We denote by T the generic regeneration cycle length, that is the distance between two arbitrary adjacent regeneration instants, for continuous-time construction (1). Also denote by $\hat{\theta}$ the corresponding quantity for the discrete-time construction (2). We note, that $\hat{\theta}$ equals the number of the customers arrived within a (continuous-time) regeneration cycle. The variable $\hat{\theta}$ is also the (discrete-time) regeneration cycle length of the embedded process $\{Q_n\}$. The regenerative process $\{Q(t)\}$ (and the queueing system) is called *positive recurrent* (stable), if $\mathsf{E}T < \infty$. This condition is a key one to establish the existence of the stationary distribution of the basic regenerative process if the system is empty at the initial instant $t = 0$ [9]. Otherwise, if $\mathsf{E}T = \infty$, then the system is called *null-recurrent* or unstable because, as it is easy to show, in this case the basic regenerative process increases unlimitedly as time increases. By the Wald's identity, it follows that

$$\mathsf{E}T = \mathsf{E}\hat{\theta}\,\mathsf{E}\tau,$$

and we may conclude that (see the work [6]) both processes $\{Q(t)\}$ and $\{Q_n\}$ are positive recurrent/null-recurrent simultaneously.

3 Stability and Performance Analysis

In this section, we first consider a classic $M/G/1$ system with an infinite capacity buffer and, using regenerative approach, we obtain the mean regeneration cycle length. This result is well-known but it allows further to simplify the stability analysis of the basic retrial system. Then we obtain the main result of this research, namely, the explicit expression for the mean regeneration period length of the stationary basic retrial system with general service time.

For arbitrary interval of time $[0, t]$ denote by: $V(t)$ the total arrived work; $B(t)$ the busy time of the server; $I(t)$ the idle time of the server. Also denote by $W(t)$ the remaining work at instant t. Then we have the following balance equation

$$V(t) = W(t) + B(t) = W(t) + t - I(t), \quad t > 0. \tag{3}$$

Denote by $\rho = \mathsf{E}S/\mathsf{E}\tau$ the traffic intensity. It is well known that $\rho < 1$ is the stability condition of the system implying in particular $\mathsf{E}T < \infty$. (It is commonly known that generic regeneration period T for this classic system also is generated by the arrival customers in an idle system [5,9].) It is easy to find, using the Strong Law of Large Numbers (SLLN) that with probability (w.p.) 1

$$\lim_{t\to\infty} \frac{V(t)}{t} = \lambda \mathsf{E}S = \rho. \tag{4}$$

Moreover, by the assumption $\rho < 1$, $\lim_{t\to\infty} W(t)/t = 0$ w.p.1, and it follows from the basic theorem of the regenerative analysis (see Chap. 1 in [9]) that

$$\lim_{t\to\infty} \frac{I(t)}{t} = \frac{\mathsf{E}I_0}{\mathsf{E}T}, \tag{5}$$

where $\mathsf{E}I_0$ is the mean idle time of the server within the regeneration cycle. Collecting the results (3)–(5), and using the memoryless property of the interarrival time, we obtain

$$\rho = 1 - \frac{\mathsf{E}I_0}{\mathsf{E}T} = 1 - \frac{1}{\lambda \mathsf{E}T}. \tag{6}$$

This immediately implies the well-known expression for the mean length of the regeneration cycle

$$\mathsf{E}T = \frac{1}{\lambda(1-\rho)} \tag{7}$$

of the classic single-server system with infinite capacity.

Remark 1. According to the Little's formula [5],

$$\lambda \mathsf{E}T = \mathsf{E}\hat{\theta}, \tag{8}$$

where $\mathsf{E}\hat{\theta}$ is the mean number of customers arrived in the system per regeneration cycle. (Note that we keep the same notation for the system with buffer.) It immediately gives the following (also well-known) expression

$$\mathsf{E}\hat{\theta} = \frac{1}{(1-\rho)}. \tag{9}$$

The analysis above indeed turns out to be useful to obtain below the corresponding expression of the mean regeneration cycle length in the retrial system under consideration.

Now we consider the original system $M/G/1$ with constant retrial rate. Recall that $\rho = \lambda \mathsf{E}S$ and θ is the retrial rate. Now we proof the following main statement.

Theorem 1. *If condition*

$$\rho < \frac{\lambda+\theta}{\mu+\lambda+\theta}, \tag{10}$$

holds then $\mathsf{E}T < \infty$, that is initially empty system is positive recurrent.

Proof. Denote, within the interval $[0,t]$, by $I_0(t)$ the server idle time when the orbit is empty, and let $I_1(t)$ be the server idle time, and simultaneously, the orbit is not empty. Note that then the total idle time is expresses as $I(t) = I_0(t)+I_1(t)$ and then the balance equation becomes (cf. (3))

$$V(t) = W(t) + t - I_0(t) - I_1(t), \quad t \geq 0, \tag{11}$$

where $W(t)$ denotes now the remaining work in the orbit and in the server, if any. Assume, by contradiction, that the system is unstable, i.e., the orbit size

$$Q(t) \Rightarrow \infty, \quad t \to \infty, \tag{12}$$

where symbol $\Rightarrow$ denotes convergence in probability. Then in particular,

$$\mathsf{P}(Q(t)=0) \to 0, \quad t \to \infty. \tag{13}$$

Denote by σ_i the i-th idle time period between the sequential busy periods of the server. Within any regeneration period, these periods are iid and we denote them by $\{\sigma_i\}$, and let σ be the generic such a period (the time between departure of a customer and the beginning of the next service). Then it is easy to see that

$$\sigma = \min(\xi, \tau), \tag{14}$$

where ξ is the generic inter-retrial time. Let $\hat{I}_1(t)$ be the total idle time of server in the renewal process generated by the iid sums $\{S_i + \sigma_i\}$, that is

$$\hat{I}_1(t) = \max_k \Big(\sum_{i=1}^k \sigma_i : \quad S_1 + \sigma_1 + S_2 + \sigma_2 + \cdots + S_k + \sigma_k \le t \Big). \tag{15}$$

We emphasize that, in this process, we 'ignore' the idle periods between regeneration periods. Alternatively, we may imagine the process $\{\hat{I}_1(t)\}$ as if it would be generated in the permanently overloaded system with 'infinitely large' orbit size. It is easy to see that the idle time process $\{I_1(t)\}$ can be interpreted as the 'lost' server capacity when the server is free but the orbit is not idle. This process can be expressed via the renewal process $\{\hat{I}_1(t)\}$ and the process $\{I_0(t)\}$, counting the time when *both orbit and server are free*, as follows:

$$I_1(t) = \hat{I}_1(t - I_0(t)). \tag{16}$$

By the SLLN,

$$\lim_{t\to\infty} \frac{\hat{I}_1(t)}{t} = \frac{\mathsf{E}\sigma}{\mathsf{E}(S+\sigma)} = \frac{\mu}{\mu+\lambda+\theta} =: \varDelta. \tag{17}$$

By the assumption (12), $\mathsf{E}I_0(t)/t \to 0$, and it is known that then there exists such a non-random sequence $z_n \to \infty$ that $I_0(z_n)/z_n \to 0$ w.p.1 [6]. Then, by (16), (17),

$$\begin{aligned}\lim_{n\to\infty} \frac{I_1(z_n)}{z_n} &= \lim_{n\to\infty} \frac{\hat{I}_1(z_n - I_0(z_n))}{z_n - I_0(z_n)} \frac{z_n - I_0(z_n)}{z_n} \\ &= \varDelta\Big(1 - \lim_{n\to\infty} \frac{I_0(z_n)}{z_n}\Big) = \varDelta.\end{aligned} \tag{18}$$

Thus, summing up the results above, we obtain

$$\lim_{n\to\infty} \frac{1}{z_n} W(z_n) = \rho - 1 + \varDelta \ge 0. \tag{19}$$

If the latter condition does not hold, that is,

$$\rho < 1 - \varDelta = \frac{\lambda + \theta}{\mu + \lambda + \theta}, \tag{20}$$

then ‘instability’ assumption (12) is violated, i.e., $Q(t) \not\Rightarrow \infty$. Then, based on the regeneration condition $\mathsf{P}(\tau > S) > 0$ (which holds automatically for the Poisson input process), one can unload system in such a way that a regeneration (the arrival of a customer in the idle system) is reached, within a finite time interval $[u_0, u_0 + D]$ with a positive probability p. (Here the instant u_0 denotes the starting point of the ‘unloading procedure’, see [9].) Since both the interval length D and the probability p do not depend on the time instant u_0 then the positive recurrence of the system follows. (Some basic details of the analysis see in [9].) Thus (10) is the sufficient stability condition, and the proof of Theorem is completed.

Remark 2. Indeed condition (20) is the stability criteria of the system. To show it assume that the opposite condition

$$\rho \geq \frac{\lambda + \theta}{\mu + \lambda + \theta} \tag{21}$$

holds and that the system remains stationary under assumption (21). Then, from the balance equation (11), we obtain easily as above that condition (20) indeed holds, implying contradiction.

Now suppose that the system is stable. Then, from the balance equation (11) similarly to the system with a buffer, we obtain

$$\rho = 1 - \frac{1}{\lambda \mathsf{E}T} - \frac{\mathsf{E}I_1}{\mathsf{E}T}, \tag{22}$$

where I_1 is the server idle time within a regeneration cycle when also the orbit is not empty. According to the Little’s formula and since all customers arrived within regeneration cycle are served in this cycle, we obtain

$$\mathsf{E}T = \mathsf{E}(\mathbb{C} + 1)\mathsf{E}\tau, \tag{23}$$

where $\mathbb{C}$ is the number of the idle intervals (of ‘type’ σ) within a regeneration cycle and is also the number of the customers served within a regeneration cycle. By the Wald identity,

$$\mathsf{E}I_1 = \mathsf{E}\mathbb{C}\mathsf{E}\sigma. \tag{24}$$

Then, from (22)–(24),

$$\mathsf{E}\mathbb{C} = \frac{\rho}{1 - \rho - \lambda \mathsf{E}\sigma}, \tag{25}$$

and by (23) and (18) we find the target mean cycle length

$$\mathsf{E}T = \frac{\theta}{\lambda(\theta - (\lambda + \theta)\rho)}. \tag{26}$$

Remark 3. Because denominator of the r.h.s. of equality (10) must be positive, it then immediately follows that (10) is the necessary stability condition as well. Note that it can be rewritten in an equivalent form as

$$\rho < \frac{\theta}{\lambda + \theta}. \tag{27}$$

Remark 4. Note that, as $\theta \to \infty$, formula (26) transforms to expression (7) while (25) becomes (9), that corresponds to the system with infinite buffer.

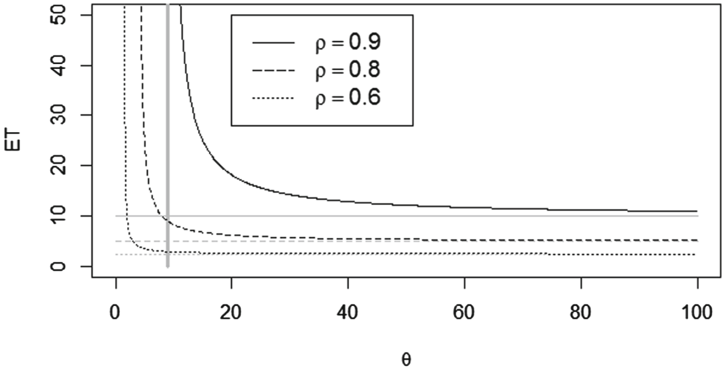


Fig. 1. The mean regeneration cycle length $\mathsf{E}T$ in the retrial system (black lines) vs. retrial rate θ; comparison with the mean regeneration cycle in the buffered system (grey lines) for $\lambda = 1$. The vertical grey line corresponds to θ^* that is the stability region boundary for $\rho = 0.9$.

Figure 1 illustrates the behavior of the $\mathsf{E}T$ depending on θ for a few values of ρ. We take $\lambda = 1$ and choose three values of traffic intensity: $\rho = 0.6$, $\rho = 0.8$ and $\rho = 0.9$. Also we compare $\mathsf{E}T$ (depending on θ in the retrial system, see (26)) with the mean regeneration cycle length $\mathsf{E}T$ in the buffered system, see (7). We note that, according to *Remark 3*, $\mathsf{E}T$ in retrial system approaches the corresponding value for the buffered system, as $\theta \to \infty$. On the other hand, as it is seen from Fig. 1, if θ approaches the stability region boundary, that is

$$\theta \downarrow \frac{\lambda\rho}{1-\rho} =: \theta^*,$$

(see condition (27)), then $\mathsf{E}T$ increases unlimitedly indicating 'increasing instability'.

4 Simulation Results

In the previous sections we have obtained explicit expressions for the mean regeneration cycle length and the mean number of customers arrived within regeneration cycle. To demonstrate these results numerically, we conduct the discrete event simulation using the 'R language'.

First we consider the retrial system and we assume that the service time has Pareto distribution

$$F(x) = 1 - \left(\frac{x_0}{x}\right)^{\alpha}, \ \ x \geq x_0.$$

In all examples, excluding Fig. 8, parameters of the Pareto service times are satisfied $\mathsf{E}S = 0.1$. We use either $x_0 = 0.05$, $\alpha = 2$ or $x_0 = 1/30$, $\alpha = 1.5$ set of parameters. Also we use Weibull distribution

$$F(x) = 1 - \exp\left\{ - \left(\frac{x}{b}\right)^{a} \right\}, \ x \geq 0,$$

with the shape parameter $a = 0.9$, while the scale parameter $b = 0.095$ which in turn implies $\mathsf{E}S = 0.1$. The input is Poisson with rate λ varying from 1 to 9, implying $\rho = 0.1, 0.2, \ldots, 0.9$, respectively. In all experiments the retrial times have exponential distribution with rate $\theta = 30$. We obtain the mean sample path of the mean $\widetilde{T}$ based on $N = 200$ paths in the fixed interval of time [0, 10000], which means that the average number of arrivals ranges from 10000 to 90000, depending on λ.

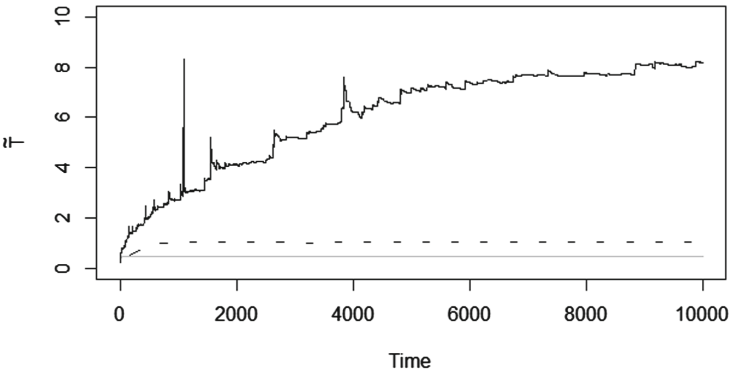


Fig. 2. The sample mean regeneration cycle length, denoted by $\widetilde{T}$, for Pareto service time with $\alpha = 2$ vs. simulation time: $\rho = 0.5$ (grey), $\rho = 0.7$ (dashed line), $\rho = 0.8$ (black). Stability condition is $\rho < 0.7913$.

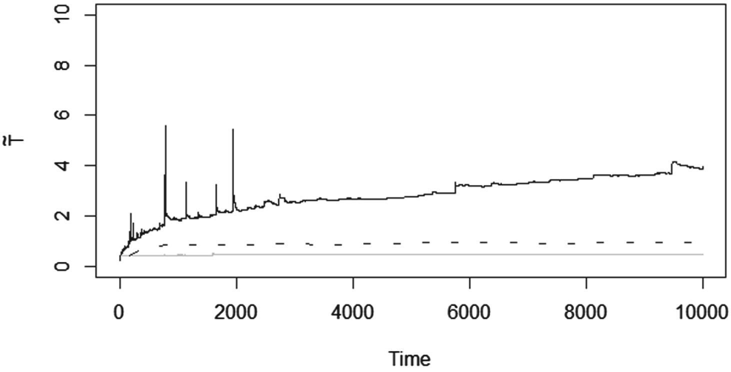


Fig. 3. The sample mean $\widetilde{T}$ for Pareto service time with $\alpha = 1.5$ vs. time; $\rho = 0.5$ (grey), $\rho = 0.7$ (dashed line), $\rho = 0.8$ (black); stability condition is $\rho < 0.7913$.

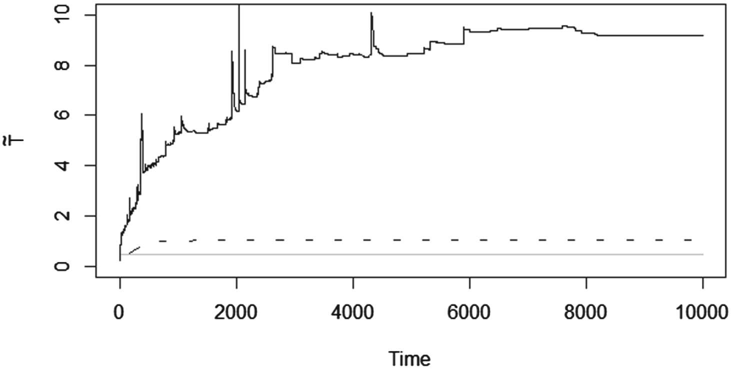


Fig. 4. The sample mean $\widetilde{T}$ for Weibull service time with $\alpha = 0.9$ vs. time; $\rho = 0.5$ (grey), $\rho = 0.7$ (dashed line), $\rho = 0.8$ (black); stability condition is $\rho < 0.7913$.

In Fig. 2, 3 and 4 we present the sample mean regeneration cycle length $\widetilde{T}$ for different service time distributions.

Then on Fig. 5, the sample mean of the number of customers arrived within a regeneration cycle, denoted by $\widetilde{\theta}$, is presented, while on Fig. 6, the sample mean of the orbit size, denoted by $\widetilde{Q}(t)$, for the different value of the traffic intensity ρ is plotted.

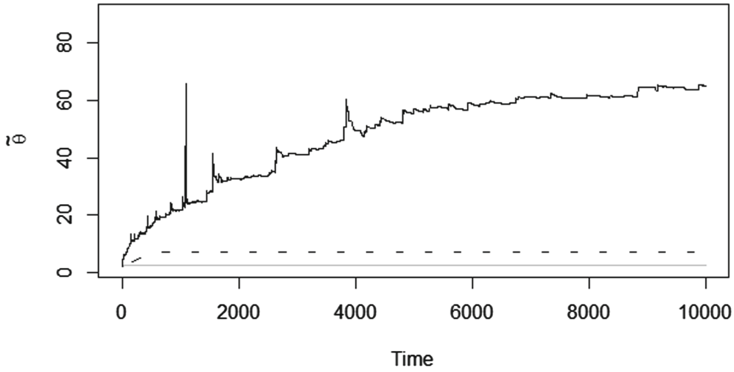


Fig. 5. The sample mean number of the customers arrived within the regeneration cycle, denoted by $\widetilde{\theta}$, for the Pareto service time with $\alpha = 2$, vs. time; $\rho = 0.5$ (grey), $\rho = 0.7$ (dashed line), $\rho = 0.8$ (black); stability condition is $\rho < 0.7913$.

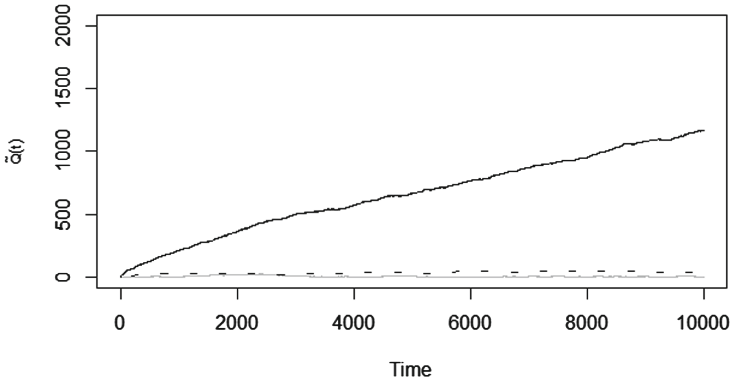


Fig. 6. The sample mean orbit size, denoted by $\widetilde{Q}(t)$, for Pareto service time with $\alpha = 2$, vs. time; $\rho = 0.5$ (grey), $\rho = 0.7$ (dashed line), $\rho = 0.8$ (black); stability condition is $\rho < 0.7913$.

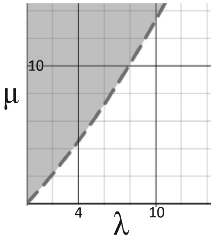


Fig. 7. The stability area (grey) of the $M/G/1$ retrial system for $\theta = 30$.

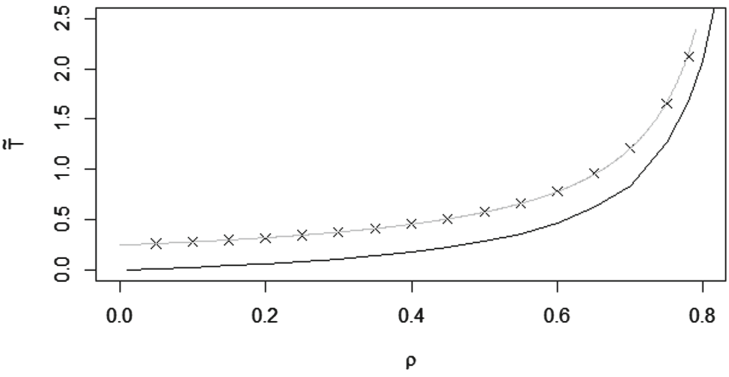


Fig. 8. The sample mean regeneration cycle length $\widetilde{T}$ (dotted line), theoretical value of $\mathsf{E}T$ (grey) and the sample mean generic busy period in regeneration cycle $\widetilde{T}_B$ (black), with Pareto service time, $\alpha = 2$, for fixed $\lambda = 4$ under stability condition $\rho < 0.8823$, vs. traffic intensity ρ.

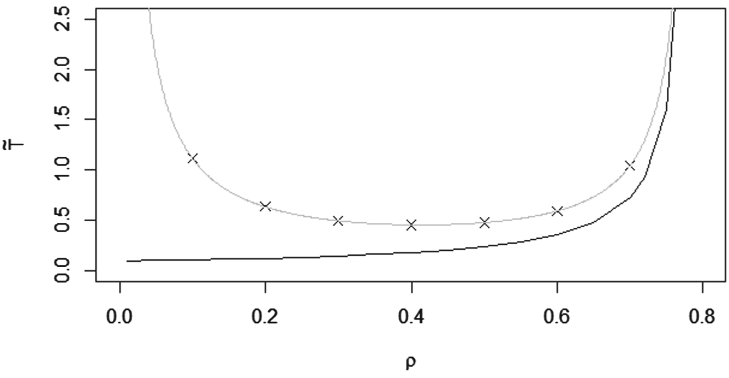


Fig. 9. The sample mean regeneration cycle length $\widetilde{T}$ (dotted line), theoretical value of $\mathsf{E}T$ (grey) and the sample mean generic busy period in regeneration cycle $\widetilde{T}_B$ (black), with Pareto service time, $\alpha = 2$, for fixed $\mu = 10$ with stability condition $\rho < 0.7913$, vs. traffic intensity ρ.

Finally, on Fig. 8, and Fig. 9, the difference between calculated mean regeneration cycle length $\widetilde{T}$ and the theoretical value mean $\mathsf{E}T$, satisfying formula (26), is demonstrated, for $\rho = 0.1, 0.2, \ldots, 0.7$. In addition, we fix $\mu = 10$ and take λ implying $\rho < 0.7913$, to satisfy stability condition (10). Also we also fix $\lambda = 4$ and take μ such that $\rho < 0.8823$. We note that in the case, where we fix $\mu = 10$, when λ approaches 0, the mean cycle length goes to infinity. It is because the intervals between arrivals become longer and as a result the 'idle part' of the regeneration period becomes longer as well.

Thus simulation results show that when stability condition (10) is satisfied (see Fig. 7) then the system remains stationary (positive recurrence), and it validates theoretical results.

5 Conclusion

We consider a $M/G/1$-type retrial system with constant retrial rate. Preliminary, we deduce a (well-known) expression for the mean regeneration cycle length in the classic $M/G/1$ queueing system with infinite buffer, and then, using a similar approach, obtain the explicit expressions for the regeneration cycle length and the number of customers arrived per cycle in the stable (positive recurrent) retrial system. As a byproduct of the analysis, using the regenerative approach, we obtain the stability criteria of the single-class retrial system. Simulation results are included to illustrate the obtained theoretical results. In a future research, it is assumed to extend the regenerative analysis to the retrial system with more general renewal input process, satisfying some monotonicity properties, to obtain the bounds for the mean regeneration cycle length.

References

1. Artalejo, J.R.: Accessible bibliography on retrial queues. Math. Comput. Modell. **30**(3–4), 1–6 (1999)
2. Artalejo, J. R.: A classified bibliography of research on retrial queues: progress in 1990-1999. TOP Off. J. Spanish Soc. Stat. Oper. Res. **7**(2), 187–211 (1999)
3. Artalejo, J.R., Falin, G.I.: Standard and retrial queueing systems: a comparative analysis. Revista Matematica Complutense **15**(1), 101–129 (2002)
4. Artalejo, J.R., Gómez-Corral, A.: Retrial Queueing Systems. A Computational Approach. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78725-9
5. Asmussen, S.: Applied Probability and Queues. Wiley, New York (1987)
6. Borovkov, A.A.: Probability Theory. Springer, New York (2013). https://doi.org/10.1007/978-1-4471-5201-9
7. Falin, G.I., Templeton, J.G.C.: Retrial Queues. Chapman and Hall, London (1997)
8. Marshall A., Olkin, I.: Life Distributions. Structure of Nonparametric, Semiparametric, and Parametric Families. Springer, New York (2007). https://doi.org/10.1007/978-0-387-68477-2
9. Morozov, E., Steyaert, B.: Stability Analysis of Regenerative Queueing Models. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-82438-9
10. Morozov, E., Delgado, R.: Stability analysis of regenerative queues. Autom. Remote Control. **70**(12), 1977–1991 (2009)
11. Serfozo, R.: Basics of Applied Stochastic Processes. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-540-89332-5
12. Sigman, K., Wolff, R.W.: A review of regenerative processes. SIAM Rev. **35**(2), 269–288 (1993)
13. Smith, W.L.: Regenerative stochastic processes. Proc. Royal Soc. Ser. A **232**(1188), 6–31 (1955)
14. Smith, W. L.: Renewal theory and its ramifications. J. Royal Stat. Soc. Ser. B (Methodol.) **20**(2), 243–302 (1958)

The Method of Marginal Asymptotic-Diffusion Analysis for Multiclass Retrial Queues

Anatoly Nazarov, Ekaterina Fedorova(✉), and Nikita Kostryukov

National Research Tomsk State University, Lenina Avenue, 36, Tomsk, Russia
moiskate@mail.ru

Abstract. In the paper, a single-server retrial queueing system with heterogeneous customers is considered. Customers of a finite number of classes arrive to the system in the Poisson stationary processes. Service times and inter-retrial times have exponential distributions with the rates depending on the customers class. A stationary probability distribution of the number of customers of each class in the orbit is found by the new method of marginal asymptotic-diffusion analysis under the condition of a long delay of customers in the orbit.

Keywords: multiclass retrial queues · heterogeneous customers · marginal asymptotic-diffusion method

1 Introduction

Retrial queueing systems are appropriate models of various communication systems (call-centers, cellular networks, LANs, etc. [1,2]), in which there are repeated attempts to get service. Usually in such systems, there is not a classical queue, where customers (clients, calls, data packages, etc.) wait. But also customers do not refuse service in the case of a busy server (operator, channel, etc.).

Retrial queues (RQ) is a class of queuing theory models. Its description is presented in [1–3]. In retrial queues, there is an "orbit", which is a some virtual place for repeated calls, where calls wait random time. In classical retrial queue, a random access protocol for calls in orbit takes place, i.e. any call has access to server at time t. In spite of the large number of studies in RQ, heterogeneous RQ is weak considered. The reason is a large dimension of the mathematical problem. When we consider a model with N types of customers, it is necessary to study a $(N+1)$-dimensional random process, the components of which can take on an infinite number of states. Therefore, it is difficult to study such models using numerical methods or simulation. Multiclass RQs (with several types of customers or several orbits) are studied by E. Morozov [4,5], A.Krishnamoorthy [6], B. Kim [7,8], etc. [9–11]. Most of them [4,5,7–9] are devoted stability analysis of multiclass retrial queues as classical, as with constant retrial rate. While

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 300–315, 2024.
https://doi.org/10.1007/978-3-031-65385-8_22

probability distributions or even means of processes under study are hardly investigated.

The novelty of the paper lies in the methodology of research for multi-class retrial queues. We propose a new method of marginal asymptotic-diffusion analysis for obtaining marginal asymptotic probability distributions of the number of calls of each class in the orbit. This method generalize the method of asymptotic-diffusion analysis (presented in [12]) for the case of multiclass queueing systems.

The rest of the paper is organized as follows. In Sect. 2, the model under study is described and the statement of the problem is formulated. Section 3 is devoted to the original marginal asymptotic-diffusion analysis method for multi-class retrial queues. In Sect. 4, we demonstrate some numerical examples with comparison of simulation and asymptotic distributions for various values of the model parameters. Section 5 consists some conclusions.

2 Mathematical Model

In the paper, a retrial queueing system with heterogeneous customers is considered. We suppose that N classes of customers exist. So, there are N arrival Poisson processes with rates λ_n, $n = \overline{1,N}$. The system has one server. The service time of the n-th class customer is exponential distributed with rate μ_n. Unserved calls go to an orbit, where they wait for random time distributed exponentially with corresponding rate σ_n. There is multiple access protocol from orbits. Orbit capacity are not limited. From the orbit, calls again turn to the server. If the server is free, a call begins its service, otherwise it returns up to the orbit to make a next attempt.

Note that it does not matter to consider a retrial queue with one common orbit or a retrial queue with several orbits (for each class of customers). It is important to distinguish a number of customers of each class. We illustrate the model structure as on Fig. 1.

Let us random processes of the number of the n-th class calls in the orbit be $i_n(t)$, where $n = \overline{1,N}$. Process $k(t)$ determines states of the server as follows:

$$k(t) = \begin{cases} 0, \text{if the server is free,} \\ n, \text{if the } n\text{-th class customer is on the server.} \end{cases}$$

Denote $P\{k(t) = k, i_1(t) = i_1, i_2(t) = i_2, ..., i_N(t) = i_N\} = P(k, \mathbf{i}, t)$ be the probability that the server is in state k and there are $\mathbf{i} = \{i_1, i_2, ..., i_N\}$ calls in the orbit at time t, where i_n is the number of the n-th class calls in the orbit ($n = \overline{1,N}$). Process $\{k(t), \mathbf{i}(t)\}$ is Markovian, so for probability distribution $P(k, \mathbf{i}, t)$, we compose the following system of Kolmogorov equations:

$$\begin{cases} \dfrac{\partial P(0,\mathbf{i},t)}{\partial t} = -\left(\sum\limits_{n=1}^{N} \lambda_n + \sum\limits_{n=1}^{N} i_n\sigma_n\right) P(0,\mathbf{i},t) + \sum\limits_{k=1}^{N} \mu_k P(k,\mathbf{i},t), \\ \frac{\partial P(k,\mathbf{i},t)}{\partial t} = -\left(\sum_{n=1}^{N} \lambda_n + \mu_k\right) P(k,\mathbf{i},t) + \lambda_k P(0,\mathbf{i},t) + \\ +(i_k+1)\sigma_k P(0,\mathbf{i}+\mathbf{e}_k,t) + \sum_{n=1}^{N} \lambda_n P(k,\mathbf{i}-\mathbf{e}_k,t),\ k = \overline{1,N}, \end{cases} \tag{1}$$

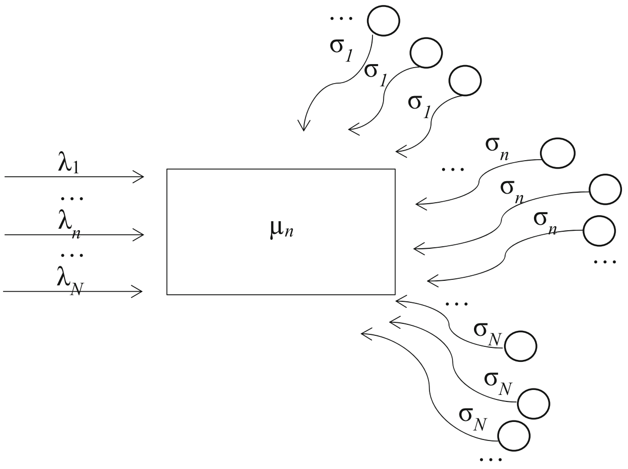


Fig. 1. Multiclass Retrial Queue

where $\mathbf{e}_k$ is a vector with the k-th element being one and others elements being zero.

Let us introduce the partial characteristic functions:

$$H(k,\mathbf{u},t) = \sum_{i_1=0}^{\infty} ... \sum_{i_N=0}^{\infty} e^{ju_1 i_1} \cdot ... \cdot e^{ju_N i_N} P(k,\mathbf{i},t).$$

By rewriting Eqs. (1) for the characteristic functions, we have

$$\begin{cases} \frac{\partial H(0,\mathbf{u},t)}{\partial t} = -H(0,\mathbf{u},t) \sum\limits_{n=1}^{N} \lambda_n + \sum\limits_{n=1}^{N} j\sigma_n \dfrac{\partial H(0,\mathbf{u},t)}{\partial u_n} \\ + \sum\limits_{k=1}^{N} \mu_k H(k,\mathbf{u},t), \\ \frac{\partial H(k,\mathbf{u},t)}{\partial t} = \left(\sum\limits_{n=1}^{N} \lambda_n (e^{ju_n} - 1) - \mu_k \right) H(k,\mathbf{u},t) \\ +\lambda_k H(0,\mathbf{u},t) - j\sigma_k e^{-ju_k} \frac{\partial H(0,\mathbf{u},t)}{\partial u_k}, \ k = \overline{1,N}. \end{cases} \tag{2}$$

3 Marginal Asymptotic-Diffusion Analysis

The marginal asymptotic-Diffusion analysis consists of several stages:

1. Derivation of "marginal" asymptotic equations for "marked" process $i_n(t)$ (a number of calls of the n-th class in the orbit).
2. Finding of asymptotic means of numbers of calls of each class and stationary probabilities of server states.
3. Implementation of Asymptotic-Diffusion Analysis for "marked" process $i_n(t)$.

Let us solve System (2) by the method of the marginal asymptotic-diffusion analysis under the limit condition of a long delay $\sigma \to 0$.

3.1 Marginal Asymptotic Equations

First of all, we mark the n-th class of calls, and try to write asymptotic equations for process $i_n(t)$.

Let us introduce infinitesimal parameter ε and substitutions

$$\sigma = \varepsilon, \quad u_v = \varepsilon w_v,$$

where

$$\sigma_v = \gamma_v \sigma,\ v = \overline{1, N},\ v \neq n,$$

Also, we denote

$$\mathbf{w}^{(n)} = \{w_1, w_2, ..., w_{n-1}, u, w_{n+1}, ..., w_N\}, \quad H(k, \mathbf{u}, t) = F(k, \mathbf{w}^{(n)}, t).$$

Then we have the following system of asymptotic equations:

$$\begin{cases} \dfrac{\partial F(0,\mathbf{w}^{(n)},t,\varepsilon)}{\partial t} = -F(0,\mathbf{w}^{(n)},t,\varepsilon)\displaystyle\sum_{v=1}^{N}\lambda_v + \sum_{\substack{v=1\\ v\neq n}}^{N} j\gamma_v \dfrac{\partial F(0,\mathbf{w}^{(n)},t,\varepsilon)}{\partial w_v} \\ +j\sigma_n \dfrac{\partial F(0,\mathbf{w}^{(n)},t,\varepsilon)}{\partial u_n} + \displaystyle\sum_{k=1}^{N}\mu_k F(k,\mathbf{w}^{(n)},t,\varepsilon), \\ \dfrac{\partial F(k,\mathbf{w}^{(n)},t,\varepsilon)}{\partial t} = \lambda_k F(0,\mathbf{w}^{(n)},t,\varepsilon) - j\gamma_k e^{-j\varepsilon w_k}\dfrac{\partial F(0,\mathbf{w}^{(n)},t,\varepsilon)}{\partial w_k} \\ +\left(\displaystyle\sum_{\substack{v=1\\ v\neq n}}^{N}\lambda_v(e^{j\varepsilon w_v}-1) + \lambda_n(e^{ju_n}-1) - \mu_k\right)F(k,\mathbf{w}^{(n)},t,\varepsilon),\ k\neq n, \\ \dfrac{\partial F(n,\mathbf{w}^{(n)},t,\varepsilon)}{\partial t} = \lambda_n F(0,\mathbf{w}^{(n)},t,\varepsilon) - j\sigma_n e^{-ju_n}\dfrac{\partial F(0,\mathbf{w}^{(n)},t,\varepsilon)}{\partial u_n} \\ +\left(\displaystyle\sum_{\substack{v=1\\ v\neq n}}^{N}\lambda_v(e^{j\varepsilon w_v}-1) + \lambda_n(e^{ju_n}-1) - \mu_n\right)F(n,\mathbf{w}^{(n)},t,\varepsilon). \end{cases}$$

By writing the equations under limit $\varepsilon \to 0$, we obtain

$$\begin{cases} \dfrac{\partial F(0,\mathbf{w}^{(n)},t)}{\partial t} = -F(0,\mathbf{w}^{(n)},t)\displaystyle\sum_{v=1}^{N}\lambda_v + \sum_{\substack{v=1\\ v\neq n}}^{N} j\gamma_v \dfrac{\partial F(0,\mathbf{w}^{(n)},t)}{\partial w_v} \\ +j\sigma_n \dfrac{\partial F(0,\mathbf{w}^{(n)},t)}{\partial u_n} + \displaystyle\sum_{k=1}^{N}\mu_k F(k,\mathbf{w}^{(n)},t), \\ \dfrac{\partial F(k,\mathbf{w}^{(n)},t)}{\partial t} = \left(\lambda_n(e^{ju_n}-1) - \mu_k\right)F(k,\mathbf{w}^{(n)},t) + \lambda_k F(0,\mathbf{w}^{(n)},t) \\ -j\gamma_k \dfrac{\partial F(0,\mathbf{w}^{(n)},t)}{\partial w_k},\ k\neq n, \\ \dfrac{\partial F(n,\mathbf{w}^{(n)},t)}{\partial t} = \left(\lambda_n(e^{ju_n}-1) - \mu_n\right)F(n,\mathbf{w}^{(n)},t) + \lambda_n F(0,\mathbf{w}^{(n)},t) \\ -j\sigma_n e^{-ju_n}\dfrac{\partial F(0,\mathbf{w}^{(n)},t)}{\partial u_n}. \end{cases} \tag{3}$$

Let the solution have the following form:

$$F(k,\mathbf{w}^{(n)},t)=H_n(k,u_n,t)\exp\left\{\sum_{v\neq n} jw_v x_v\right\}.$$

System (3) is rewritten as follows

$$\begin{cases} \dfrac{\partial H_n(0,u_n,t)}{\partial t}=-H_n(0,u_n,t)\left(\displaystyle\sum_{v=1}^{N}\lambda_v+\sum_{\substack{v=1\\ v\neq n}}^{N}\gamma_v x_v\right)\\ +j\sigma_n\frac{\partial H_n(0,u_n,t)}{\partial u_n}+\sum\limits_{k=1}^{N}\mu_k H_n(k,u_n,t),\\ \frac{\partial H_n(k,u_n,t)}{\partial t}=\big(\lambda_n(e^{ju_n}-1)-\mu_k\big)H_n(k,u_n,t)\\ +H_n(0,u_n,t)(\lambda_k+\gamma_k x_k),\ k\neq n,\\ \frac{\partial H_n(n,u_n,t)}{\partial t}=\big(\lambda_n(e^{ju_n}-1)-\mu_n\big)H_n(n,u_n,t)+\lambda_n H_n(0,u_n,t)\\ -j\sigma_n e^{-ju_n}\frac{\partial H_n(0,u_n,t)}{\partial u_n}. \end{cases} \tag{4}$$

Thus, we obtained the system of equations for the marginal asymptotic characteristic functions of the number of calls of the n-th class.

3.2 Asymptotic Means

The next step of the study is finding of parameters x_k, $k=\overline{1,N}$. Let us write System (2) in the steady state.

$$\begin{cases} -H(0,\mathbf{u})\sum\limits_{n=1}^{N}\lambda_n+\sum\limits_{n=1}^{N}j\sigma_n\frac{\partial H(0,\mathbf{u})}{\partial u_n}+\sum\limits_{k=1}^{N}\mu_k H(k,\mathbf{u})=0,\\ \left(\sum\limits_{n=1}^{N}\lambda_n(e^{ju_n}-1)-\mu_k\right)H(k,\mathbf{u})+\lambda_k H(0,\mathbf{u})\\ -j\sigma_k e^{-ju_k}\frac{\partial H(0,\mathbf{u})}{\partial u_k}=0,\ k=\overline{1,N}. \end{cases} \tag{5}$$

The sum of this equations gives us an additional equation:

$$\sum_{n=1}^{N} j\sigma_n(1-e^{-ju_n})\frac{\partial H(0,\mathbf{u})}{\partial u_n}+\sum_{k=1}^{N}H(k,\mathbf{u})\sum_{n=1}^{N}\lambda_n(e^{ju_n}-1)=0. \tag{6}$$

We use the following substitutions:

$$\sigma_n=\gamma_n\sigma,\ \sigma=\varepsilon,\ u_n=\varepsilon w_n,\ H(k,\mathbf{u})=F(k,\mathbf{w},\varepsilon),\ n=\overline{1,N}.$$

Thus, the following system is obtained.

$$\begin{cases} -F(0,\mathbf{w},\varepsilon)\sum\limits_{n=1}^{N}\lambda_n + \sum\limits_{n=1}^{N} j\gamma_n \frac{\partial F(0,\mathbf{w},\varepsilon)}{\partial w_n} + \sum\limits_{k=1}^{N}\mu_k F(k,\mathbf{w},\varepsilon) = O(\varepsilon), \\ \left(\sum\limits_{n=1}^{N}\lambda_n(e^{j\varepsilon w_n}-1)-\mu_k\right)F(k,\mathbf{w},\varepsilon)+\lambda_k F(0,\mathbf{w},\varepsilon) \\ -j\gamma_k e^{-j\varepsilon w_k}\frac{\partial F(0,\mathbf{w},\varepsilon)}{\partial w_k} = O(\varepsilon),\ k=\overline{1,N}, \\ \sum\limits_{n=1}^{N} j\gamma_n j w_n \frac{\partial F(0,\mathbf{w},\varepsilon)}{\partial w_n} + \sum_{k=1}^{N} F(k,\mathbf{w},\varepsilon)\sum\limits_{n=1}^{N}\lambda_n j w_n = O(\varepsilon). \end{cases} \tag{7}$$

For $\varepsilon \to 0$, we have

$$\begin{cases} -F(0,\mathbf{w})\sum\limits_{n=1}^{N}\lambda_n + \sum\limits_{n=1}^{N} j\gamma_n \frac{\partial F(0,\mathbf{w})}{\partial w_n} + \sum\limits_{k=1}^{N}\mu_k F(k,\mathbf{w}) = 0, \\ -\mu_k F(k,\mathbf{w}) + \lambda_k F(0,\mathbf{w}) - j\gamma_k \frac{\partial F(0,\mathbf{w})}{\partial w_k} = 0,\ k=\overline{1,N}, \\ \sum\limits_{n=1}^{N} j\gamma_n j w_n \frac{\partial F(0,\mathbf{w})}{\partial w_n} + \sum\limits_{k=1}^{N} F(k,\mathbf{w})\sum\limits_{n=1}^{N}\lambda_n j w_n = 0. \end{cases} \tag{8}$$

Let us find the solution in the following form:

$$F(k,\mathbf{w}) = r_k \cdot \exp\left\{\sum_{n=1}^{N} j w_n x_n\right\}, \quad k=\overline{1,N}.$$

It's easy to obtain the system

$$\begin{cases} -r_0\sum\limits_{n=1}^{N}\lambda_n - \sum\limits_{n=1}^{N}\gamma_n r_0 x_n + \sum\limits_{k=1}^{N}\mu_k r_k = 0, \\ -\mu_k r_k + \lambda_k r_0 + \gamma_k r_0 x_k = 0,\ k=\overline{1,N}, \\ \sum\limits_{n=1}^{N} j w_n(-\gamma_n x_n r_0 + \lambda_n \sum\limits_{k=1}^{N} r_k) = 0. \end{cases} \tag{9}$$

So, r_k are expressed as

$$r_k = r_0\frac{\lambda_k+\gamma_k x_k}{\mu_k},\ k=\overline{1,N}.$$

From the normalization condition, we have

$$r_0 = \frac{1}{1+\sum\limits_{k=1}^{N}\frac{\lambda_k+\gamma_k x_k}{\mu_k}}. \tag{10}$$

Denote

$$\kappa_v = \lambda_v + \gamma_v x_v,$$

then we get

$$r_0 = \frac{\lambda_n}{\kappa_n}, \quad r_k = \frac{\lambda_k}{\kappa_k},\ k=\overline{1,N.} \tag{11}$$

From the last equation of System (9), we obtain

$$\lambda_n(1-r_0)-\gamma_n x_n r_0=0.$$

Then κ_n are as follows

$$\kappa_n=\frac{\lambda_n}{1-\sum\limits_{k=1}^{N}\frac{\lambda_k}{\mu_k}}. \tag{12}$$

3.3 Marginal Asymptotic-Diffusion Probabilities

Let us solve asymptotic Eqs. (4) for marked the n-th process. First of all, we summarize the equations for getting an additional equation:

$$\frac{\partial H_n(u_n,t)}{\partial t}=(e^{ju_n}-1)\left(j\sigma_n e^{-ju_n}\frac{\partial H_n(0,u_n,t)}{\partial u_n}+\lambda_n\sum_{k=1}^{N}H_u(k,u_n,t)\right). \tag{13}$$

First Asymptotics. Let us denote

$$\sigma_n=\varepsilon,\ \sigma_n t=\varepsilon t=\tau,\ u_n=\varepsilon w,\ H_n(k,u_n,t)=F_n(k,w,\tau,\varepsilon),\ k=\overline{1,N}.$$

By substituting the notations in Eqs. (4), we have

$$\begin{cases}\varepsilon\dfrac{\partial F_n(0,w,\tau,\varepsilon)}{\partial\tau}=-F_n(0,w,\tau,\varepsilon)\left(\lambda_n+\sum\limits_{\substack{v=1\\v\neq n}}^{N}\kappa_v\right)+j\dfrac{\partial F_n(0,w,\tau,\varepsilon)}{\partial w}\\+\sum\limits_{k=1}^{N}\mu_k F_n(k,w,\tau,\varepsilon),\\\varepsilon\dfrac{\partial F_n(k,w,\tau,\varepsilon)}{\partial\tau}=\left(\lambda_n(e^{j\varepsilon w}-1)-\mu_k\right)F_n(k,w,\tau,\varepsilon)\\+F_n(0,w,\tau,\varepsilon)\kappa_k,\ \ k\neq n,\\\varepsilon\dfrac{\partial F_n(n,w,\tau,\varepsilon)}{\partial\tau}=\left(\lambda_n(e^{j\varepsilon w}-1)-\mu_n\right)F_n(n,w,\tau,\varepsilon)+\lambda_n F_n(0,w,\tau,\varepsilon)\\-je^{-j\varepsilon w}\dfrac{\partial F_n(0,w,\tau,\varepsilon)}{\partial w}.\end{cases} \tag{14}$$

And from Eq. (13), there is the following equation

$$\varepsilon\sum_{k=0}^{N}\frac{\partial F_n(k,w,\tau,\varepsilon)}{\partial\tau}$$
$$=(1-e^{-j\varepsilon w})\left(j\frac{\partial F_n(0,w,\tau,\varepsilon)}{\partial w}+\lambda_n e^{j\varepsilon w}\sum_{k=1}^{N}F_n(k,w,\tau,\varepsilon)\right). \tag{15}$$

Under the limit $\varepsilon \to 0$, Eqs. (14) are rewritten as

$$\begin{cases} -F_n(0,w,\tau)\left(\lambda_n + \sum\limits_{\substack{v=1\\ v\neq n}}^{N}\kappa_v\right) + j\dfrac{\partial F_n(0,w,\tau)}{\partial w} + \sum\limits_{k=1}^{N}\mu_k F_n(k,w,\tau) = 0, \\ -\mu_k F_n(k,w,\tau) + F_n(0,w,\tau)\kappa_k = 0,\ \ k \neq n, \\ -\mu_n F_n(n,w,\tau) + \lambda_n F_n(0,w,\tau) - j\dfrac{\partial F_n(0,w,\tau)}{\partial w} = 0. \end{cases} \tag{16}$$

Let us find the solution in the following form

$$F_n(k,w,\tau) = R_k(x(\tau))e^{jwx(\tau)}. \tag{17}$$

For simplicity, in further derivation, we omit τ, so we use notation $x = x(\tau)$.

Substituting into Eqs. (16), we get:

$$R_k(x) = R_0(x)\frac{\kappa_k}{\mu_k},\ \text{for}\ \ k \neq n, \tag{18}$$

$$R_n(x) = R_0(x)\frac{\lambda_n + x}{\mu_n}.$$

From the normalization condition:

$$R_0(x) = \left(1 + \frac{\lambda_n + x}{\mu_n} + \sum_{\substack{k=1\\ k\neq n}}^{N}\frac{\kappa_k}{\mu_k}\right)^{-1}.$$

Returning to Eq. (15), we use Maclaurin series.

$$\sum_{k=0}^{N}\frac{\partial F_n(k,w,\tau,\varepsilon)}{\partial \tau}$$
$$= jw\left(j\frac{\partial F_n(0,w,\tau,\varepsilon)}{\partial w} + \lambda_n\sum_{k=1}^{N}F_n(k,w,\tau,\varepsilon)\right) + O(\varepsilon).$$

Under the limit $\varepsilon \to 0$, we obtain the following equation

$$\sum_{k=0}^{N}\frac{\partial F_n(k,w,\tau)}{\partial \tau} = jw\left(j\frac{\partial F_n(0,w,\tau)}{\partial w} + \lambda_n\sum_{k=1}^{N}F_n(k,w,\tau)\right). \tag{19}$$

Substituting (17) into (19), we get:

$$\frac{dx(\tau)}{d\tau} = \left(-R_0(x)\cdot x + \lambda_n\sum_{k=1}^{N}R_k(x)\right).$$

A derived equation is an equation for the asymptotic mean $x(\tau)$ of the number of calls of the n-th class with a transfer coefficient:

$$a(x(\tau)) = \lambda_n(1 - R_0(x(\tau))) - R_0(x(\tau)) \cdot x(\tau). \tag{20}$$

Second Asymptotics. We suppose that

$$H_n(k, u_n, t) = H_n^{(2)}(k, u_n, t) \exp\left\{\frac{ju_n}{\sigma_n}x\right\}, \tag{21}$$

where $x = x(\sigma_n t)$. Substituting (21) into System (4), we have:

$$\begin{cases} \dfrac{\partial H_n^{(2)}(0, u_n, t)}{\partial t} + ju_n a(x) H_n^{(2)}(0, u_n, t) = j\sigma_n \dfrac{\partial H_n^{(2)}(0, u_n, t)}{\partial u_n} \\ -H_n^{(2)}(0, u_n, t)\left(\lambda_n + \displaystyle\sum_{\substack{v=1 \\ v \neq n}}^{N} \kappa_v\right) - H_n^{(2)}(0, u_n, t)x \\ + \displaystyle\sum_{k=1}^{N} \mu_k H_n^{(2)}(k, u_n, t), \\ \dfrac{\partial H_n^{(2)}(k, u_n, t)}{\partial t} + ju_n a(x) H_n^{(2)}(k, u_n, t) = \\ + \left(\lambda_n(e^{ju_n} - 1) - \mu_k\right) H_n^{(2)}(k, u_n, t) + H_n^{(2)}(0, u_n, t)\kappa_k, \ k \neq n, \\ \dfrac{\partial H_n^{(2)}(n, u_n, t)}{\partial t} + ju_n a(x) H_n^{(2)}(n, u_n, t) = \lambda_n H_n^{(2)}(0, u_n, t) \\ + \left(\lambda_n(e^{ju_n} - 1) - \mu_n\right) H_n^{(2)}(n, u_n, t) \\ -j\sigma_n e^{-ju_n} \dfrac{\partial H_n^{(2)}(0, u_n, t)}{\partial u_n} + e^{-ju_n} H_n^{(2)}(0, u_n, t)x. \end{cases} \tag{22}$$

By summing of Eq. (22), we get one more equation:

$$\frac{\partial H_n^{(2)}(u_n, t)}{\partial t} + ju_n a(x) H_n^{(2)}(u_n, t) = (e^{ju_n} - 1)$$
$$\times \left(j\sigma_n e^{-ju_n} \frac{\partial H_n^{(2)}(0, u_n, t)}{\partial u_n} - e^{-ju_n} H_n^{(2)}(0, u_n, t)x + \lambda_n \sum_{k=1}^{N} H_n^{(2)}(k, u_n, t)\right). \tag{23}$$

Let us introduce the following notations:

$$\sigma_n = \varepsilon^2, \quad \sigma_n t = \varepsilon^2 t = \tau, \quad u_n = \varepsilon w, \quad H_n^{(2)}(k, u_n, t) = F_n^{(2)}(k, w, \tau, \varepsilon).$$

Substituting into System (22), we have

$$\begin{cases} \varepsilon^2 \dfrac{\partial F_n^{(2)}(0,w,\tau,\varepsilon)}{\partial \tau} + j\varepsilon w a(x) F_n^{(2)}(0,w,\tau,\varepsilon) = j\varepsilon \dfrac{\partial F_n^{(2)}(0,w,\tau,\varepsilon)}{\partial w} \\ -F_n^{(2)}(0,w,\tau,\varepsilon) \left(\lambda_n + \displaystyle\sum_{\substack{v=1 \\ v\neq n}}^{N} \kappa_v \right) - F_n^{(2)}(0,w,\tau,\varepsilon)x + \displaystyle\sum_{k=1}^{N} \mu_k F_n^{(2)}(k,w,\tau,\varepsilon), \\ \varepsilon^2 \dfrac{\partial F_n^{(2)}(k,w,\tau,\varepsilon)}{\partial \tau} + j\varepsilon w a(x) F_n^{(2)}(k,w,\tau,\varepsilon) = \\ + \left(\lambda_n (e^{j\varepsilon w} - 1) - \mu_k \right) F_n^{(2)}(n,w,\tau,\varepsilon) + F_n^{(2)}(0,w,\tau,\varepsilon)\kappa_k, \;\; k \neq n, \\ \varepsilon^2 \dfrac{\partial F_n^{(2)}(n,w,\tau,\varepsilon)}{\partial \tau} + j\varepsilon w a(x) F_n^{(2)}(n,w,\tau,\varepsilon) = \lambda_n F_n^{(2)}(0,w,\tau,\varepsilon) \\ + \left(\lambda_n (e^{j\varepsilon w} - 1) - \mu_n \right) F_n^{(2)}(n,w,\tau,\varepsilon) - j\varepsilon e^{-j\varepsilon w} \dfrac{\partial F_n^{(2)}(0,w,\tau,\varepsilon)}{\partial w} \\ + e^{-j\varepsilon w} F_n^{(2)}(0,w,\tau,\varepsilon)x, \end{cases} \tag{24}$$

And from (23), we obtain

$$\varepsilon^2 \frac{\partial F_n^{(2)}(w,\tau,\varepsilon)}{\partial \tau} + j\varepsilon w a(x) F_n^{(2)}(w,\tau,\varepsilon) = (e^{j\varepsilon w} - 1) \\ \times \left(j\varepsilon e^{-j\varepsilon w} \frac{\partial F_n^{(2)}(0,w,\tau,\varepsilon)}{\partial w} - x e^{-j\varepsilon w} F_n^{(2)}(0,w,\tau,\varepsilon) + \lambda_n \sum_{k=1}^{N} F_n^{(2)}(k,w,\tau,\varepsilon) \right). \tag{25}$$

Let us find the solution in the following form

$$F_n^{(2)}(k,w,\tau,\varepsilon) = \Phi_n(x,w,\tau)(R_k(x) + jw\varepsilon f_k(x)) + O(\varepsilon^2). \tag{26}$$

By substituting (26) into System (24), we obtain

$$\begin{cases} \varepsilon \dfrac{\partial \Phi_n(w,\tau)}{\partial \tau} R_0(x) + jwa(x)\Phi_n(w,\tau)(R_0(x) + jw\varepsilon f_0(x)) = \\ -\Phi_n(w,\tau) jw f_0(x) \left(\lambda_n + \displaystyle\sum_{\substack{v=1 \\ v\neq n}}^{N} \kappa_v \right) + j \dfrac{\partial \Phi_n(w,\tau)}{\partial w} (R_0(x) + jw\varepsilon f_0(x)) \\ -\varepsilon \Phi_n(w,\tau) f_0(x) - \Phi_n(w,\tau) jw f_0(x) \cdot x + \displaystyle\sum_{k=1}^{N} \mu_k \Phi_n(w,\tau) jw f_k(x) + O(\varepsilon^2), \\ \varepsilon \dfrac{\partial \Phi_n(w,\tau)}{\partial \tau} R_k(x) + jwa(x)\Phi_n(w,\tau)(R_k(x) + jw\varepsilon f_k(x)) = \\ + \lambda_n \left(jw + \varepsilon \dfrac{(jw)^2}{2} \right) \Phi_n(w,\tau) R_k(x) + (\lambda_n j\varepsilon w - \mu_k)\Phi_n(w,\tau) jw f_k(x) \\ + \Phi_n(w,\tau) jw f_0(x)\kappa_k + O(\varepsilon^2), \;\; k \neq n, \\ \varepsilon \dfrac{\partial \Phi_n(w,\tau)}{\partial \tau} R_n(x) + jwa(x)\Phi_n(w,\tau)(R_n(x) + jw\varepsilon f_n(x)) = \\ + \lambda_n \left(jw + \varepsilon \dfrac{(jw)^2}{2} \right) \Phi_n(w,\tau) R_n(x) + (\lambda_n j\varepsilon w - \mu_n)\Phi_n(w,\tau) jw f_n(x) \\ + \lambda_n \Phi_n(w,\tau) jw f_0(x) - j(1 - j\varepsilon w) \dfrac{\partial \Phi_n(w,\tau)}{\partial w} (R_0 + jw\varepsilon f_0(x)) \\ + \Phi_n(w,\tau)\varepsilon f_0(x) + \left(-jw + \varepsilon \dfrac{(jw)^2}{2} \right) \Phi_n(w,\tau) R_0(x) \cdot x \\ + (1 - j\varepsilon w)\Phi_n(w,\tau) jw f_0(x) \cdot x + O(\varepsilon^2). \end{cases}$$

For $\varepsilon \to 0$, we have the following equations

$$\begin{cases} f_0(x)\left(\lambda_n + \sum\limits_{\substack{v=1\\ v\neq n}}^{N} \kappa_v + x\right) - \sum\limits_{k=1}^{N} \mu_k f_k(x) = \\ -a(x)R_0(x) + R_0(x)\dfrac{\partial \Phi_n(w,\tau)/\partial w}{w\Phi_n(w,\tau)}, \\ f_0(x)\kappa_k - \mu_k f_k(x) = R_k(x)(a(x) - \lambda_n),\ k \neq n, \\ f_0(x)(\lambda_n + x) - \mu_n f_n(x) = \\ R_n(x)(a(x) - \lambda_n) + R_0(x)\cdot x + R_0(x)\dfrac{\partial \Phi_n(w,\tau)/\partial w}{w\Phi_n(w,\tau)}. \end{cases} \tag{27}$$

Comparing this equations and System (19), we conclude that

$$f_k(x) = CR_k(x) + g_k(x) - \phi_k(x)\frac{\partial \Phi_n(w,\tau)/\partial w}{w\Phi_n(w,\tau)}, \tag{28}$$

where $C = const.$

Substituting (28) into System (27), we obtain equations for $\phi(x)$:

$$\begin{cases} -\phi_0(x)\left(\lambda_n + \sum\limits_{\substack{v=1\\ v\neq n}}^{N} \kappa_v + x\right) + \sum\limits_{k=1}^{N} \mu_k \phi_k(x) = R_0(x), \\ \phi_0(x)\kappa_k - \mu_k \phi_k(x) = 0,\ k \neq n, \\ \phi_0(x)(\lambda_n + x) - \mu_n \phi_n(x) = -R_0(x). \end{cases} \tag{29}$$

For the solution uniqueness, we suppose

$$\sum_{k=0}^{N} \phi_k(x) = 0.$$

By comparing Eqs. (29) and (19), it is obvious that

$$\phi_k(x) = \frac{dR_k(x)}{dx}. \tag{30}$$

Also from (27), we obtain the following equations for g_k:

$$\begin{cases} -g_0(x)\left(\lambda_n + \sum\limits_{\substack{v=1\\ v\neq n}}^{N} +\kappa_v + x\right) + \sum\limits_{k=1}^{N} \mu_k g_k(x) = a(x)R_0(x), \\ g_0(x)\kappa_k - \mu_k g_k(x) = R_k(x)(a(x) - \lambda_n),\ k \neq n, \\ g_0(x)(\lambda_n + x) - \mu_n g_n(x) = R_n(x)(a(x) - \lambda_n) + R_0(x)\cdot x. \end{cases} \tag{31}$$

Let us add an additional condition for the solution uniqueness:

$$\sum_{k=0}^{N} g_k(x) = 0.$$

Next step is to find the form of function $\varPhi_n(w,\tau)$. For this purpose, we substitute Expression (26) into Eq. (25) and divide by ε and write the equation under $\varepsilon \to 0$.

$$\begin{aligned}\frac{\partial \varPhi_n(w,\tau)}{\partial \tau} &= \frac{(jw)^2}{2}\varPhi_n(w,\tau) \times (a(x) + 2x(R_0(x) - g_0(x)) \\ -2\lambda_n g_0(x) &+ (2\phi_0(x)(x+\lambda_n) + 2R_0(x))\frac{\partial \varPhi_n(w,\tau)}{\partial w}\frac{1}{w\varPhi_n(w,\tau)}\bigg).\end{aligned} \tag{32}$$

So, we obtain the equation

$$\frac{\partial \varPhi_n(w,\tau)}{\partial \tau} = w\frac{\partial \varPhi_n(w,\tau)}{\partial w}a'(x) + \frac{(jw)^2}{2}\varPhi_n(w,\tau)b(x), \tag{33}$$

where

$$b(x) = a(x) + 2x(R_0(x) - g_0(x)) - 2\lambda_n g_0(x).$$

Let us denote the probability distribution density:

$$P(y,\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-jwy}\varPhi(w,\tau)dw.$$

Then we have the following Fokker-Planck equation:

$$\frac{\partial P(y,\tau)}{\partial \tau} = -\frac{\partial}{\partial y}\left(P(y,\tau)ya'(x)\right) + \frac{1}{2}\frac{\partial^2}{\partial y^2}(P(y,\tau)b(x))$$

for the probability distribution density of the diffusion process $y(\tau)$:

$$dy(\tau) = y(\tau)a^*(x)d\tau + \sqrt{b(x)}dw(\tau).$$

Combining the results of two asymptotics, we introduce

$$z(\tau) = x(\tau) + \varepsilon y(\tau),$$

which is diffusion random process satisfying the following Fokker-Planck equation

$$\frac{\partial P(z,\tau)}{\partial \tau} = -\frac{\partial}{\partial z}\left(P(z,\tau)a(z)\right) + \frac{1}{2}\frac{\partial^2}{\partial z^2}(P(z,\tau)\sigma b(z)).$$

In steady state, we obtain the following expression for probability distribution density of $z(\tau)$:

$$P(z) = \frac{C}{b(z)}\exp\left(\frac{\sigma}{2}\int_0^z \frac{a(x)}{b(x)}dx\right).$$

Returning to the notation, we substitute $z = \sigma_n i_n$:

$$P_n(i_n) = \frac{C}{b(\sigma_n i_n)}\exp\left(\frac{\sigma_n}{2}\int_0^{\sigma_n i_n} \frac{a(x)}{b(x)}dx\right) \tag{34}$$

where $C = \text{const}$ obtained from the normalization condition $\sum_i P(i) = 1$.

Thus, we obtain the formula for calculation of the asymptotic stationary distribution of the one-dimensional process of the number of customers on the n-th orbit.

Note, that for each class of customers, parameters $a(x)$ and $b(x)$, $g_k(x)$, etc. will be different.

4 Numerical Examples

For demonstrating the proposed asymptotic-diffusion method application area, we present the comparison of asymptotic $P_n(i)$ and empirical $D_n(i)$ (calculated by simulation) distributions for different values of the model parameters. Let us system parameters be the following

$$N = 2, \quad \lambda_1 = 0.3, \quad \lambda_2 = 0.5, \quad \mu_1 = \mu_2 = 1, \quad \sigma_1 = \sigma_2 = 0.1.$$

In Fig. 2, the comparison of the distributions is demonstrated.

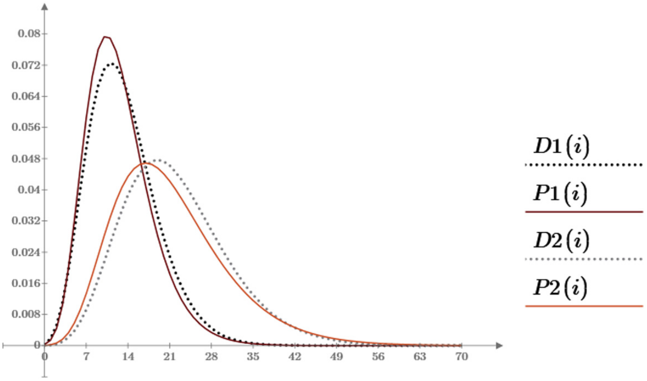


Fig. 2. Comparison of asymptotic and simulate probability distributions of numbers of the first and the second class of customers for $\sigma = 0.1$

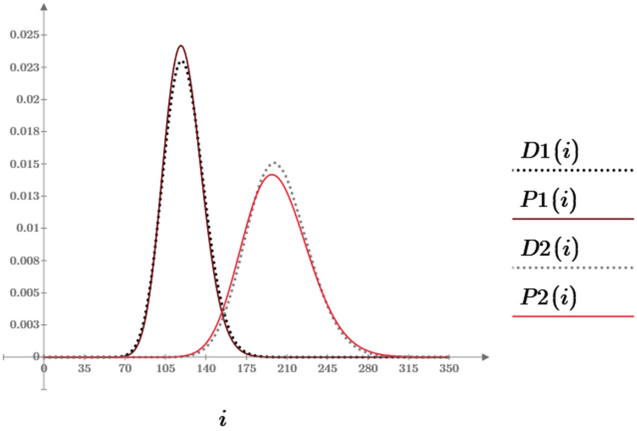


Fig. 3. Comparison of asymptotic and simulate probability distributions of numbers of the first and the second class of customers for $\sigma = 0.01$

In the next example (Fig. 3), we propose $\sigma_1 = \sigma_2 = 0.01$.

For the estimation of the asymptotic accuracy, we use

- Kolmogorov distance (Table 1).
- The relative error of asymptotic means (Table 2).

Table 1. The Kolmogorov distances for different σ

σ	1	0.1	0.01
$\delta_1(t)$	0.135	0.053	0.021
$\delta_2(t)$	0.155	0.051	0.026

Table 2. The relative error of asymptotic means for different σ

σ	1	0.1	0.01
m_1	0.150	0.032	0.004
m_2	0.240	0.047	0.005

Thus, we conclude that the method accuracy increases with decreasing of retrial rates.

Let us demonstrate one more example for four classes of customers with different retrial rates:

$$N = 4, \quad \lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 0.2, \quad \mu_1 = \mu_2 = \mu_3 = \mu_4 = 1,$$
$$\sigma_1 = 0.02, \quad \sigma_2 = 0.03, \quad \sigma_3 = 0.05, \quad \sigma_4 = 0.07.$$

In Fig. 4, the comparison of the distributions is demonstrated.

Also, we consider examples for four classes of customers with different arrival rates and the same other parameters:

$$N = 4, \quad \lambda_1 = 0.1, \quad \lambda_2 = 0.2, \quad \lambda_3 = 0.3, \quad \lambda_4 = 0.15,$$
$$\mu_1 = \mu_2 = \mu_3 = \mu_4 = 1, \quad \sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 0.01.$$

In Fig. 5, the comparison of the distributions is demonstrated.

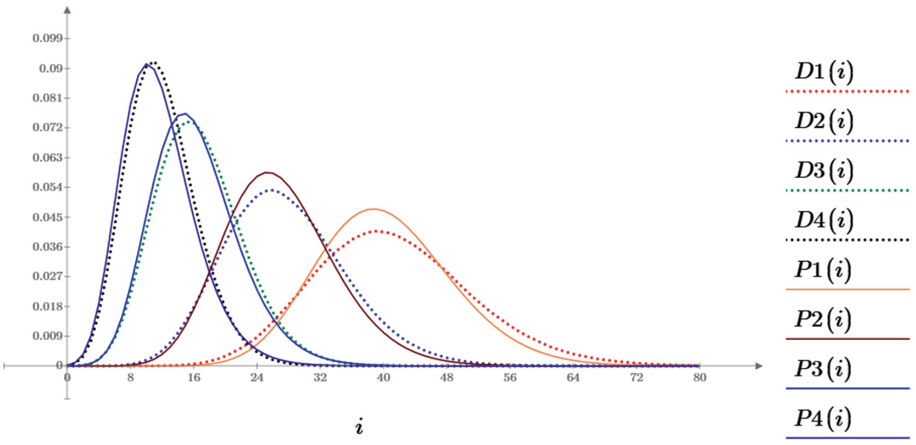


Fig. 4. Comparison of asymptotic and simulation distributions of the numbers of calls of the n-th class for $\sigma_1 = 0.02$, $\sigma_2 = 0.03$, $\sigma_3 = 0.05$, $\sigma_4 = 0.07$

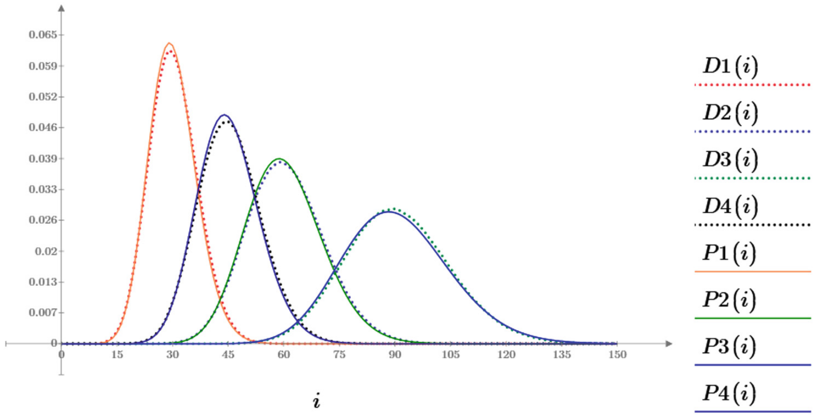


Fig. 5. Comparison of asymptotic and simulation distributions of the numbers of calls of the n-th class for $\lambda_1 = 0.1$, $\lambda_2 = 0.2$, $\lambda_3 = 0.3$, $\lambda_4 = 0.15$

We see that the proposed method has maximum accuracy when all retrial rates are small.

5 Conclusion

In the study, we have considered a multi-class retrial queue. The new method of marginal asymptotic analysis under the condition of a long delay has been proposed. The numerical analysis has been performed for different system parameters, which shown that the method accuracy increases with decreasing of retrial rates.

References

1. Artalejo, J.R., Gomez-Corral, A.: Retrial Queueing Systems. Springer, Berlin (2008). https://doi.org/10.1007/978-3-540-78725-9. 267 p.
2. Phung-Duc, T.: Retrial queueing models: a survey on theory and applications. In: Stochastic Operations Research in Business and Industry, pp. 1–26. World Scientific Publisher (2017)
3. Falin, G.I., Templeton, J.G.C.: Retrial Queues, p. 320. Chapman and Hall, London (1997)
4. Avrachenkov, K., Morozov, E., Nekrasova, R.: Optimal and equilibrium retrial rates in single-server multi-orbit retrial systems. In: Jonsson, M., Vinel, A., Bellalta, B., Tirkkonen, O. (eds.) MACOM 2015. LNCS, vol. 9305, pp. 135–146. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23440-3_11
5. Morozov, E., Rumyantsev, A., Dey, S., Deepak, T.G.: Performance analysis and stability of multiclass orbit queue with constant retrial rates and balking. Perform. Eval. **134**, 102005 (2019)
6. Krishnamoorthy, A., Joshua, V.C., Mathew, A.P.: A retrial queueing system with multiple hierarchial orbits and orbital search. In: Vishnevskiy, V.M., Kozyrev, D.V. (eds.) DCCN 2018. CCIS, vol. 919, pp. 224–233. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99447-5_19
7. Kim, B., Kim, J.: Proof of the conjecture on the stability of a multi-class retrial queue with constant retrial rates. Queueing Syst. **104**, 175–185 (2023)
8. Kim, B., Kim, J.: Stability of a multi-class multi-server retrial queueing system with service times depending on classes and servers. Queueing Syst. **94**, 129–146 (2020)
9. Avrachenkov, K.: Stability and partial instability of multi-class retrial queues. Queueing Syst. **100**(3–4), 177–179 (2022)
10. Shin, Y.W., Moon, D.H.: M/M/c retrial queue with multiclass of customers. Methodol. Comput. Appl. Probab. **16**, 931–949 (2014)
11. Falin, G.: On a multiclass batch arrival retrial queue. Adv. Appl. Probab. **20**(2), 483–487 (1988)
12. Nazarov, A., Phung-Duc, T., Paul, S., Lizura, O.: Asymptotic-diffusion analysis for retrial queue with batch poisson input and multiple types of outgoing calls. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) DCCN 2019. LNCS, vol. 11965, pp. 207–222. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-36614-8_16

Simulating Retrial Queues with Finite Source, Two-Way Communication to the Orbit, Backup Server, and Impatient Customers

Ádám Tóth[1](✉), János Sztrik[1], and Avtandil Bardavelidze[2]

[1] University of Debrecen, University Square 1, Debrecen 4032, Hungary
{toth.adam,sztrik.janos}@inf.unideb.hu
[2] Akaki Tsereteli State University, Kutaisi, Georgia
avtandil.bardavelidzec@atsu.edu.ge

Abstract. This paper explores a retrial queuing system with two-way communication and an unreliable server that may encounter random breakdowns. The system is of the finite-source M/M/1//N type, where the idle server can initiate calls to customers in the orbit, termed as secondary customers. Both primary and secondary customer service times are characterized by exponential distributions, with rates denoted as μ_1 and μ_2, respectively. The novelty of this study lies in its investigation of various failure time distributions and their impact on critical performance metrics, such as the mean response time of a random customer, while utilizing a backup server with impatient customers. The backup server can be likened to a primary server operating at a reduced rate during maintenance intervals. To ensure a valid comparison, a fitting process equalizes the mean and variance across all distributions. The outcomes are visually depicted through the utilization of our self-made simulation program.

Keywords: Finite-source queuing system · Retrial queues · Two-way communication · Sensitivity analysis · Simulation · Impatience

1 Introduction

Nowadays, the analysis of telecommunication systems and the creation of optimal designs for these schemes have become formidable endeavors due to the immense traffic and escalating number of users. Information exchange pervades every facet of contemporary life, underscoring the need to develop mathematical and simulation models for telecommunication systems or adapt existing ones to keep pace with these dynamic changes. Retrial queues stand out as potent and fitting tools for modeling real-world challenges that arise in telecommunications, networks, mobile networks, call centers, and similar systems. A plethora of literature, exemplified by works like [1,5,6,10], delves into the examination of various retrial queuing systems characterized by recurring calls.

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 316–328, 2024.
https://doi.org/10.1007/978-3-031-65385-8_23

We are currently exploring a retrial queuing system endowed with two-way communication capabilities, a research area that has gained substantial prominence owing to its striking resemblance to certain real-world systems. This correspondence is particularly pronounced in the context of call centers, where service units often perform multitasking, engaging in activities such as sales, promotions, and product advertising alongside handling incoming calls. In our investigation, the primary server, following a random idle interval, calls customers in from the orbit, called secondary customers. The system's utilization of the service unit is under scrutiny and has undergone extensive examination in prior works, exemplified by studies like [4,9,13].

In various research scenarios, some assume that service units remain continuously available, but real-world events like failures or unexpected incidents can occur during their operation, resulting in the rejection of incoming customers. Devices used across different industries are prone to breakdowns, and relying on their uninterrupted operation is often overly optimistic and unrealistic. Likewise, in wireless communication, multiple factors can affect transmission rates, resulting in disruptions during packet transmission. The inherent lack of reliability in retrial queuing systems has a substantial impact on system operations and performance metrics.

Furthermore, ceasing production entirely is not a feasible choice, as it may result in delays in order fulfillment. Therefore, in the event of such failures, it becomes crucial to keep machines or operators with lower processing rates operational to ensure a continuous workflow. Additionally, the authors investigated the option of implementing a backup server that could provide services at a reduced rate when the primary server is inaccessible. This approach has attracted significant attention in recent research, with studies such as [8,12] being notable examples.

In the service sector, it's not uncommon for service providers to encounter disruptions for various reasons, including difficulties in accessing their databases to address customer requests. When such disruptions transpire, service providers frequently employ alternative measures, such as resorting to backup systems or gathering additional information from customers to meet their needs.

Numerous research papers explore the performance of systems with the objective of improving service by integrating a backup server, as demonstrated in studies such as [2,11,15,16]. These inquiries provide insights into strategies and approaches for sustaining service quality in challenging scenarios.

The primary aim of this investigation is to assess how the system's unreliable operation affects performance measures, such as the mean response time of a customer or service unit utilization, by comparing various failure time distributions while the customers may depart after a random long enough waiting. This study builds upon the authors' earlier research [17], where the system incorporated an unreliable server. In the current configuration, in the event of server unavailability, a backup server takes over the processing of incoming requests.

To acquire the desired performance metrics, we developed a simulation model utilizing SimPack [7], which encompasses a collection of C/C++ libraries and executable programs tailored for computer simulation. Simulation serves as an excellent alternative for approximating performance metrics when deriving precise formulas proves problematic or nearly impossible. This paper introduces a sensitivity analysis of different failure time distributions' impact on key performance measures. We elucidate these findings by means of graphical representations that highlight intriguing facets of sensitivity-related issues.

2 System Model

The system is a retrial queuing system characterized by an unreliable server and a finite source of customers which is shown in Fig. 1. Within the source, there exist N customers, each generating primary requests at an exponential rate denoted by λ. Consequently, the inter-arrival times adhere to an exponential distribution parameterized by λ. Notably, our model does not contain waiting queues; thus, incoming customers can only occupy the server when it is available and idle. The service time for primary customers follows an exponential distribution with a parameter of μ_1. Following the successful completion of a service, the customer returns to the source. However, if an incoming customer (whether from the source or orbit) encounters a server in a busy or failed state, its request is redirected to the orbit. While within the orbit, a customer may attempt to fulfill its service requirement after an exponentially distributed random time with a parameter of σ.

The system assumes the presence of an unreliable server prone to failures, which can occur according to different distributions-such as gamma, hypoexponential, hyper-exponential, Pareto, and lognormal. Each distribution comes with distinct parameters while sharing the same mean value. The repair process initiates immediately upon the server's failure, with the repair time following an exponential distribution characterized by parameter γ_2. If the server is busy and subsequently fails, the customer is promptly transferred to the orbit. Regardless of the service unit's availability, all customers within the source can generate requests. However, these requests are directed to the backup server, which operates at a reduced rate-an exponentially distributed random variable with parameter μ_3-when the primary server is unavailable. Importantly, the backup server is assumed to be reliable and functions solely during periods of primary server unavailability. In cases where the backup server is busy, incoming requests are placed into the orbit. Yet, during idle periods, the main server can initiate outgoing calls to customers within the orbit after a random time interval, characterized by an exponential distribution with a rate of ν. The service time for these secondary customers follows an exponential distribution with parameters μ_2. Customers in the orbit, after waiting an exponentially distributed time with parameter τ, may choose to leave the system without getting their service.

Throughout the model's creation, the fundamental assumption is maintained that all random variables remain entirely independent of each other.

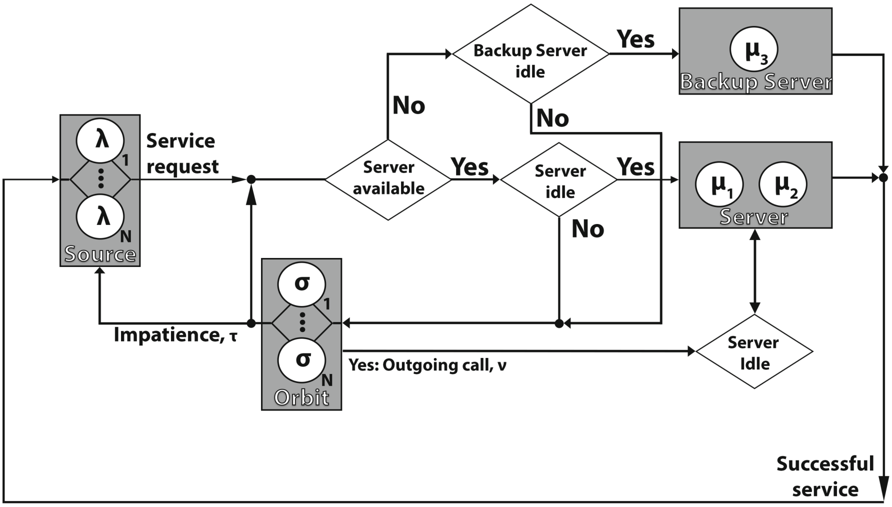


Fig. 1. System model

3 Simulation Results

We employed a statistical module class that incorporates a statistical analysis tool, enabling us to quantitatively estimate both the mean and variance values of observed variables via the batch mean method. This method involves aggregating n consecutive observations from a steady-state simulation to generate a sequence of independent samples. The batch mean method is a widely utilized technique for establishing confidence intervals concerning the steady-state mean of a process. It is important to note that, in order to ensure that the sample averages exhibit approximate independence, the use of sizable batches is imperative. Further details on the batch mean method can be found in [3,14]. In our simulations, we conducted operations with a confidence level of 99.9%, and the simulation run concluded when the relative half-width of the confidence interval reached the threshold of 0.00001.

3.1 First Scenario

In Table 1 the used values of input parameters are presented. The parameters of the failure time are presented in the following table (Table 2). To ensure a valid comparison, parameters are selected to have the same mean and variance values. The simulation program was executed with various parameter values, and this paper will highlight the most significant results. As indicated in the table, the squared coefficient of variation is greater than one in this scenario, enabling the examination of the impact of specific random variables. Additionally, we present results with a different set of parameters when the squared coefficient of variation is less than one.

Table 1. Used numerical values of model parameters

N	λ	γ_2	σ	μ_1	μ_2	ν	μ_3	τ
100	0.01	1	0.01	1	1.2	0.02	0.1	0.001

Table 2. Parameters of failure time

Distribution	Gamma	Hyper-exponential	Pareto	Lognormal
Parameters	$\alpha = 0.6$ $\beta = 0.5$	$p = 0.25$ $\lambda_1 = 0.41667$ $\lambda_2 = 1.25$	$\alpha = 2.2649$ $k = 0.67018$	$m = -0.3081$ $\sigma = 0.99037$
Mean	1.2			
Variance	2.4			
Squared coefficient of variation	1.6666666667			

The steady-state distribution, corresponding to different failure time distributions, is visually represented in Fig. 2. In this graph, the X-axis is labeled as i, which denotes the number of customers present in the system, while the Y-axis is labeled as $P(i)$, indicating the probability of precisely i customers being in the system. A closer examination of the curves reveals that all of them closely resemble the normal distribution. Notably, the Pareto distribution seems to exhibit a lower number of customers in the system. Nevertheless, when comparing the different distributions examined in our study, no significant disparities emerge.

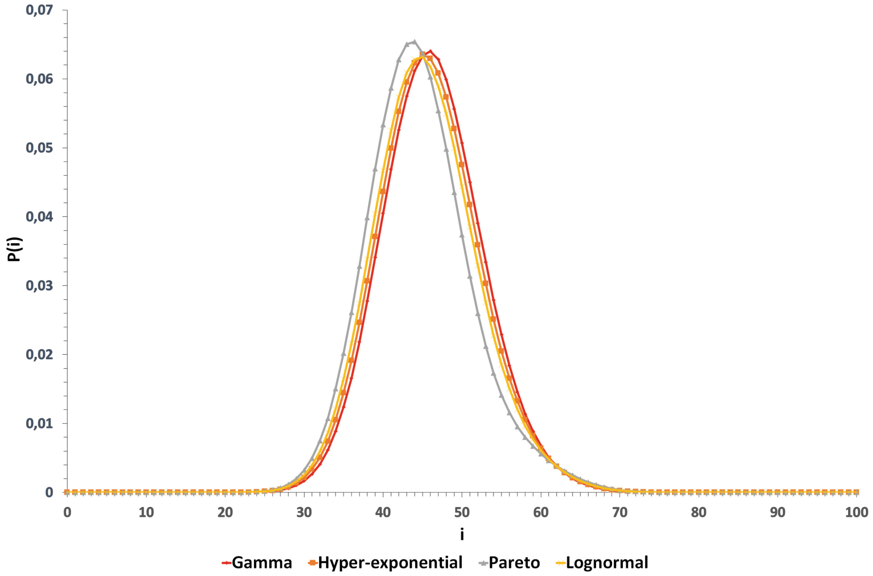


Fig. 2. Comparison of steady-state distributions

Figure 3 provides an illustration of the correlation between the mean response time of customers and the arrival intensity. On the contrary to the patterns observed in Fig. 2, the highest mean response time is associated with the Pareto distribution. However, the distinctions among the other distribution types become more pronounced. Remarkably, the gamma distribution stands out by yielding the lowest mean response time. A noteworthy phenomenon is that, as the arrival intensity increases, the mean response time initially experiences an uptrend, but subsequently, it starts to decrease after reaching a specific threshold. This behavior is a distinctive characteristic of retrial queuing systems with a finite source, and it tends to manifest when appropriate parameter configurations are applied.

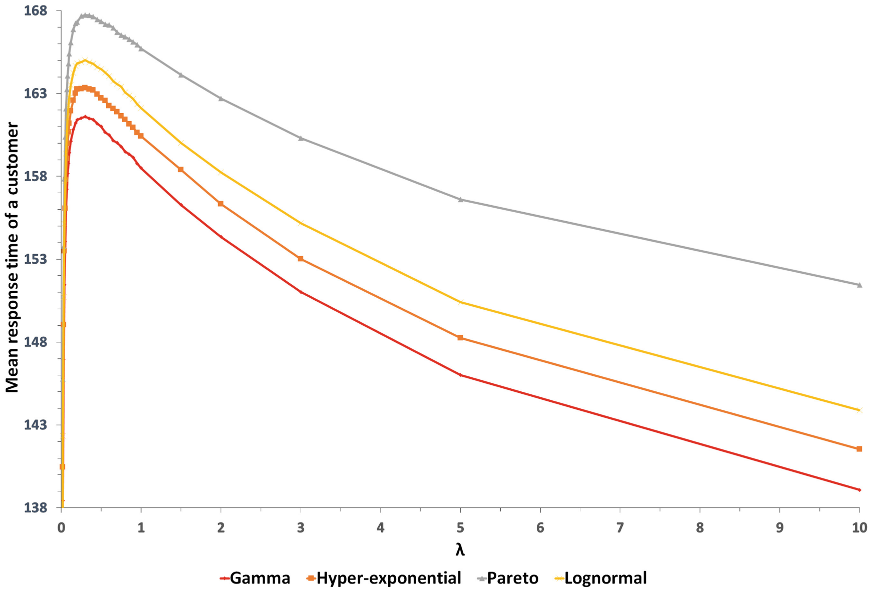


Fig. 3. Mean response time vs. arrival intensity

The variance of the response time is presented in the function of the arrival intensity of the incoming customers in Fig. 4. Looking at the results, it can be said that there are differences in this indicator as well, considering the used failure distributions. Similar trends to the previous chart are observed, with the smallest values occurring in the gamma distribution and the largest values in the Pareto distribution. However, at higher arrival intensity values, we find the smallest numbers in the Pareto distribution, which is an interesting development and requires further experiments and runs to explain this change.

The utilization of the backup service unit is shown in Fig. 5 besides the arrival of the incoming primary customers. In choosing the parameters, we aimed to simulate an environment where as many interruptions or failures as possible

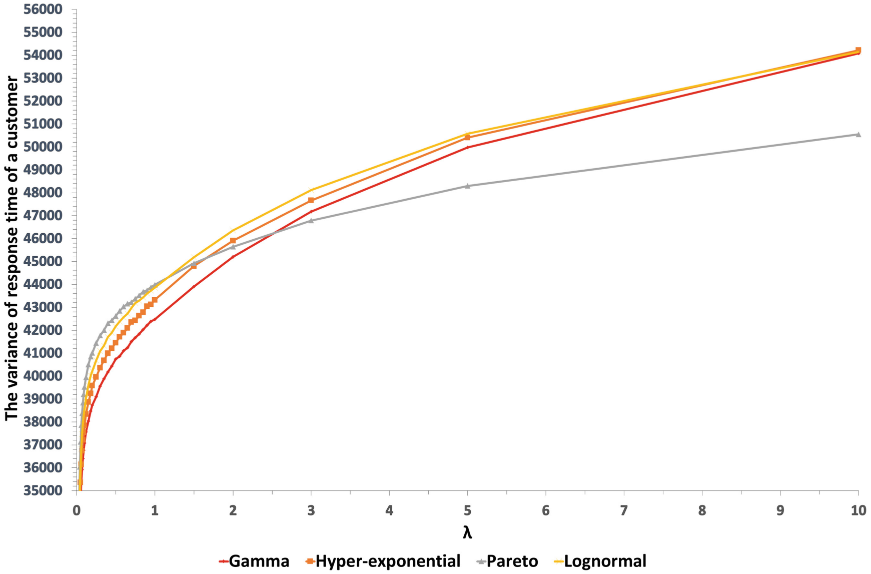


Fig. 4. Variance of the response time vs. arrival intensity

occur. Thus, in the chart, the utilization of the backup server unit becomes significant, and it is evident that this server is busy for most of the time. There are no significant differences among the used failure distributions, but with the Pareto distribution, the utilization is higher compared to the other distributions.

Figure 6 demonstrates the development of the probability of abandonment of a primary customer besides increasing arrival intensity. This metric indicates the likelihood of any given primary customer exiting the system during the orbit, signifying that the request does not meet its specified service requirement (impatient customers). As λ increases, the value of this performance measure also starts to increase, and this holds true for every utilized distribution, but the discrepancy among them is relatively significant. In the case of the gamma distribution, the inclination to exit the system earlier is much lower than in the others, especially when compared to the Pareto distribution.

3.2 Second Scenario

Upon analyzing the outcomes from the previous section, our keen interest was focused on understanding how modifications to the failure time parameters would impact the performance measures. In this scenario, the parameters were selected to ensure that the squared coefficient of variation remains below one. Instead of employing a hyper-exponential distribution, we opt for a hypo-exponential distribution. This choice is motivated by the fact that, in the case of a hypo-exponential distribution, the squared coefficient of variation is always less than one. The identical performance measures will be visually presented as earlier,

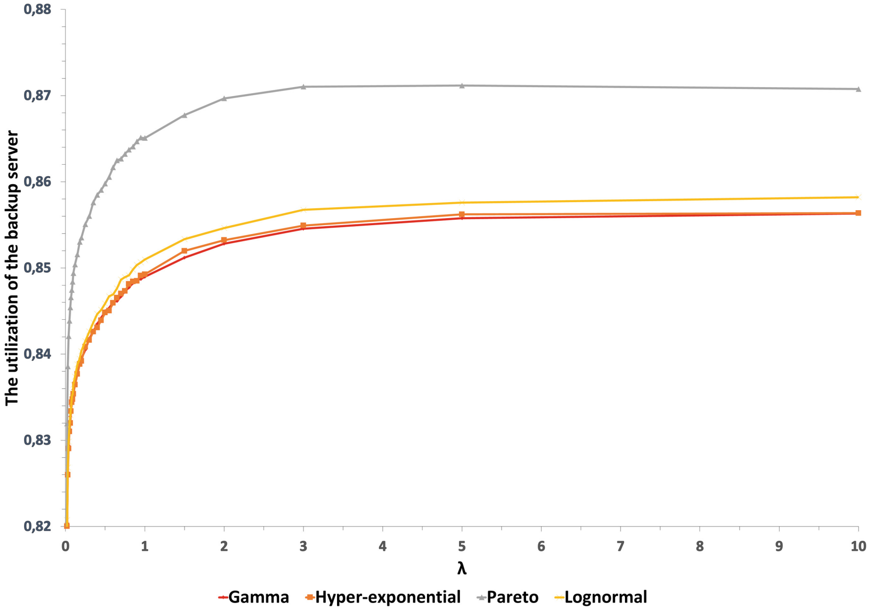


Fig. 5. Utilization of the backup server vs. arrival intensity

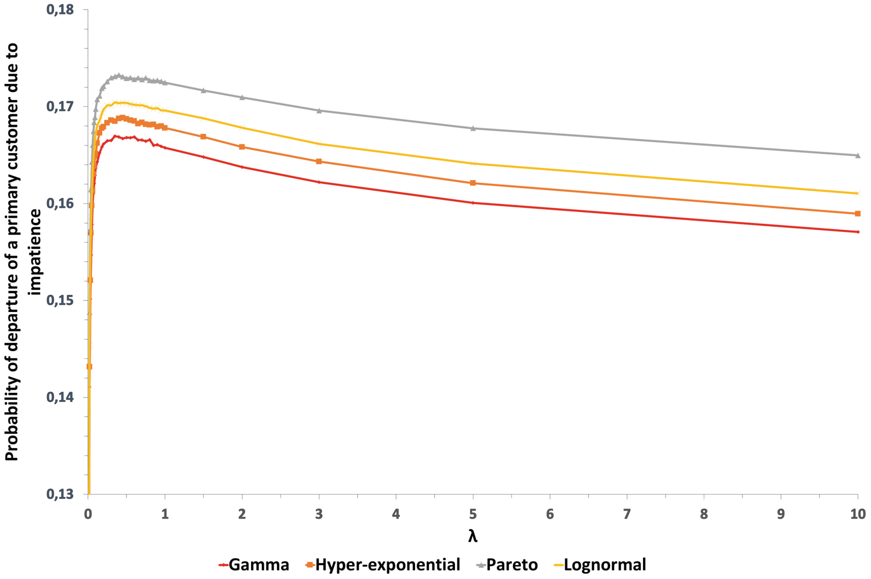


Fig. 6. The probability of the departure of a primary customer vs. arrival intensity

but with the incorporation of the new failure time parameters, as indicated in Table 2. The remaining parameters remain unchanged, as depicted in Table 1 (Table 3).

Table 3. Parameters of failure time

Distribution	Gamma	Hypo-exponential	Pareto	Lognormal
Parameters	$\alpha = 1.3846$ $\beta = 1.1538$	$\mu_1 = 1$ $\mu_2 = 5$	$\alpha = 2.5442$ $k = 0.7283$	$m = -0.08948$ $\sigma = 0.7373$
Mean	1.2			
Variance	1.04			
Squared coefficient of variation	0.72222222			

We will examine the same figures but with the updated parameter setting. Initially, Fig. 7 is related to the distribution of the number of customers in the system. Upon closer analysis of the curves, the obtained values are much more similar. Concerning the shape of the curves, they align with a normal distribution. Nevertheless, there isn't much disparity observed. As evident, the curves are nearly identical. The mean number of customers is slightly higher compared to the previous scenario.

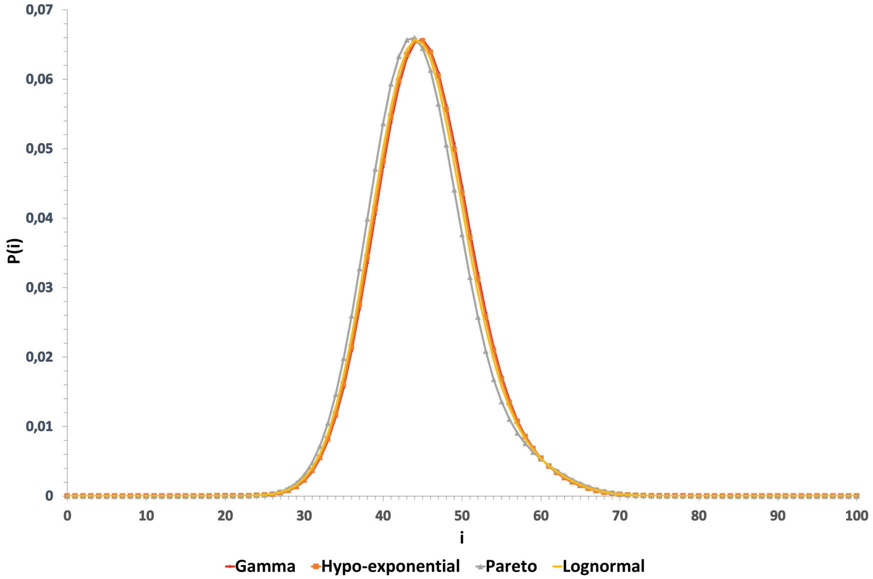


Fig. 7. Comparison of steady-state distributions

Figure 8 illustrates the evolution of the mean response time for a successfully served customer as the arrival intensity increases. In this situation, the mean value remains constant, but the variance is substantially reduced. The difference in the average mean response time among the distributions is not very pronounced, except for Pareto, where the values are notably higher. Therefore, it appears that variance has a noteworthy impact on performance measures, with larger values potentially leading to greater disparities in performance measures.

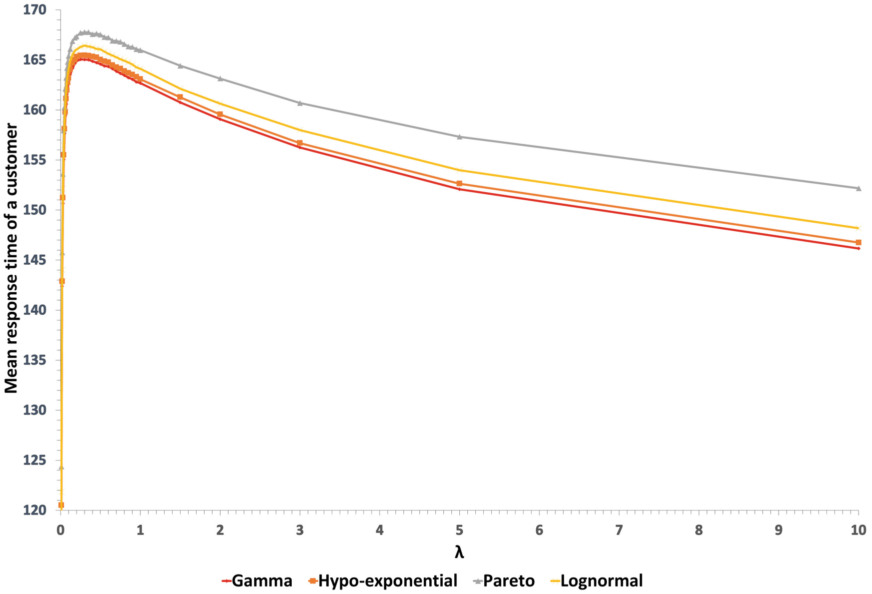


Fig. 8. Mean response time vs. arrival intensity

In the next, in Fig. 9 the variance of the response time is presented with the increment of the arrival intensity of the incoming customers. In comparison to the previous scenario, perhaps the difference is most evident in this figure with the use of newly employed parameters. In practice, the lines overlap completely, with prominent values only occurring in the log-normal distribution for higher arrival intensity values. What may be worth mentioning is that the values obtained in this scenario are smaller compared to the previous one.

Figure 10 depicts the comparison of the utilization of the backup server as a function of the arrival intensity. As expected, considering the results from the previous scenario, the differences in the obtained values are relatively close to each other, even in the case of Pareto distribution. It can be concluded that with this parameter setting, the distinctions among the distributions are not prominent. Regardless of the distribution, the utilization is nearly the same, meaning that the backup server is occupied approximately 87% of the simulation time.

Finally, Fig. 11 illustrates the variations in the abandonment probability with the increase in arrival intensity. The values are more closely aligned compared to

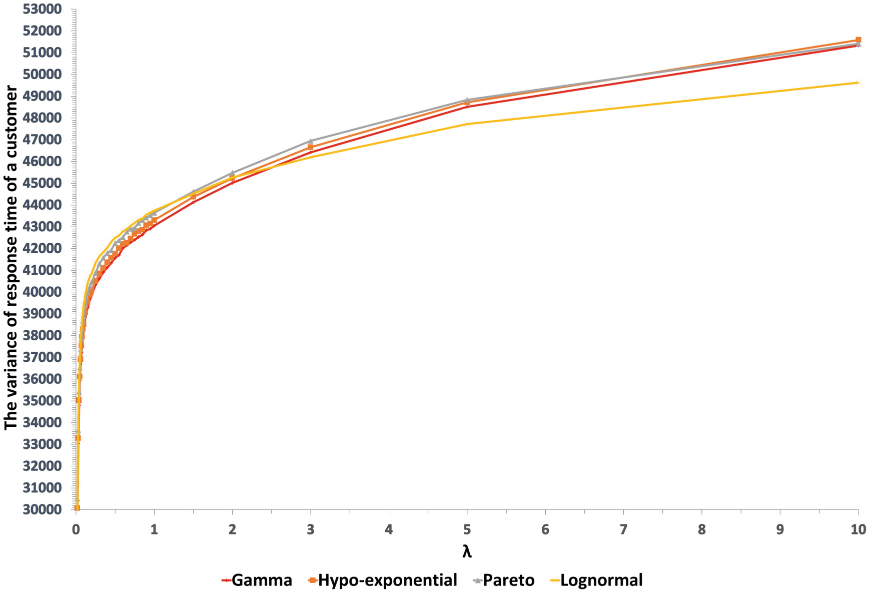


Fig. 9. Variance of the response time vs. arrival intensity

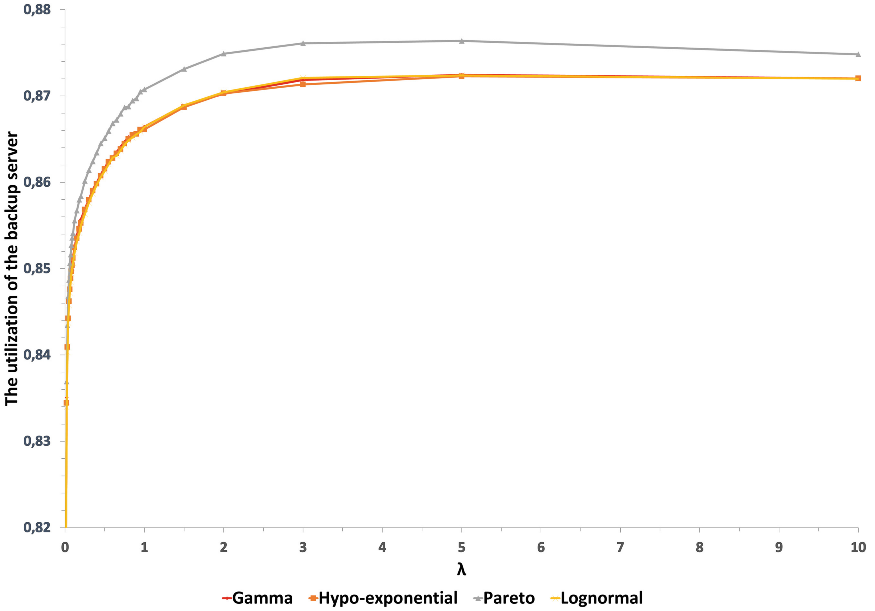


Fig. 10. Utilization of the backup server vs. arrival intensity

the alternative scenario. However, the highest values are observed in the case of the Pareto distribution; otherwise, the difference is minimal. The values obtained for one distribution do not stand out compared to the others; in each case, approximately 17% of incoming requests decide to abandon the system without being served.

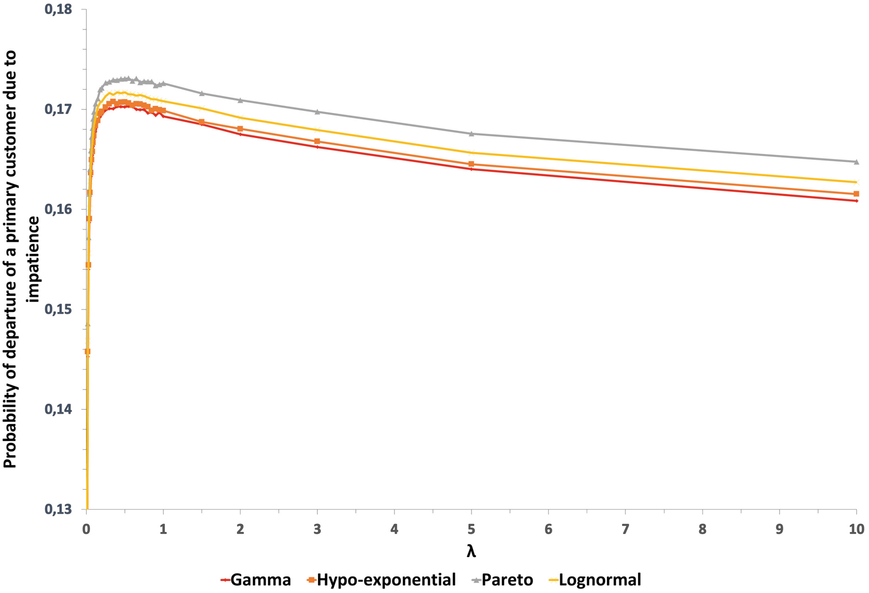


Fig. 11. The probability of the departure of a primary customer vs. arrival intensity

4 Conclusion

We introduced a retrial queuing system characterized by a finite source and two-way communication with impatient customers. Within this system, a primary server exhibits unreliability, and during periods of malfunction, a secondary service unit takes over. Furthermore, we conducted a sensitivity analysis utilizing a range of random number generators to investigate how different distributions of failure time impact performance metrics, such as the mean response time of any given customer. It's worth noting that when the squared coefficient of variation exceeds one, we observed variations in the mean response time among the values. Results also suggest that there is minimal difference among the measured values when the squared coefficient of variation is below one. The authors intend to further their research, delving into the observed phenomenon with greater scrutiny and enhancing their model by incorporating additional elements such as collisions and conducting additional sensitivity analyses on various random variables.

References

1. Artalejo, J., Gomez-Corral, A.: Retrial Queueing Systems: A Computational Approach. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78725-9
2. Chakravarthy, S.R., Shruti, Kulshrestha, R.: A queueing model with server breakdowns, repairs, vacations, and backup server. Oper. Res. Pers. **7**, 100131 (2020). https://doi.org/10.1016/j.orp.2019.100131. https://www.sciencedirect.com/science/article/pii/S2214716019302076
3. Chen, E.J., Kelton, W.D.: A procedure for generating batch-means confidence intervals for simulation: checking independence and normality. SIMULATION **83**(10), 683–694 (2007)
4. 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
5. Dragieva, V.I.: Number of retrials in a finite source retrial queue with unreliable server. Asia-Pac. J. Oper. Res. **31**(2), 23 (2014). https://doi.org/10.1142/S0217595914400053
6. Fiems, D., Phung-Duc, T.: Light-traffic analysis of random access systems without collisions. Ann. Oper. Res. **277**, 1–17 (2017)
7. Fishwick, P.A.: SimPack: getting started with simulation programming in C and C++. In: 1992 Winter Simulation Conference, pp. 154–162 (1992)
8. 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)
9. Gómez-Corral, A., Phung-Duc, T.: Retrial queues and related models. Ann. Oper. Res. **247**(1), 1–2 (2016). https://doi.org/10.1007/s10479-016-2305-2
10. Kim, J., Kim, B.: A survey of retrial queueing systems. Ann. Oper. Res. **247**(1), 3–36 (2016). https://doi.org/10.1007/s10479-015-2038-7
11. Klimenok, V., Dudin, A., Semenova, O.: Unreliable retrial queueing system with a backup server, pp. 308–322 (2021). https://doi.org/10.1007/978-3-030-92507-9_25
12. Krishnamoorthy, A., Pramod, P.K., Chakravarthy, S.R.: Queues with interruptions: a survey. TOP **22**(1), 290–320 (2014). https://doi.org/10.1007/s11750-012-0256-6
13. Kuki, A., Sztrik, J., Tóth, Á., Bérczes, T.: A contribution to modeling two-way communication with retrial queueing systems. In: Dudin, A., Nazarov, A., Moiseev, A. (eds.) ITMM/WRQ -2018. CCIS, vol. 912, pp. 236–247. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-97595-5_19
14. Law, A.M., Kelton, W.D.: Simulation Modeling and Analysis. McGraw-Hill Education, New York (1991)
15. Liu, Y., Zhong, Q., Chang, L., Xia, Z., He, D., Cheng, C.: A secure data backup scheme using multi-factor authentication. IET Inf. Secur. **11**(5), 250–255 (2017). https://doi.org/10.1049/iet-ifs.2016.0103
16. Satheesh R, K., Praba S, K.: A multi-server with backup system employs decision strategies to enhance its service. Research Square, pp. 1–31 (2023). https://doi.org/10.21203/rs.3.rs-2498761/v1
17. Sztrik, J., Tóth, Á., Pintér, Á., Bács, Z.: The effect of operation time of the server on the performance of finite-source retrial queues with two-way communications to the orbit. J. Math. Sci. **267**, 196–204 (2022). https://doi.org/10.1007/s10958-022-06124-z

Author Index

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 329–330, 2024.
https://doi.org/10.1007/978-3-031-65385-8