Alexander Dudin
Valentina Klimenok
Gennadiy Tsarenkov
Sergey Dudin (Eds.)

Communications in Computer and Information Science 356

Modern Probabilistic Methods for Analysis of Telecommunication Networks

Belarusian Winter Workshops
in Queueing Theory, BWWQT 2013
Minsk, Belarus, January 2013
Proceedings

Springer

Communications in Computer and Information Science 356

Alexander Dudin Valentina Klimenok
Gennadiy Tsarenkov Sergey Dudin (Eds.)

Modern Probabilistic Methods for Analysis of Telecommunication Networks

Belarusian Winter Workshops
in Queueing Theory, BWWQT 2013
Minsk, Belarus, January 28-31, 2013
Proceedings

Springer

Volume Editors

Alexander Dudin
Valentina Klimenok
Gennadiy Tsarenkov
Sergey Dudin

Belarusian State University
Laboratory of Applied Probabilistic Analysis
Minsk, Belarus

E-mail:
dudin@bsu.by
klimenok@bsu.by
gtsarenkov@tut.by
dudin85@mail.ru

ISSN 1865-0929 e-ISSN 1865-0937
ISBN 978-3-642-35979-8 e-ISBN 978-3-642-35980-4
DOI 10.1007/978-3-642-35980-4
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: Applied for

CR Subject Classification (1998): G.3, C.2.0, C.2.4, C.4

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The Belarusian Workshops on Queueing Theory were started in 1985. Their initiation as a continuation of the series of All-Union conferences in queueing theory in the former Soviet Union was triggered by the famous scientist B.V. Gnedenko and the founder of the Belarusian scientific school in probability theory G.A. Medvedev. These workshops were organized annually until 1999 and biennially since 2001 as scientific conferences on queueing theory and its various applications. The workshops became the main forum of researchers in queueing theory in the former Soviet Union and independent countries previously united in the Soviet Union. Since 1995, the representatives of many other countries (Austria, Algeria, Belgium, Bulgaria, Canada, China, France, Germany, Hungary, India, Italy, Japan, Korea, Mexico, The Netherlands, Portugal, Poland, Spain, Sweden, Turkey, USA) have participated in these workshops.

The Belarusian workshops on queueing theory achieved the status of international conferences, with each conference having its own subtitle. The 22nd Belarusian Workshop on Queueing Theory (BWWQT 2013) was held at the Belarusian State University, Minsk, Belarus, during January 28–31, 2013, as the international conference "Modern Probabilistic Methods for Analysis, Design and Optimization of Information Telecommunication Networks."

The proceedings of the Belarusian Workshops on Queueing Theory were regularly published by the Belarusian University Publishers as volumes in the series "*Queues: Flows, Systems, Networks.*" This year, a collection of selected papers among those accepted to the program of the workshop are published in Springer's *Communications in Computer and Information Science* (CCIS) series.

This volume presents new results in the study and optimization of information transmission models in telecommunication networks using different approaches, mainly based on theories of queueing systems and queueing networks.

This volume is aimed at specialists in probabilistic theory, random processes, mathematical modeling, and mathematical statistics as well as engineers engaged in logical and technical design and operational management of telecommunication and computer networks, databases, contact centers, health care, security, custom, border control, etc.

January 2013 Alexander Dudin

Organization

The workshops have been organized by the Department of Probability Theory and Mathematical Statistics and the Research Laboratory of Applied Probabilistic Analysis of the Belarusian State University since 1985. The Belarusian workshops are traditionally conducted as scientific conferences.

In 2013 the workshop was conducted as the International Conference "Modern Probabilistic Methods for Analysis, Design, and Optimization of Information and Telecommunication Networks."

Program Committee

A. Dudin (Belarus), Chair
G. Medvedev (Belarus)
M. Neuts (USA), Co-chairs
A. Andronov (Latvia)
V. Anisimov (UK)
D. Baum (Germany)
K. Al-Begain (UK)
S. Chakravarthy (USA)
B. Choi (Korea)
G. Falin (Russia, UK)
A. Gomez-Corral (Spain)
A. Gortsev (Russia)
C. Kim (Korea)
V. Klimenok (Belarus)
I. Kovalenko (Ukraine)
A. Krishnamurthy (India)
A. Latkov (Latvia)
E. Lebedev (Ukraine)
M. Lee (Korea)
A. Melikov (Azerbaijan)
Y. Malinkovsky (Belarus)
M. Matalytsky (Belarus)
E. Morozov (Russia)
A. Nazarov (Russia)
R. Nobel (The Netherlands)
V. Rykov (Russia)
S. Stepanov (Russia)
J. Sztrik (Hungury)
H. Tijms (The Netherlands)
O. Tikhonenko (Poland)
V. Vishnevsky (Russia)
A. Zeifman (Russia)

Local Organizing Committee

A. Dudin (Co-chair)
G. Medvedev
V. Klimenok
G. Tsarenkov
V. Mushko
A. Kazimirsky
O. Dudina
S. Dudin

Table of Contents

Analysis of Queueing System with Constant Service Time for SIP Server Hop-by-Hop Overload Control

Pavel Abaev[1], Alexander Pechinkin[2,1], and Rostislav Razumchik[2,1]

[1] Peoples Friendship University, Ordzhonikidze str., 3, 117198 Moscow, Russia
[2] Institute of Informatics Problems of Russian Academy of Sciences, Vavilova str., 44-2, 119333 Moscow, Russia
{pabaev,rrazumchik}@ieee.org, apechinkin@ipiran.ru

Abstract. Consideration is given to the analysis of queueing system $M_2|D|1|R$ with bi-level hysteretic input load control that can model signalling hop-by-hop overload control mechanism for SIP servers described in RFC 6357. Bi-level hysteretic input load control implies that system may be in three states (normal, overloaded, blocking), depending on the total number of customers present in it, and upon each state change input flow rate is adjusted. New approach that allows fast computation of joint stationary probability distribution is proposed, expressions for important performance characteristics are given.

Keywords: SIP server, hop-by-hop mechanism, loss-based overload control, constant service rate, queueing model.

1 Introduction

In [1] there was developed a hop-by-hop signaling load control mechanism based on the loss-based scheme for SIP server networks (described in [2]) and constructed the applicable threshold-based queuing system to analyze the performance characteristics of the mechanism. This study is devoted to the analysis of one of the generalizations of that queueing system, namely consideration is given to similar queueing system but in which customers that enter the queue are served for constant time as opposed to exponentially distributed service times as assumed in [1]. Threshold-based and hysteric queueing systems have been subject of extensive research, for example [3]–[7], just to mention a few. In this paper we propose new approach that allows fast computation of joint stationary probability distribution for significant practical use values of thresholds.

2 Description of the System

Consider the queueing system with Poisson incoming flows of customers (say type 1 and type 2) with rate λ_1 and λ_2, finite queue of size $R-1<\infty$, and one server. Denote $\lambda=\lambda_1+\lambda_2$. Type 1 customers have relative priority over

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 1–10, 2013.

customers of type 2 (i.e. no service interruptions are allowed and type 2 customer enters server only when it becomes free and there are no type 1 customers in the queue). If arriving customer sees R customers in the system, it is considered to be lost. All customers are served for constant time $0 < T < \infty$. The hysteric mechanism operates as follows. The system during operation changes its state depending on the total number of customers present in it. Choose arbitrary numbers L and H such that $0 < L < H < R$. When the system starts to work it is empty and as long as the total number of customers in the system remains below $H-1$, system is considered to be in "normal" state. When total number of customers exceeds $H-1$ for the first time, the system changes its state to "overload" and stays in it as long as the number of customers remains between L and $R-1$. Moreover when overloaded, system accepts only type 1 customers. Being in "overload" state, system waits till the number of customers drops down below L after which it changes its state back to "normal", or exceeds $R-1$ after which it changes its state to "blocking". In the "blocking" state systems does not accept new arriving customers until the total number of customers drops down below $H+1$, after which system's state changes back to "overload".

Consider random process $\{X(t) = (\xi(t), \eta(t), \nu(t)),\ \ t \geq 0\}$, where $\xi(t)$ is the total number of customers in the system at an instant t, $\eta(t)$ is the elapsed service time of the customer in server at instant t, $\nu(t)$ is the state of the system at instant t. When $\xi(t) = 0$, components $\eta(t)$ and $\nu(t)$ are omitted. Process $X(t)$ defined in such a way is Markov process. The state space of $X(t)$ can be represented as $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1 \cup \mathcal{X}_2$, where $\mathcal{X}_0$ is the set of "normal" states, $\mathcal{X}_1$ is the set of "overload" states, and $\mathcal{X}_2$ is the set of "blocking" states. These sets are

$$\begin{aligned}\mathcal{X}_0 &= \{0\} \cup \{(n, x, 0) : 0 < n \leq H-1, x \in [0, T]\}, \\ \mathcal{X}_1 &= \{(n, x, 1) : L \leq n \leq R-1, x \in [0, T]\}, \\ \mathcal{X}_2 &= \{(n, x, 2) : H+1 \leq n \leq R, x \in [0, T]\}\ .\end{aligned}$$

Let us introduce the following notation:

$$p_0(t) = \mathbf{P}\{\xi(t) = 0\}, \quad P_{ns}(x, t) = \mathbf{P}\{\xi(t) = n, \eta(t) < x, \nu(t) = s\},\ s = 0, 1, 2\ .$$

As $\mathcal{X}$ is finite and all states intercommunicate then limiting probabilities exist and coincide with stationary probabilities which we denote by $p_0 = \lim\limits_{t\to\infty} p_0(t)$, $P_{ns}(x) = \lim\limits_{t\to\infty} P_{ns}(x, t)$. It can be shown that for all allowable values of n and s derivatives $dP_{ns}(x)/dx$ exist and further they are denoted by $p_{ns}(x)$. In order to shed some light on the above made notations let us describe the meaning of $p_{ns}(x)$. Now then $p_{n0}(x)$, $\ n = \overline{1, H-1}$, is the stationary probability density of the fact, that total number of customers in the system is n, *elapsed* service time of currently served customer equals x and system accepts all (type 1 and type 2) arriving customers. Onwards, $p_{n1}(x)$, $\ n = \overline{L, R-1}$, is the stationary probability density of the fact, that total number of customers in the system is n, *elapsed* service time of currently served customer equals x and system accepts *only* type 1 arriving customers (type 2 customers are dropped). Finally, $p_{n2}(x)$, $n = \overline{H+1, R}$, is the stationary probability density of the fact, that total number

of customers in the system is n, *elapsed* service time of currently served customer equals x and system *does not accept any* new arriving customers.

In order to write out the equations for $p_{ns}(x)$ auxiliary functions are needed. We introduce them in the next section.

3 Auxiliary Functions

Assume that at arbitrary time instant τ the total number of customers in the system is n, $n = \overline{H+1, R-1}$, *remaining* service time of currently served customer is x, and only type 1 customers are allowed to enter system (i.e. the system is in "overload" state). Denote by $\alpha_n(x)$ the probability of the fact that until the moment of time when the total number of customers in the system equals $n-1$ for the first time, there will never be R customers in the system (or, alternatively, the total number of customer in the system will reach $n-1$ earlier that R).

Henceforth notation $\alpha_n = \alpha_n(T)$, $n = \overline{L, R-1}, n \neq H$ is used. By definition $\alpha_R(x) \equiv 0$, $\forall x \in [0, T]$. Let us show, that other probabilities $\alpha_n(x)$ satisfy the system of equations

$$\alpha_n'(x) = -\lambda_1 \alpha_n(x) + \lambda_1 \alpha_{n+1}(x)\alpha_n, \quad n = \overline{H+1, R-1} \ . \tag{1}$$

Indeed, consider time instant $\tau - \triangle$, where $\triangle$ is a small amount of time. Then, the total number of customer in the system will reach $n-1$ earlier that R given that at time instant $\tau - \triangle$ the total number of customers in the system is n, $n = \overline{H+1, R-1}$, *remaining* service time of currently served customer is x, and only type 1 customers are allowed to enter system if the following conditions are

1. in time $\triangle$ type 1 customer did not arrive (with probability $1 - \lambda_1 \triangle + o(\triangle)$) and the total number of customer in the system will reach $n-1$ earlier that R given at time instant τ the total number of customers in the system is n, *remaining* service time of currently served customer is $x - \triangle$, and only type 1 customers are allowed to enter system (with probability $\alpha_n(x - \triangle)$),
2. and in time $\triangle$ type 1 customer arrived (with probability $\lambda_1 \triangle + o(\triangle)$), the total number of customer in the system will reach n earlier that R given at time instant τ the total number of customers in the system is $n+1$, *remaining* service time of currently served customer is $x - \triangle$, and only type 1 customers are allowed to enter system (with probability $\alpha_{n+1}(x - \triangle)$) and total number of customers will reach $n-1$ earlier that R, given that remaining service time is T (with probability α_n).

Thus is holds

$$\alpha_n(x) = (1 - \lambda_1 \triangle)\alpha_n(x - \triangle) + \lambda_1 \triangle \alpha_{n+1}(x - \triangle)\alpha_n + o(\triangle), \quad n = \overline{H+1, R-1} \ .$$

Now, subtracting $\alpha_n(x)$ from both sides, dividing by $\triangle$ and taking the limit as $\triangle \to 0$ we obtain (1). Note that the boundary conditions for (1) are $\alpha_n(0) = 1$, $n = \overline{H+1, R-1}$.

The solution of (1) is straightforward. Let $\alpha_n(x) = e^{-\lambda_1 x}\beta_n(x), \quad n = \overline{H+1, R}$. Making such substitution into (1) we get

$$\beta_n'(x) = \lambda_1 \alpha_n \beta_{n+1}(x), \quad n = \overline{H+1, R-1} \ . \tag{2}$$

Using boundary conditions for (1) we find that $\beta_n(0) = 1, \quad n = \overline{H+1, R-1}$. Moreover from the fact, that $\alpha_R(x) \equiv 0, \ \forall x \in [0,T]$ it follows, that $\beta_R(x) \equiv 0, \ \forall x \in [0,T]$.

The solution of the system of differential equations (2) has the form

$$\beta_n(x) = \sum_{i=0}^{R-n-1} c_{n,i} x^i, n = \overline{H+1, R-1} \ . \tag{3}$$

Thus $\beta_{R-1}(x) = c_{R-1,0} = 1$ and from (2) for $n = \overline{H+1, R-2}$ we get

$$\beta_n(x) = c_{n,0} + \lambda_1 \alpha_n \int_0^x \beta_{n+1}(y)\, dy = c_{n,0} + \lambda_1 \alpha_n \sum_{i=0}^{R-n-2} \frac{x^{i+1}}{i+1} c_{n+1,i} \ . \tag{4}$$

From the fact, that $\beta_n(0) = 1, \ n = \overline{H+1, R-2}$ it follows, that $c_{n,0} = 1, \quad n = \overline{H+1, R-2}$. Remember, that $\forall x \in [0,T] \ \alpha_n(x) = e^{-\lambda_1 x}\beta_n(x), \quad n = \overline{H+1, R}$ and thereby $\beta_n(T) = e^{\lambda_1 T}\alpha_n(T)$. Now, by putting $x = T$ in (4), the expression for α_n is found:

$$\alpha_n = \left(e^{\lambda_1 T} - \lambda_1 \sum_{i=0}^{R-n-2} \frac{T^{i+1}}{i+1} c_{n+1,i} \right)^{-1}, \quad n = \overline{H+1, R-2} \ . \tag{5}$$

Comparing (3) and (4) we obtain the following recurrence relations for computation of coefficients $c_{n,i}$ in (3):

$$c_{n,0} = 1, \ n = \overline{H+1, R-1}, c_{n,i} = \frac{\lambda_1 \alpha_n}{i} c_{n+1,i-1}, \ i = \overline{1, R-n-1}, n = \overline{H+1, R-2}\,.$$

Thus the probability $\alpha_n(x) \ \forall x \in [0,T]$ can be determined by computing $\beta_n(x)$, using (3), and then multiplying it by $e^{-\lambda_1 x}$. Now we proceed to the definition of another auxiliary function.

Assume that at arbitrary time instant the total number of customers in the system is $n, \ n = \overline{L, H-1}$, *remaining* service time of currently served customer is x, and all arriving customers are allowed to enter system (i.e. the system is in "normal" state). Denote by $\alpha_n(x)$ the probability of the fact that until the moment of time when the total number of customers in the system equals $n-1$ for the first time, there will never be H customers in the system (alternatively, the total number of customer in the system will reach $n-1$ earlier that H).

By definition $\alpha_H(x) \equiv 0, \ \forall x \in [0,T]$. Using the same argument as above, it can be shown that probabilities $\alpha_n(x), \quad n = \overline{L, H-1}$ satisfy the system of equations

$$\alpha_n'(x) = -\lambda \alpha_n(x) + \lambda \alpha_{n+1}(x) \alpha_n \ , \tag{6}$$

with boundary condition $\alpha_n(0) = 1, \quad n = \overline{L, H-1}$. The solution of (6) is found by analogy with (1) and therefore is stated below without detailed explanation:

$$\alpha_n(x) = e^{-\lambda x} \sum_{i=0}^{H-n-1} c_{n,i} x^i, \quad n = \overline{L, H-1},$$

$$c_{n,0} = 1, \ n = \overline{L, H-1}, \ c_{n,i} = \frac{\lambda \alpha_n}{i} c_{n+1,i-1}, \ i = \overline{1, H-n-1}, \ n = \overline{L, H-2},$$

$$\alpha_n = \left(e^{\lambda T} - \lambda \sum_{i=0}^{H-n-2} \frac{T^{i+1}}{i+1} c_{n+1,i} \right)^{-1}, \quad n = \overline{L, H-2} \ .$$

Having introduced all necessary auxiliary functions, we proceed in the next section to the derivation and solution of equations for the stationary probability densities $p_{ns}(x)$.

4 Stationary Probability Distribution

Let us start with the derivation of equations for stationary probability densities $p_{n0}(x), \quad n = \overline{1, L-1}$. Let $n = \overline{2, L-1}$. Consider time instants $t - \triangle$ and t. For the process $X(t)$ to be in state $(n, x, 0)$ at time instant t, the following conditions must hold

1. at the instant $t-\triangle$ the process $X(t)$ is in state $(n, x-\triangle, 0)$, and no customer arrives in the time interval $\triangle$ (with probability $1 - \lambda_1 \triangle + o(\triangle)$);
2. at the instant $t-\triangle$ the process $X(t)$ is in state $(n-1, x-\triangle, 0)$, one customer (with probability $\lambda_1 \triangle + o(\triangle)$) arrives in the time interval $\triangle$.

Since all other events are of probability $o(\triangle)$, using the law of total probability, we obtain

$$p_{n0}(x,t) = p_{n0}(x-\triangle, t-\triangle)(1-\lambda\triangle) + p_{n-1,0}(x-\triangle, t-\triangle)\lambda\triangle + o(\triangle), \quad n = \overline{2, L-1} \ .$$

Taking the limit as $t \to \infty$ in the previous equation, it can be rewritten as

$$p_{n0}(x) = p_{n0}(x-\triangle)(1-\lambda\triangle) + p_{n-1,0}(x-\triangle)\lambda\triangle + o(\triangle), \quad n = \overline{2, L-1} \ .$$

Now, subtracting $p_{n0}(x-\triangle)$ from both sides, dividing by $\triangle$ and making $\triangle$ tend to zero, we obtain

$$p'_{n0}(x) = -\lambda p_{n0}(x) + \lambda p_{n-1,0}(x), \quad n = \overline{2, L-1} \ . \tag{7}$$

When $n = 1$, following the similar argument, one can show that it holds

$$p'_{10}(x) = -\lambda p_{10}(x) \ . \tag{8}$$

In order to solve (7) and (8) let us make the substitution $p_{n0}(x) = e^{-\lambda x} q_{n0}(x)$, $n = \overline{1, L-1}$. Then (7) and (8) will take the form

$$q'_{10}(x) = 0, \quad q'_{n0}(x) = \lambda q_{n-1,0}(x), \quad n = \overline{2, L-1} \ . \tag{9}$$

Solution of the system (9) are functions $q_n(x)$ such that

$$q_{10}(x) = c_1, \quad q_{n0}(x) = c_n + \lambda \int_0^x q_{n-1,0}(y)\, dy = \sum_{i=0}^{n-1} \frac{(\lambda x)^i}{i!} c_{n-i}, \quad n = \overline{2, L-1} \ .$$

In order to determine coefficients $c_n = p_{n0}(0), n = \overline{1, L-1}$ we will use the elimination method (see, for example, [8, Chapter 1]). Let us begin with probability $p_1(0)$. Notice that when the process $X(t)$ is in state $(1, x, 0), x \in [0, T]$ and any customer arrives, it leaves this state but with probability 1 comes back to it and moreover elapsed service time of the customer in server will always be 0 (in other words, the process $X(t)$ leaves the state $(1, x, 0)$ with arrival of a customer and with probability 1 comes back to state $(1, 0, 0)$). Thus, using the global balance principle, the probability of the state $(1, 0, 0)$ is

$$p_1(0) = c_1 = \lambda p_0 + \int_0^T \lambda p_1(x)\, dx = \lambda p_0 + c_1(1 - e^{-\lambda T}) \ ,$$

wherefrom it follows that $c_1 = \lambda e^{\lambda T} p_0$. If we consider probability $p_n(0), \quad n = \overline{2, L-1}$ and apply the same probabilistic argument, we obtain

$$p_n(0) = c_n = \int_0^T \lambda p_n(x)\, dx = c_n(1 - e^{-\lambda T}) + \lambda \int_0^T e^{-\lambda x} \sum_{i=1}^{n-1} \frac{(\lambda x)^i}{i!} c_{n-i} dx \ ,$$

whence

$$c_n = \lambda e^{\lambda T} \int_0^T e^{-\lambda x} \sum_{i=1}^{n-1} \frac{(\lambda x)^i}{i!} c_{n-i} dx, \quad n = \overline{2, L-1} \ .$$

Now, having found expressions for $p_{n0}(x), \ n = \overline{1, L-1}$, we proceed to derivation of equations for $p_{n0}(x), \quad n = \overline{L, H-1}$. Using completely the same reasoning which was used above for $p_{n0}(x), \ n = \overline{2, L-1}$, one can verify that it holds

$$p'_{n0}(x) = -\lambda p_{n0}(x) + \lambda p_{n-1,0}(x), \quad n = \overline{L, H-1} \ . \tag{10}$$

Substitution $p_{n0}(x) = e^{-\lambda x} q_{n0}(x), \ n = \overline{L, H-1}$ into (10) yields

$$q'_{n0}(x) = \lambda q_{n-1,0}(x), \quad n = \overline{L, H-1} \ .$$

Functions $q_{n0}(x)$ that satisfy the previous system of differential equations have the form

$$q_{n0}(x) = c_n + \lambda \int_0^x q_{n-1,0}(y)\, dy = \sum_{i=0}^{n-1} \frac{(\lambda x)^i}{i!} c_{n-i}, \quad n = \overline{L, H-1} \ . \tag{11}$$

Now one needs to determine coefficients $c_n = p_{n0}(0), n = \overline{L, H-1}$. Taking into consideration, that process $X(t)$ never visits state $(H-1, 0, 0)$, it holds that

$c_{H-1} = p_{H-1,0}(0) = 0$. Using elimination method and the same probabilistic argument as used above for obtaining equations for $c_n, n = \overline{1, L-1}$, we get the following equations for other coefficients $c_n, \quad n = \overline{L, H-2}$:

$$p_{n0}(0) = c_n = \int_0^T \lambda\alpha_{n+1}(T-x)p_{n0}(x)\,dx =$$

$$= \lambda c_n e^{-\lambda T} \sum_{j=0}^{H-n-2} \frac{T^{j+1}}{j+1} c_{n+1,j} + \lambda e^{-\lambda T} \sum_{j=0}^{H-n-2} \sum_{i=1}^{n-1} \frac{j!\,\lambda^i\,T^{i+j+1}}{(i+j+1)!} c_{n+1,j}c_{n-i} \ ,$$

whence

$$c_n = \lambda\alpha_n \sum_{j=0}^{H-n-2} \sum_{i=1}^{n-1} \frac{j!\,\lambda^i\,T^{i+j+1}}{(i+j+1)!} c_{n+1,j}c_{n-i}, \quad n = \overline{L, H-2} \ .$$

Having found expressions for c_n, $n = \overline{L, H-1}$, stationary probability densities $p_{n0}(x), \ n = \overline{L, H-1}$ are considered to be found too.

Let us dwell on the derivation of equations for $p_{n1}(x), \quad n = \overline{L, H-1}$. Considering the process $X(t)$ at time instants $t - \triangle$ and t one can verify that the following differential equations for $p_{n1}(x)$ hold

$$p'_{L1}(x) = -\lambda_1 p_{L1}(x) \ ,$$

$$p'_{n1}(x) = -\lambda_1 p_{n1}(x) + \lambda_1 p_{n-1,1}(x), \quad n = \overline{L+1, H-1} \ .$$

This system can be solved in the same manner as system (7)–(8). After substitution $p_{n1}(x) = e^{-\lambda_1 x} q_{n1}(x)$, $n = \overline{L, H-1}$ it yields to

$$q'_{L1}(x) = 0, \quad q'_{n1}(x) = \lambda_1 q_{n-1,1}(x), \quad n = \overline{L+1, H-1} \ . \tag{12}$$

Solution of the system (12) has the form

$$q_{n1}(x) = c_n^* + \lambda_1 \int_0^x q_{n-1,1}(y)\,dy = \sum_{i=0}^{n-L} \frac{(\lambda_1 x)^i}{i!} c^*_{n-i}, \quad n = \overline{L, H-1} \ . \tag{13}$$

Using again the elimination method to determine coefficients $c_n^* = p_{n1}(0), n = \overline{L, H-1}$ one can find, that

$$p_{n1}(0) = c_n^* = \int_0^T \lambda_1 p_{n1}(x)\,dx + \int_0^T \lambda p_{H-1,0}(x)\,dx \ ,$$

wherefrom after some algebraic manipulations it follows

$$c_n^* = \lambda_1 e^{\lambda_1 T} \int_0^T e^{-\lambda_1 x} \sum_{i=1}^{n-L} \frac{(\lambda_1 x)^i}{i!} c^*_{n-i} dx + \lambda e^{\lambda_1 T} p_{H-1,0}, \quad n = \overline{L, H-1} \ .$$

In the latter equality we assume $\sum_{i=1}^{0} \equiv 0$.

Whereas probabilities $p_{n1}(x)$, $n = \overline{L, H-1}$. are now considered to be found, we proceed to the determination of probability $p_{H1}(x)$. Considering the state change on the process $X(t)$ in the small time period $\triangle$ one can verify that it holds

$$p'_{H1}(x) = -\lambda_1 p_{H1}(x) + \lambda p_{H-1,0}(x) + \lambda_1 p_{H-1,1}(x) \ .$$

After substitution $p_{H1}(x) = e^{-\lambda_1 x} q_{H1}(x)$ the previous equation yields to

$$q'_{H1}(x) = \lambda e^{-\lambda_2 x} q_{H-1,0}(x) + \lambda_1 q_{H-1,1}(x) \ . \tag{14}$$

The solution of (14) has the form

$$q_{H1}(x) = c_H + \lambda \int_0^x e^{-\lambda_2 y} q_{H-1,0}(y)\, dy + \lambda_1 \int_0^x q_{H-1,1}(y)\, dy =$$

$$= c_H + \lambda \int_0^x e^{-\lambda_2 y} \sum_{i=0}^{H-2} \frac{(\lambda y)^i}{i!} c_{H-i-1} dy + \sum_{i=0}^{H-L-1} \frac{(\lambda_1 x)^{i+1}}{(i+1)!} c^*_{H-i-1} \ ,$$

where the constant $c_H = p_H(0)$ is found from equation $p_{H1}(0) = \int_0^T \lambda_1 p_{H1}(x)\, dx$ i. e. equals

$$c_H = \lambda e^{\lambda_1 T} \int_0^T e^{-\lambda y} \sum_{i=0}^{H-2} \frac{(\lambda y)^i}{i!} c_{H-i-1} dy - \lambda \int_0^T e^{-\lambda_2 y} \sum_{i=0}^{H-2} \frac{(\lambda y)^i}{i!} c_{H-i-1} dy +$$

$$+ \lambda_1 e^{\lambda_1 T} \int_0^T e^{-\lambda_1 x} \sum_{i=0}^{H-L-1} \frac{(\lambda_1 x)^{i+1}}{(i+1)!} c^*_{H-i-1} dx \ .$$

The equations for stationary probability densities $p_{n1}(x)$, $n = \overline{H+1, R-1}$, and their solution are almost identical to the ones for $p_{n0}(x)$, $n = \overline{L, H-1}$, that is it holds

$$p'_{n1}(x) = -\lambda_1 p_{n1}(x) + \lambda_1 p_{n-1,1}(x), \quad n = \overline{H+1, R-1} \ .$$

If we make substitution $p_{n1}(x) = e^{-\lambda_1 x} q_{n1}(x)$, $n = \overline{H+1, R-1}$, then the previous system of equations can be rewritten as

$$q'_{n1}(x) = \lambda_1 q_{n-1,1}(x), \quad n = \overline{H+1, R-1} \ ,$$

whose solution is

$$q_{n1}(x) = c_n + \lambda_1 \int_0^x q_{n-1,1}(y)\, dy, \quad n = \overline{H+1, R-1} \ .$$

In order to determine the functions $q_{n1}(x)$ completely one needs to find terms $c_n = p_{n1}(0)$, $n = \overline{H+1, R-1}$. As process $X(t)$ never visits state $(R-1, 0, 1)$, then $c_{R-1} = p_{R-1,1}(0) = 0$. Other terms c_n are found in the similar manner as it is done throughout the paper (i.e. using elimination method). Thereby for $n = \overline{H+1, R-2}$ it holds

$$c_n = p_{n1}(0) = \int_0^T \lambda_1 \alpha_{n+1}(T-x) p_{n1}(x)\, dx \ ,$$

wherefrom remembering that $\alpha_n(x) = e^{-\lambda_1 x}\beta_n(x)$ and $\beta_n(x)$, $n = \overline{H+1, R-2}$, is defined by (4)–(5), one finds that

$$c_n = \lambda_1 \int_0^T q_{n-1}(y)\,[\beta_n(T-y) - 1]dy, \quad n = \overline{H+1, R-2} \ .$$

The last system of equations that is left to be found is for $p_{n2}(x)$, $n = \overline{H+1, R}$. One can readily verify that these stationary probability densities satisfy

$$p'_{R2}(x) = \lambda_1 p_{R-1,1}(x), \quad p'_{n2}(x) = 0, \quad n = \overline{H+1, R-1} \ .$$

The solution of this system is $p_{n2}(x) = c^*_n$, $n = \overline{H+1, R-1}$ and

$$p_{R2}(x) = c^*_R + \lambda_1 \int_0^x p_{R-1,1}(y)\, dy = c^*_R + \lambda_1 P_{R-1,1}(x) \ .$$

The terms $c^*_n = p_{n2}(0)$, $n = \overline{H+1, R}$ can be found in the following way. Note that as process $X(t)$ never visits state $(R, 0, 2)$, then $c^*_R = p_{R2}(0) = 0$. Now as the system in "blocking" state does not accept any new arriving customers, then it holds

$$p_{n2}(0) = p_{n+1,2}(T), \quad n = \overline{H+1, R-1} \ .$$

Thus, we have shown how to determine all stationary probability densities $p_{ns}(x)$. The probability p_0 is found, as usual, from the normalization condition

$$p_0 = \left(1 + \sum_{n,s} \int_0^T p_{ns}(x)dx\right)^{-1} \ .$$

Using the above results, we may calculate performance characteristics of the system. Server utilization is simply $1 - p_0$. Loss probability of type 1 and type 2 customers, π_1 and π_2 respectively, is

$$\pi_1 = \sum_{n=H+1}^{R} \int_0^T p_{n2}(x)dx, \quad \pi_2 = \pi_1 + \sum_{n=L}^{R-1} \int_0^T p_{n1}(x)dx \ .$$

Load, served by the system, equals $\lambda^* = (1-\pi_1)\lambda_1 + (1-\pi_2)\lambda_2$ and thus mean waiting time is $V = Q/\lambda^*$, where Q – mean number of customers in the queue which can be computed as follows

$$Q = \sum_{n=1}^{H-1}(n-1)P_{n0} + \sum_{n=L}^{R-1}(n-1)P_{n1} + \sum_{n=H+1}^{R}(n-1)P_{n2}, \quad P_{ns} = \int_0^T p_{ns}(x)dx \ .$$

5 Conclusion

In this study consideration is given to finite $M_2|D|1|R$ queue with bi-level hysteretic load control which can serve as an alternative model for hop-by-hop overload control in SIP server networks. New approach is proposed which allows fast computation of the joint stationary distribution. Numerical experiments show that the proposed algorithm allows accurate computations of stationary distribution for high values of L and H ($L > H > 200$) in reasonable time. Our further research will be devoted to analysis of one of the important performance characteristic of hysteric mechanism in the considered queueing system – mean return time to normal operation state and verification of obtained results by comparing them with simulation based on real time traffic.

Acknowledgments. The reported study was partially supported by RFBR, research project No. 12-07-00108 and No. 11-07-00112.

References

1. Abaev, P., Gaidamaka, Y., Pechinkin, A., Razumchik, R., Shorgin, S.: Simulation of overload control in SIP server networks. In: Proc. of the 26th European Conference on Modelling and Simulation, pp. 533–539 (2012)
2. Hilt, V., Noel, E., Shen, C., Abdelal, A.: Design Considerations for Session Initiation Protocol (SIP) Overload Control. RFC6357 (2011)
3. Abaev, P., Gaidamaka, Y., Samouylov, K.E.: Queuing Model for Loss-Based Overload Control in a SIP Server Using a Hysteretic Technique. In: Andreev, S., Balandin, S., Koucheryavy, Y. (eds.) NEW2AN/ruSMART 2012. LNCS, vol. 7469, pp. 371–378. Springer, Heidelberg (2012)
4. Zhernoviy, K.Y., Zhernoviy, Y.V.: Queueing system $M^\theta|G|1|m$ with bi-level hysteric strategy of service rates switch-over. J. Information Processes. 12(2), 127–140 (2012) (in Russian)
5. Avrachenkov, K., Dudin, A., Klimenok, V., Nain, P., Semenova, O.: Optimal threshold control by the robots of web search engines with obsolescence of documents. J. Computer Networks 55(8), 1880–1893 (2011)
6. Semenova, O.: Optimal hysteresis control for BMAP/SM/1 queue with MAP-input of disasters. J. Quality Technology and Quantitative Management 4(3), 395–405 (2007)
7. Kim, C., Klimenok, V., Birukov, A., Dudin, A.: Optimal multi-threshold control by the BMAP/SM/1 retrial system. Annals of Operations Research 141(1), 193–210 (2006)
8. Bocharov, P., D'Apice, C., Pechinkin, A., Salerno, S.: Queueing Theory. VSP Publishing, Utrecht (2003)

On Mean Return Time in Queueing System with Constant Service Time and Bi-level Hysteric Policy

Pavel Abaev[1], Alexander Pechinkin[2,1], and Rostislav Razumchik[2,1]

[1] Peoples Friendship University, Ordzhonikidze str., 3, 117198 Moscow, Russia
[2] Institute of Informatics Problems of Russian Academy of Sciences, Vavilova str., 44-2, 119333 Moscow, Russia
{pabaev,rrazumchik}@ieee.org, apechinkin@ipiran.ru

Abstract. Single server queueing system with two Poisson input flows of rate λ_1 and λ_2, finite queue of size $R-1<\infty$ and bi-level hysteretic policy is considered. Customers of λ_1 flow are served with relative priority. Customers of both flows are served with the same constant service time. Bi-level hysteretic policy implies that system may be in three states (normal, overload, blocking), depending on the total number of customers present in it. New method for calculation of mean return time to normal operation state is proposed.

Keywords: SIP, hysteric control, constant service rate, queueing system, mean return time.

1 Introduction

Threshold load control is a well-known and reliable tool for preventing SS7 signalling link congestion [1]. In [2] it was shown that the same technique is applicable to overload control problems in a SIP server signalling network, that was stated in recent IETF RFCs and drafts (see, for example, [3]) and remains unsolved. Again in [2] Markov queuing system with bi-level hysteretic policy that can model overload control was introduced and thoroughly studied. In particular there was stated and numerically solved the problem of optimal choice of threshold values that minimize mean return time of the system to normal operation state given certain restrictions on blocking probabilities.

In this study we consider queueing system with bi-level hysteretic policy which in Kendall's notation is denoted by $M_2|D|1|R$. The main goal of the paper is to find expression for mean return time to normal operation state. The next section starts with the detailed description of the system and performance characteristic of interest, and introduces some auxiliary functions. In section 3 we propose new approach for the calculation of this performance characteristic. Conclusion sums up the results of the paper and outlines plans of further study.

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 11–19, 2013.

2 Description of the System

Consider the queueing system with two Poisson flows of rate λ_1 and λ_2, finite queue of size $R-1<\infty$, and one server. Further use notation $\lambda=\lambda_1+\lambda_2$. Customers of flow λ_1 have relative priority over customers of flow λ_2 (i.e. no service interruptions are allowed and customer of λ_2 flow enters server only when it becomes free and there are no customers of flow λ_1 in the queue). If arriving customer sees R customers in the system, it is considered to be lost. All customers are served for constant time $0<T<\infty$. The bi-level hysteric policy implies the following. The system during operation changes its state depending on the total number of customers present in it. Choose arbitrary numbers L and H such that $0<L<H<R$. When the system starts to work it is empty and as long as the total number of customers in the system remains below $H-1$, system is considered to be in "normal" state. When total number of customers exceeds $H-1$ for the first time, the system changes its state to "overloaded" and stays in it as long as the number of customers remains between L and $R-1$. Moreover when overloaded, system accepts only type 1 customers. Being in "overloaded" state, system waits till the number of customers drops down below L after which it changes its state back to "normal", or exceeds $R-1$ after which it changes its state to "blocking". In the "blocking" state systems does not accept new arriving customers until the total number of customers drops down below $H+1$, after which system's state changes back to "overloaded".

The important (as it is mentioned in [2]) performance characteristic of the system with bi-level hysteric policy is mean time it takes the system to return from "overloaded" or "blocking" state back to "normal" state. Assume at an arbitrary moment of time there are total of n, $n=\overline{L,R-1}$ customers in the considered queueing system and it is in "overloaded" state. Then denote by M_n, $n=\overline{L,R-1}$ – mean time to the time instant when the total number of customers in the system becomes equal $L-1$ for the first time. Now assume that at an arbitrary moment of time there are total of n, $n=\overline{H+1,R}$ customers in the considered queueing system and it is in "blocking" state. Denote by M_n^*, $n=\overline{H+1,R}$ – mean time to the time instant when the total number of customers in the system becomes equal $L-1$ for the first time. The goal is to obtain analytic expressions that allow fast computation of M_n and M_n^*.

In order to achieve this goal new approach was developed. Before moving to its explanation one needs to define the following auxiliary functions. Their purpose will become clear in Section 3.

Assume that at arbitrary time instant the total number of customers in the system is n, $n=\overline{H+1,R-1}$, *remaining* service time of currently served customer is x, and system is in "overload" state. Denote by $\alpha_n(x)$ the probability of the fact that until the moment of time when the total number of customers in the system equals $n-1$ for the first time, there will never be R customers in the system. Introduce notation $\alpha_n=\alpha_n(T)$, $n=\overline{L,R-1}$, $n\neq H$. As it is show in [4] for $\alpha_n(x)$ it holds

$$\alpha_n(x)=e^{-\lambda_1 x}\beta_n(x),\ n=\overline{H+1,R-1}\ , \tag{1}$$

where

$$\beta_n(x) = \sum_{i=0}^{R-n-1} c_{n,i}x^i,\ n = \overline{H+1, R-1}, \quad c_{n,0} = 1,\ n = \overline{H+1, R-1}, \quad (2)$$

$$c_{n,i} = \frac{\lambda_1 \alpha_n}{i} c_{n+1,i-1},\ i = \overline{1, R-n-1},\ n = \overline{H+1, R-2}, \quad (3)$$

$$\alpha_n = \left(e^{\lambda_1 T} - \lambda_1 \sum_{i=0}^{R-n-2} \frac{T^{i+1}}{i+1} c_{n+1,i}\right)^{-1}, \quad n = \overline{H+1, R-2}\ . \quad (4)$$

Now assume that at arbitrary time instant the total number of customers in the system is n, $n = \overline{L, H-1}$, *remaining* service time of currently served customer is x, and system is in "normal" state. Denote by $\alpha_n(x)$ the probability of the fact that until the moment of time when the total number of customers in the system equals $n-1$ for the first time, there will never be H customers in the system (alternatively, the total number of customer in the system will reach $n-1$ earlier that H. In [4] it was found that $\alpha_n(x)$, $n = \overline{L, H-1}$ have the form $\alpha_n(x) = e^{-\lambda x}\beta_n(x)$, $n = \overline{L, H-1}$, where

$$\beta_n(x) = \sum_{i=0}^{H-n-1} c_{n,i}x^i, \quad n = \overline{L, H-1}, \quad c_{n,0} = 1,\ n = \overline{L, H-1} \quad (5)$$

$$c_{n,i} = \frac{\lambda \alpha_n}{i} c_{n+1,i-1},\ i = \overline{1, H-n-1},\ n = \overline{L, H-2}, \quad (6)$$

$$\alpha_n = \left(e^{\lambda T} - \lambda \sum_{i=0}^{H-n-2} \frac{T^{i+1}}{i+1} c_{n+1,i}\right)^{-1}, \quad n = \overline{L, H-2}\ . \quad (7)$$

Having introduced auxiliary functions $\alpha_n(x)$, we proceed in the next section to the detailed explanation of the approach that allows calculation of mean return times M_n and M_n^*.

3 Calculation of Mean Return Times

Denote by $m_n(x)$, $n = \overline{H+1, R-1}$, – mean time to the time instant when the total number of customers becomes equal $n-1$ for the first time and until that time instant total number of customers will never reach R, provided that at *arbitrary* time instant (say τ) there are n customers in the system, *remaining* service time of currently served customer is x, and system is in "overloaded" state.

As opposed to $m_n(x)$ denote by $m_n^*(x)$, $n = \overline{H+1, R-1}$, – mean time to the time instant when the total number of customers becomes equal R for the first time and until that time instant total number of customers will never be less than n, provided that at *arbitrary* time instant (say τ) there are n customers in the system, *remaining* service time of currently served customer is x, and system is in "overloaded" state.

Henceforth, the following notation is used:

$$\alpha_n^*(x) = 1 - \alpha_n(x),\ n = \overline{H+1, R-1}, \quad \alpha_n^* = \alpha_n^*(T),\ n = \overline{H+1, R-1},$$
$$m_n = m_n(T),\ n = \overline{H+1, R-1}, \quad m_n^* = m_n^*(T),\ n = \overline{H+1, R-1}\ . \quad (8)$$

One can verify that for $m_n(x)$, $n = \overline{H+1, R-1}$ the following equations hold:

$$m'_{R-1}(x) = \alpha_{R-1}(x) - \lambda_1 m_{R-1}(x), \quad (9)$$
$$m'_n(x) = \alpha_n(x) - \lambda_1 m_n(x) +$$
$$+\lambda_1[m_{n+1}(x)\alpha_n + \alpha_{n+1}(x)m_n], \quad n = \overline{H+1, R-2}\ . \quad (10)$$

Indeed, let $n = \overline{H+1, R-2}$ and consider time instant $\tau - \triangle$, where $\triangle$ is a small amount of time. Then, the mean time to reach $n-1$ without visiting R, given that at time instant $\tau - \triangle$ total number of customer in the system is n, *remaining* service time of currently served customer is x and system is in "overloaded" state equals

1. $\triangle$ if eventually we will reach $n-1$ without visiting R, given that at time instant τ total number of customer in the system is n, *remaining* service time of currently served customer is $x - \triangle$ and system is in "overloaded" state (which happens with probability $\alpha_n(x)$);
2. mean time to reach $n-1$ without visiting R, given that at time instant τ total number of customer in the system is n, *remaining* service time of currently served customer is $x - \triangle$ and system is in "overloaded" (which happens with probability $1 - \lambda_1\triangle + o(\triangle)$);
3. sum of two terms (both with probability $\lambda_1\triangle + o(\triangle)$)
   - mean time to reach n without visiting R, given that at time instant τ total number of customer in the system is $n+1$, *remaining* service time of currently served customer is $x - \triangle$ and system is in "overloaded" state (which happens with probability $\alpha_n(T) = \alpha_n$) and
   - mean time to reach $n-1$ without visiting R, given that total number of customer in the system is n, *remaining* service time of currently served customer is T and system is in "overloaded" state if until the moment of time when the total number of customers in the system becomes equal n for the first time there will never be R customers, provided that at time instant τ the total number of customers in the system is $n+1$, *remaining* service time of currently served customer is $x - \triangle$ and system is in "overload" state (which happens with probability $\alpha_{n+1}(x)$).

Using the law of total expectation, we obtain

$$m_n(x) = \triangle\alpha_n(x - \triangle) + (1 - \lambda_1\triangle)m_n(x - \triangle) +$$
$$+\lambda_1\triangle[m_{n+1}(x - \triangle)\alpha_n + \alpha_{n+1}(x - \triangle)m_n] + o(\triangle), \quad n = \overline{H+1, R-2}\ .$$

Subtracting $m_n(x - \triangle)$ from both sides, dividing by $\triangle$ and taking the limit as $\triangle \to 0$ we obtain (10). For $n = R-1$ one can readily see, that

$$m_{R-1}(x) = \triangle\alpha_{R-1}(x - \triangle) + m_{R-1}(x - \triangle)(1 - \lambda_1\triangle) + o(\triangle)\ ,$$

so that by subtracting $m_{R-1}(x-\triangle)$ from both sides, dividing by $\triangle$ and making $\triangle$ tend to zero, we arrive at (9). Evidently the boundary conditions for (9) and (10) are $m_n(0)=0$, $n=\overline{H+1,R-1}$.

In order to solve (9)-(10) let us introduce functions $u_n(x)$, $n=\overline{H+1,R-1}$ such that

$$m_n(x)=e^{-\lambda_1 x}u_n(x), \quad n=\overline{H+1,R-1}\ . \tag{11}$$

Substitution of (11) into (9) and (10), seeing (1), yields

$$u'_{R-1}(x)=\beta_{R-1}(x), \tag{12}$$

$$u'_n(x)=\beta_n(x)+\lambda_1[u_{n+1}(x)\alpha_n+\beta_{n+1}(x)m_n], \quad n=\overline{H+1,R-2}\ . \tag{13}$$

Note, that functions $\beta_n(x)$ and values of α_n that enter the above system of equations are known and given by (2)–(4). The solution of (12)–(13) has the form

$$u_n(x)=\sum_{i=1}^{R-n} r_{n,i}x^i, \quad n=\overline{H+1,R-1}\ . \tag{14}$$

Coefficients $r_{n,i}$ can be found by substitution of (14) into (13) and direct integration. For $n=R-1$ integration from 0 to x of (14) and use of (2) leads to

$$u_{R-1}(x)=\int_0^x \beta_{R-1}(y)\,dy=c_{R-1,0}x=x\ ,$$

wherefrom it follows that $r_{R-1,1}=c_{R-1,0}=1$. For $n=\overline{H+1,R-2}$ if one integrates (13) from 0 to x, substitutes (14) into the result and uses (2), one obtains

$$u_n(x)=\int_0^x(\beta_n(y)+\lambda_1[u_{n+1}(y)\alpha_n+\beta_{n+1}(y)m_n])\,dy=$$

$$=\sum_{i=0}^{R-n-1} c_{n,i}\frac{x^{i+1}}{i+1}+\lambda_1\left[\alpha_n\sum_{i=1}^{R-n-1} r_{n+1,i}\frac{x^{i+1}}{i+1}+m_n\sum_{i=0}^{R-n-2} c_{n+1,i}\frac{x^{i+1}}{i+1}\right]\ . \tag{15}$$

By comparing coefficients of equal powers of x in (14) and (15) one can obtain the following formulas for $r_{n,i}$:

$$r_{n,1}=c_{n,0}+\lambda_1 m_n c_{n+1,0}, \quad n=\overline{H+1,R-2},$$

$$r_{n,i}=\frac{1}{i}(c_{n,i-1}+$$

$$+\lambda_1[\alpha_n r_{n+1,i-1}+m_n c_{n+1,i-1}]), \quad i=\overline{2,R-n-1}, \quad n=\overline{H+1,R-3},$$

$$r_{n,R-n}=\frac{1}{R-n}(c_{n,R-n-1}+\lambda_1\alpha_n r_{n+1,R-n-1}), \quad n=\overline{H+1,R-2}\ .$$

The lattermost terms that are left unknown in expressions for $r_{n,i}$ are m_n. They can be found, using relation (11). Indeed, if we substitute (15) into (11) and put

$x = T$, we arrive at the following relation for $n = \overline{H+1, R-2}$:

$$m_n = e^{-\lambda_1 T}\left[\sum_{i=0}^{R-n-1} c_{n,i}\frac{T^{i+1}}{i+1} + \lambda_1\alpha_n \sum_{i=1}^{R-n-1} r_{n+1,i}\frac{T^{i+1}}{i+1} + \lambda_1 m_n \sum_{i=0}^{R-n-2} c_{n+1,i}\frac{T^{i+1}}{i+1}\right] ,$$

whence, after collecting the common terms and seeing (4), it follows that

$$m_n = \alpha_n\left[\sum_{i=0}^{R-n-1} c_{n,i}\frac{T^{i+1}}{i+1} + \lambda_1\alpha_n \sum_{i=1}^{R-n-1} r_{n+1,i}\frac{T^{i+1}}{i+1}\right] .$$

Thus we have found all relations needed to compute coefficients $r_{n,i}$, functions $u_n(x)$ and ultimately quantities $m_n(x)$, $n = \overline{H+1, R-1}$.

Applying the same argument, which was used for $m_n(x)$, one can verify, that for $m_n^*(x)$, $n = \overline{H+1, R-1}$ it holds

$$m_{R-1}^{*\prime}(x) = \alpha_{R-1}^*(x) - \lambda_1 m_{R-1}^*(x), \tag{16}$$

$$m_n^{*\prime}(x) = \alpha_n^*(x) - \lambda_1 m_n^*(x) + \\ +\lambda_1[m_{n+1}^*(x) + m_{n+1}(x)\alpha_n^* + \alpha_{n+1}(x)m_n^*], \quad n = \overline{H+1, R-2} , \tag{17}$$

with boundary conditions $m_n^*(0) = 0$, $n = \overline{H+1, R-1}$. Substitution of

$$m_n^*(x) = e^{-\lambda_1 x}u_n^*(x), \quad n = \overline{H+1, R-1} , \tag{18}$$

into (16) and (17), seeing (1), yields

$$u_{R-1}^{*\prime}(x) = e^{\lambda_1 x} - \beta_{R-1}(x), \tag{19}$$

$$u_n^{*\prime}(x) = e^{\lambda_1 x} - \beta_n(x) + \\ +\lambda_1[u_{n+1}^*(x) + u_{n+1}(x)\alpha_n^* + \beta_{n+1}(x)m_n^*], \quad n = \overline{H+1, R-2} . \tag{20}$$

Note, that functions $\beta_n(x)$ and values of α_n^* that enter the above system of equations are known and given by (2)–(4) and (8). The solution of (19)–(20) has the form

$$u_n^*(x) = \sum_{i=0}^{R-n} r_{n,i}^* x^i + t_n e^{\lambda_1 x}, \quad n = \overline{H+1, R-1} . \tag{21}$$

For $n = R-1$ integration from 0 to x of (19) and use of (2) yields

$$u_{R-1}^*(x) = \int_0^x [e^{\lambda_1 x} - \beta_{R-1}(x)]dx = \frac{1}{\lambda_1}e^{\lambda_1 x} - \frac{1}{\lambda_1} - c_{R-1,0}x ,$$

whence it follows that $r_{R-1,0}^* = -1/\lambda_1$, $r_{R-1,1}^* = -c_{R-1,0} = 1$, $t_{R-1} = 1/\lambda_1$. Other coefficients $r_{n,i}^*$ and t_n can be found by substitution of (21) into (20), seeing (14) and direct integration. Thus it holds

$$u_n^*(x) = \int_0^x (e^{\lambda_1 y} - \beta_n(y) + \lambda_1[u_{n+1}^*(y) + u_{n+1}(y)\alpha_n^* + \beta_{n+1}(y)m_n^*])\, dy =$$

$$= \frac{(1+\lambda_1 t_{n+1})}{\lambda_1}(e^{\lambda_1 x}-1) - \sum_{i=0}^{R-n-1} c_{n,i}\frac{x^{i+1}}{i+1} + \lambda_1\Bigg[\sum_{i=0}^{R-n-1} r^*_{n+1,i}\frac{x^{i+1}}{i+1} +$$

$$+\alpha^*_n \sum_{i=1}^{R-n-1} r_{n+1,i}\frac{x^{i+1}}{i+1} + m^*_n \sum_{i=0}^{R-n-2} c_{n+1,i}\frac{x^{i+1}}{i+1}\Bigg], \quad n=\overline{H+1,R-2} \ .(22)$$

By comparing coefficients of equal powers of x in (22) and (21) one can obtain the following formulas for $r^*_{n,i}$ and t_n:

$$r^*_{n,0} = -\frac{1+\lambda_1 t_{n+1}}{\lambda_1}, \quad t_n = \frac{1+\lambda_1 t_{n+1}}{\lambda_1}, \quad n=\overline{H+1,R-2},$$

$$r^*_{n,1} = -c_{n,0} + \lambda_1[r^*_{n+1,0} + m^*_n c_{n+1,0}], \quad n=\overline{H+1,R-2},$$

$$r^*_{n,i} = -c_{n,i-1}\frac{1}{i} + \lambda_1\Bigg[\frac{r^*_{n+1,i-1}}{i} +$$

$$+\alpha^*_n\frac{r_{n+1,i-1}}{i} + m^*_n\frac{c_{n+1,i-1}}{i}\Bigg], \quad i=\overline{2,R-n-1}, \quad n=\overline{H+1,R-3},$$

$$r^*_{n,R-n} = -\frac{c_{n,R-n-1}}{R-n} + \lambda_1\left[\frac{r^*_{n+1,R-n-1}}{R-n} + \alpha^*_n\frac{r_{n+1,R-n-1}}{R-n}\right], \quad n=\overline{H+1,R-2} \ .$$

In expressions for $r^*_{n,i}$ and t_n the only unknown quantities are m^*_n, $n=\overline{H+1,R-2}$. They can be found, as well as case of m_n, using relation (18) when $x=T$. We omit these manipulations and state the final expression for m^*_n:

$$m^*_n = \alpha_n\Bigg(\frac{1+\lambda_1 t_{n+1}}{\lambda_1}(e^{\lambda_1 T}-1) - \sum_{i=0}^{R-n-1} c_{n,i}\frac{T^{i+1}}{i+1} +$$

$$+\lambda_1\Bigg[\sum_{i=0}^{R-n-1} r^*_{n+1,i}\frac{T^{i+1}}{i+1} + \alpha^*_n\sum_{i=1}^{R-n-1} r_{n+1,i}\frac{T^{i+1}}{i+1}\Bigg]\Bigg), \quad n=\overline{H+1,R-2} \ .$$

Hereon all relations for computation of quantities $m^*_n(x)$ are found.

Let us now denote by L_n, $n=\overline{H+1,R-1}$, – mean time mean time to the time instant when the total number of customers in the system becomes equal H for the first time, provided that *at an arbitrary* time instant there are total of n customers in the system and system is in "overloaded" state. Additionally denote by L^*_n, $n=\overline{H+1,R}$ mean time to the time instant when the total number of customers in the system becomes equal H, provided that *at an arbitrary* time instant there are total of n customers in the system and system is in "blocking" state. We assume that if $n=\overline{H+1,R-1}$, then at *at an arbitrary* time instant customer in server has *remaining* service time T, and if $n=R$ then customer in server has *remaining* service time $T/2$. This assumption leads to certain calculation error, but this error is insignificant already for small values of $R-H$. For example, when $R-H=10$ the error is estimated at 5%, and for $R-H=50$ it does not exceed 1%. It is possible to obtain exact expressions for L_n and L^*_n but slight increase of accuracy will lead to serious complication

of calculations. This is seen to be impractical. Thus taking this reasoning into consideration, we can write out expressions for L_n and L_n^*:

$$L_n^* = (n-H)T, \quad n = \overline{H+1, R-1}, \quad L_R^* = \left(R - H - \frac{1}{2}\right)T,$$
$$L_{H+1} = m_{H+1} + m_{H+1}^* + \alpha_{H+1}^* L_R^*,$$
$$L_n = m_n + \alpha_n L_{n-1} + m_n^* + \alpha_n^* L_R^*, \quad n = \overline{H+2, R-1} \ .$$

The lattermost functions that have to be introduced are $m_n^*(x), \ n = \overline{L,H}$, – mean time to the time instant when the total number of customers becomes equal $n-1$ for the first time, provided that at *arbitrary* time instant there are n customers in the system, *remaining* service time of currently served customer is x, and system is in "overloaded" state. Denote $m_n^* = m_n^*(T), \ n = \overline{L,H}$. Repeating the same arguments, that we used to obtain equations for $m_n(x)$, one can verify, that it holds

$$m_H^{*\prime}(x) = 1 - \lambda_1 m_H^*(x) + \lambda_1(m_{H+1}(x) + m_{H+1}^*(x) + \alpha_{H+1}^*(x)L_R^* + m_H^*), \quad (23)$$
$$m_n^{*\prime}(x) = 1 - \lambda_1 m_n^*(x) + \lambda_1(m_{n+1}^*(x) + m_n^*), \quad n = \overline{L, H-1} \ , \quad (24)$$

with boundary conditions $m_n^*(0) = 0, \ n = \overline{L,H}$. By analogy with (16)–(17) the solution of (23)–(24) has the form $m_n^*(x) = e^{-\lambda_1 x} u_n^*(x), \ n = \overline{L,H}$, where

$$u_n^*(x) = \sum_{i=0}^{R-n} r_{n,i}^* x^i + t_n e^{\lambda_1 x}, \quad n = \overline{L,H} \ .$$

Coefficients $r_{n,i}^*$ and t_n for $n = \overline{L,H}$ completely in the same way as it is done when solving (16)–(17). Due to the lack of space we do not state here intermediate calculations and provide final expressions for the coefficients:

$$t_H = \frac{(1 + \lambda_1[L_R^* + t_{H+1} + m_H^*])}{\lambda_1}, t_n = \frac{(1+\lambda_1[t_{n+1} + m_n^*])}{\lambda_1}, \quad n = \overline{H-1, L},$$
$$r_{H,0}^* = -\frac{(1 + \lambda_1[L_R^* + t_{H+1} + m_H^*])}{\lambda_1}, \quad r_{H,1}^* = \lambda_1\left(r_{H+1,0}^* - L_R^* c_{H+1,0}\right),$$
$$r_{H,i}^* = \lambda_1\left(\frac{r_{H+1,i-1}}{i} + \frac{r_{H+1,i-1}^*}{i} - L_R^* \frac{c_{H+1,i-1}}{i}\right), \ i = 2, R-H-1,$$
$$r_{H,R-H}^* = \lambda_1\left(\frac{r_{H+1,R-H-1}}{R-H} + \frac{r_{H+1,R-H-1}^*}{R-H}\right),$$
$$r_{n,0}^* = -\frac{(1 + \lambda_1[t_{n+1} + m_n^*])}{\lambda_1}, \quad r_{n,i}^* = \lambda_1 \frac{r_{n+1,i-1}^*}{i}, \quad i = \overline{1, R-n}, \quad n = \overline{H-1, L} \ .$$

In the above expressions quantities $m_n^*, \ n = \overline{L,H}$, remain unknown. They can be found from equation $m_n^*(x) = e^{-\lambda_1} u_n^*(x)$, if one puts $x = T$. Omitting intermediate calculations we arrive at the following expressions for m_n^*:

$$m_H^* = \frac{(1 + \lambda_1[L_R^* + t_{H+1}])}{\lambda_1}(e^{\lambda_1 T} - 1) +$$

$$+\lambda_1\left(\sum_{i=1}^{R-H-1} r_{H+1,i}\frac{T^{i+1}}{i+1}+\sum_{i=0}^{R-H-1}\frac{r^*_{H+1,i}T^{i+1}}{i+1}-L^*_R\sum_{i=0}^{R-H-2}c_{H+1,i}\frac{T^{i+1}}{i+1}\right),$$

$$m^*_n=\frac{(1+\lambda_1 t_{n+1})}{\lambda_1}(e^{\lambda_1 T}-1)+\lambda_1\sum_{i=0}^{R-n-1}r^*_{n+1,i}\frac{T^{i+1}}{i+1},\quad n=\overline{H-1,L}\ .$$

Now everything is ready for writing out expressions for M_n and M^*_n. Recall that M_n, $n=\overline{L,R-1}$ is mean time to the time instant when the total number of customers in the system becomes equal $L-1$ for the first time, given that at an arbitrary moment of time there are total of n, $n=\overline{L,R-1}$ customers in the considered queueing system and it is in "overloaded" state. We still assume, that at arbitrary moment of time remaining service time of customer in service is T. Then it is easy to see, that $M_L=m^*_L$, $M_n=m^*_n+M_{n-1}$, $n=\overline{L+1,H}$, $M_n=L_n+M_H$, $n=\overline{H+1,R-1}$ Remembering, that M^*_n, $n=\overline{H+1,R}$ is mean time to the time instant when the total number of customers in the system becomes equal $L-1$ for the first time, given that at an arbitrary moment of time there are total of n, $n=\overline{H+1,R}$ customers in the considered queueing system and it is in "blocking" state, it becomes clear that $M^*_n=L^*_n+M_H,\ \ n=\overline{H+1,R}$.

4 Conclusion

Consideration is given to $M_2|D|1|R<\infty$ queueing system with bi-level hysteretic policy. The policy implies that system may be in three states (normal, overloaded, blocking). One of the important performance characteristics of hysteric policy and system itself – mean return time to normal state of operation – is being analyzed. New method that allows fast computation of this characteristic is proposed. All theoretical results were compared with simulation results, obtained with the use of GPSS, and showed good accuracy. Further research in this area will be concentrated on the verification of obtained results by comparing them with simulation based on real time SIP traffic.

Acknowledgments. The reported study was partially supported by RFBR, research project No. 12-07-00108 and No. 11-07-00112.

References

1. ITU-T Recommendation Q.704. Signalling System No.7 – Message Transfer Part, Signalling network functions and messages (1996)
2. Abaev, P., Gaidamaka, Y., Pechinkin, A., Razumchik, R., Shorgin, S.: Simulation of overload control in SIP server networks. In: Proc. of the 26th European Conference on Modelling and Simulation, pp. 533–539 (2012)
3. Hilt, V., Noel, E., Shen, C., Abdelal, A.: Design Considerations for Session Initiation Protocol (SIP) Overload Control. RFC6357 (2011)
4. Abaev, P., Pechinkin, A., Razumchik, R.: Analysis of queueing system with constant service time for SIP server hop-by-hop overload control. In: Dudin, A., et al. (eds.) BWWQT 2013. CCIS, vol. 356, pp. 1–10. Springer, Heidelberg (2013)

Discrete-Time Queueing System with Expulsions

Iván Atencia, Inmaculada Fortes*, and Sixto Sánchez

E.T.S. Ingeniería Informática
Dept. Matemática Aplicada, Málaga, Spain
iatencia@ctima.uma.es

Abstract. In this paper we analyze a discrete-time queueing system in which an arriving customer can decide, with a certain probability, to go directly to the server expelling out of the system the customer that is currently in service or to join the queue in the last place. The arrivals are assumed to be geometrical and the service times are arbitrarily distributed. We present some numerical examples in order to illustrate the effect of the parameters on several performance characteristics.

Keywords: Discrete-time, expulsions, recurrent formulae.

1 Introduction

The standard models of classical queueing theory are systems operating in continuos time. But in practice there are many systems which shows an inherent generic slotted time scale (for example time–shared computing systems) and demands a serious study of discrete time systems. One of the advantages of dealing with discrete-time models is that they have been found more appropriate that their continuous-time counterpart for modelling computer and telecommunication systems. The discrete time scale often reflects the nature of an underlying application: for example, the clock time unit in a computer system fixed size data units (bits, bytes, fixed length packets) on a communication channel, etc.

The study of discrete-time queues was initiated by Meisling [7], Birdsall et al. [2], and also by Powell et al. [10]. Reference works and more detailed applications on discrete–time queueing theory include the monographs [3,11]. Further, a detailed treatment regarding this subject can be found in a two–volume book on applied probability [5,6].

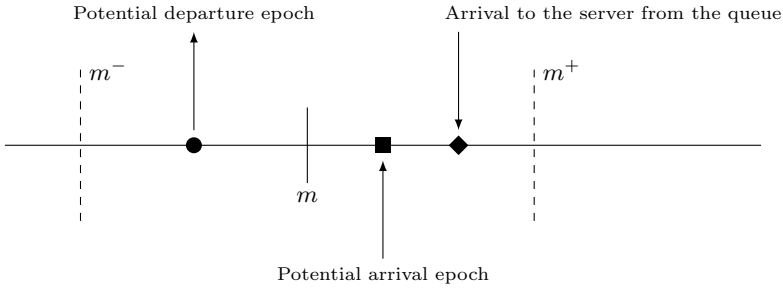
In our work we consider a discrete-time single-server queueing system with expulsions. The expulsions can be controlled by the server and decides weather the new incoming work or customer is worth to update the server and expel out of the system the current one in service or continue servicing and the new arrival joins the queue in order to be served later on. Let us note that the customers arriving from outside has priority on others. In order to avoid trivial cases we will suppose $0 < a < 1$.

* The work of I. Fortes and S. Sánchez is partially supported by the Junta of Andalucía [P09-FQM-5233].

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 20–25, 2013.

2 The Mathematical Model

We consider a single-server discrete-time queueing system where the time axis is divided into a sequence of equal time intervals (called slots) and it is assumed that all queueing activities (arrivals and departures) take place at the slot boundaries. For mathematical convenience, we will suppose that the departures occur at the moment immediately before the slot boundaries and the arrivals occur at the moment immediately after the slot boundaries, that is:



Customers arrive according to a geometric arrival process with rate a, that is, a is the probability that a customer arrives at a slot. If, upon arrival, the service is idle, the service of the arriving customer begins immediately, otherwise, the arriving customer either with probability θ expels the customer that is currently being served out of the system and starts immediately its service, or with complementary probability $\bar{\theta} = 1 - \theta$ joins the last place of the queue.

Service times are governed by an arbitrary distribution $\{s_i\}_{i=1}^{\infty}$, with generating functions $S(x) = \sum_{i=1}^{\infty} s_i\, x^i$. We will denote by $S_k = \sum_{i=k}^{\infty} s_i;\ \ k \geq 1$, the probability that the service lasts not less than k slots.

3 The Markov Chain

At time m^+ the system can be described by the Markov process $\{Y_m\,,\ m \in \mathbb{N}\}$ with $Y_m = (C_m, \xi_m, N_m)$ where C_m takes the values 0, or 1 according to the server is free or busy and N_m is the number of customers in the queue. If $C_m = 1$, ξ_m corresponds to the remaining service time.

It can be shown that $\{Y_m\,,\ m \in \mathbb{N}\}$ is the Markov chain of our queueing system, whose states space is

$$\{(0); (1, i, k) : i \geq 1, k \geq 0\}$$

Our goal is to determine the stationary distribution

$$\pi_0 = \lim_{m\to\infty} P[C_m = 0, N_m = 0]$$
$$\pi_{1,i,k} = \lim_{m\to\infty} P[C_m = 1, \xi_m = i, N_m = k]$$

The Kolmogorov equations for the stationary distribution of the system are given by

$$\pi_0 = \bar{a}\pi_0 + \bar{a}\pi_{1,1,0} \Longleftrightarrow a\pi_0 = \bar{a}\pi_{1,1,0} \tag{1}$$

$$\pi_{1,i,0} = as_i\pi_0 + \bar{a}\pi_{1,i+1,0} + \bar{a}s_i\pi_{1,1,1} + as_i\pi_{1,1,0} + a\theta s_i \sum_{j=2}^{\infty} \pi_{1,j,0} \tag{2}$$

$$\begin{aligned}\pi_{1,i,k} = {} & \bar{a}\pi_{1,i+1,k} + \bar{a}s_i\pi_{1,1,k+1} + as_i\pi_{1,1,k} + \\ & + a\theta s_i \sum_{j=2}^{\infty} \pi_{1,j,k} + a\bar{\theta}\pi_{1,i+1,k-1},\ k \geq 1\end{aligned} \tag{3}$$

Eqs. (2) and (3) can be written in the following way:

$$\begin{aligned}\pi_{1,i,k} = {} & \delta_{0k}\, as_i\pi_0 + \bar{a}\pi_{1,i+1,k} + \bar{a}s_i\pi_{1,1,k+1} + as_i\pi_{1,1,k} + \\ & + a\theta s_i \sum_{j=2}^{\infty} \pi_{1,j,k} + (1-\delta_{0,k})\, a\bar{\theta}\pi_{1,i+1,k-1},\ k \geq 0\end{aligned} \tag{4}$$

where $\delta_{a,b}$ is the Kronecker's symbol and the normalizing condition is

$$\pi_0 + \sum_{i=1}^{\infty}\sum_{k=0}^{\infty} \pi_{1,i,k} = 1.$$

With the aim of solving Eq. (4) we introduce the following generating function

$$\varphi(x,z) = \sum_{i=1}^{\infty}\sum_{k=0}^{\infty} \pi_{1,i,k} x^i z^k = \sum_{i=1}^{\infty} \varphi_i(z) x^i$$

where $\varphi_i(z)$ is the auxiliary function

$$\varphi_i(z) = \sum_{k=0}^{\infty} \pi_{1,i,k} z^k.$$

Multiplying equation (4) by z^k and summing over k, we have

$$\varphi_i(z) = (\bar{a} + a\bar{\theta}z)\varphi_{i+1}(z) + \frac{\bar{a} + a\bar{\theta}z}{z} s_i\varphi_1(z) + a\theta s_i\varphi(1,z) - \frac{1-z}{z} a\pi_0 s_i \tag{5}$$

Multiplying the former equation by x^i and summing over i we get:

$$\begin{aligned}\frac{x - (\bar{a} + a\bar{\theta}z)}{x}\varphi(x,z) = {} & (\bar{a} + a\bar{\theta}z)\frac{S(x) - z}{z}\varphi_1(z) + \\ & + a\theta S(x)\varphi(1,z) - \frac{1-z}{z} aS(x)\pi_0\end{aligned} \tag{6}$$

Setting $x = 1$ and $x = \bar{a} + a\bar{\theta}z$ in the above equation, respectively we obtain

$$\varphi_1(z) = \frac{S(\bar{a} + a\bar{\theta}z)(1 - \bar{\theta}z)}{(\bar{a} + a\bar{\theta}z)[S(\bar{a} + a\bar{\theta}z) - \bar{\theta}z]}\, a\pi_0 \tag{7}$$

$$\varphi(1,z) = \frac{1 - S(\bar{a} + a\bar{\theta}z)}{S(\bar{a} + a\bar{\theta}z) - \bar{\theta}z}\, \pi_0 \tag{8}$$

By substituting (7) and (8) in (6) yields

$$\varphi(x,z) = \frac{S(x) - S(\bar{a} + a\bar{\theta}z)}{x - (\bar{a} + a\bar{\theta}z)} \cdot \frac{xz(1-\bar{\theta}z)}{S(\bar{a} + a\bar{\theta}z) - \bar{\theta}z} a\pi_0 \tag{9}$$

The normalization condition, that can be written as $\pi_0 + \varphi(1,1) = 1$, allow us to find out the unknown constant π_0:

$$\pi_0 = \frac{S(\bar{a} + a\bar{\theta}) - \bar{\theta}}{1 - \bar{\theta}}$$

Therefore, the necessary condition for the stability of the system is $S(\bar{a}+a\bar{\theta}) > \bar{\theta}$.

We summarize the above results in the following theorem:

Theorem 1. *The generating functions of the stationary distribution of the chain are given by*

$$\varphi_1(z) = \frac{S(\bar{a} + a\bar{\theta}z)(1-\bar{\theta}z)}{(\bar{a} + a\bar{\theta}z)[S(\bar{a} + a\bar{\theta}z) - \bar{\theta}z]} a\pi_0$$

$$\varphi(1,z) = \frac{1 - S(\bar{a} + a\bar{\theta}z)}{S(\bar{a} + a\bar{\theta}z) - \bar{\theta}z} \pi_0$$

$$\varphi(x,z) = \frac{S(x) - S(\bar{a} + a\bar{\theta}z)}{x - (\bar{a} + a\bar{\theta}z)} \cdot \frac{xz(1-\bar{\theta}z)}{S(\bar{a} + a\bar{\theta}z) - \bar{\theta}z} a\pi_0$$

where $\pi_0(z)$ is given by

$$\pi_0 = \frac{S(\bar{a} + a\bar{\theta}) - \bar{\theta}}{1 - \bar{\theta}}$$

In order to design a responsive system, a probabilistic assessment of factors like queue length, timing, and composition must be made. Queueing theory provides some powerful tools to help make this assessment, and is an absolutely essential part of any communication design.

Lemma 1. *1. The GF of the number of customers in the system is:*

$$\Phi(z) = \pi_0 + z\varphi(1,z) = \frac{S(\bar{a} + a\bar{\theta}z)(1-z) + z\theta}{S(\bar{a} + a\bar{\theta}z) - \bar{\theta}z} \pi_0$$

2. The GF of the number of customers in the queue is:

$$\Psi(z) = \pi_0 + \varphi(1,z) = \frac{1 - \bar{\theta}z}{S(\bar{a} + a\bar{\theta}z) - \bar{\theta}z} \pi_0$$

3. The mean number of the customers in the system is given by

$$E[L] = \Phi'(1) = \frac{S(\bar{a} + a\bar{\theta})[1 - S(\bar{a} + a\bar{\theta})] - a\theta\bar{\theta}S'(\bar{a} + a\bar{\theta})}{\theta[S(\bar{a} + a\bar{\theta}) - \bar{\theta}]}$$

4. *The mean number of the customers in the queue is given by*

$$E[N] = \Psi'(1) = \frac{\bar{\theta}[1 - S(\bar{a} + a\bar{\theta})] - a\theta\bar{\theta}S'(\bar{a} + a\bar{\theta})}{\theta[S(\bar{a} + a\bar{\theta}) - \bar{\theta}]}$$

5. *The mean sojourn time of a customer in the system and in the queue are given by*

$$\bar{v} = \frac{E[L]}{a}$$
$$\bar{w} = \frac{E[N]}{a}.$$

4 Numerical Results

This section is devoted to illustrate the effect of the parameters on several performance characteristics. Throughout this section, we assume that the mean service time is equal to 3 of a $BN(2, 0.4)$ for the service time. Of course, in all the below examples, the parametric values are chosen so as to satisfy the stability condition.

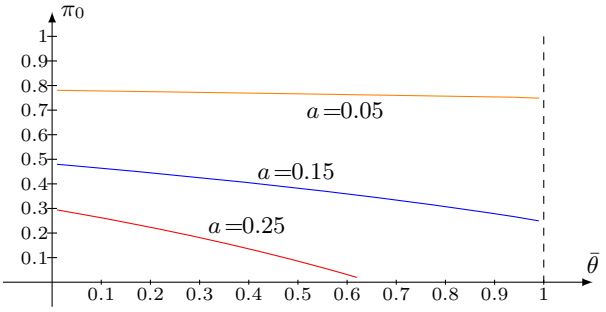


Fig. 1. The probability that the system is empty

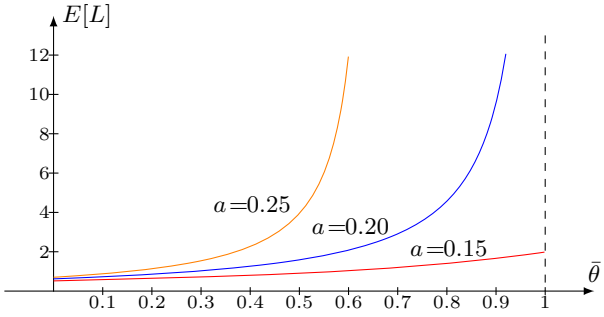


Fig. 2. The behavior of $E[L]$ against the parameter $\bar{\theta}$

In figure 1 the probability that the system is empty is plotted against the parameter $\bar{\theta}$ for different values of a ($a = 0.05, 0.15, 0.25$). As we expected, π_0 decreases with increasing values of $\bar{\theta}$ depending also on the arrival rate a.

The graphic plotted in fig 2 illustrates the behavior of $E[L]$ against the parameter $\bar{\theta}$. As intuition tells us, $E[L]$ increases with increasing values of $\bar{\theta}$ depending also on the arrival rate.

References

1. Atencia, I., Pechinkin, A.V.: A discrete-time queueing system with optional LCFS discipline. Ann. Oper. Res. (2012), doi:10.1007/s10479-012-1097-2
2. Birdsall, T., Ristenbatt, M., Weinstein, S.: Analysis of asynchronous time multiplexing of speech sources. IRE Transactions on Communication Systems 10, 390–397 (1962)
3. Bruneel, H., Kim, B.G.: Discrete-time models for communication systems including ATM. Kluwer Academic Publishers, Boston (1993)
4. Cascone, A., Manzo, P., Pechinkin, A., Shorgin, S.: A $Geom/G/1/n$ system with a LIFO discipline without interruptions in the service and with a limitation for the total capacity for the customers. Avtomatika i Telemejanika 1, 107–120 (2011) (in Russian)
5. Hunter, J.: Mathematical Techniques of Applied Probability. Operations Research and Industrial Engineering, vol. 1. Academic Press, New York (1983)
6. Hunter, J.: Mathematical Techniques of Applied Probability, Discrete-Time Models: Techniques and Applications. Operations Research and Industrial Engineering, vol. 2. Academic Press, New York (1983)
7. Meisling, T.: Discrete time queueing theory. Oper. Res. 6, 96–105 (1958)
8. Pechinkin, A., Svischeva, T.: The stationary state probability in the BMAP/G/1/r queueing system with inverse discipline and probabilistic priority. In: Transactions of XXIV International Seminar on Stability Problems for Stochastic Models, Jurmala, Latvia, September 10-17, pp. 141–174 (2004)
9. Pechinkin, A., Shorgin, S.: A $Geo/G/1/\infty$ system with a one non-standard discipline for the service. Informatics and it Applications 2, 55–62 (2008) (in Russian)
10. Powell, B., Avi-Itzhak, B.: Queueing systems with enforced idle times. Operations Research 15(6), 1145–1156 (1967)
11. Takagi, H.: Queueing analysis: A foundation of performance evaluation. Discrete-Time Systems, vol. 3. North-Holland, Amsterdam (1993)

Stationary Distribution Invariance of an Open Queueing Network with Temporarily Non-active Customers

Julia Bojarovich* and Yury Malinkovsky

Gomel State University
juls1982@list.ru, malinkovsky@gsu.by

Abstract. This paper considers stationary functioning of an open queueing network with temporarily non-active customers. Non-active customers are in a system queue and do not get service. Customers can pass from non-active state into state, when they can get their service and vice versa. Stationary distribution invariance with reference to service time distribution functional form is obtained.

Keywords: queueing network, temporarily non-active customers, stationary distribution invariance.

1 Introduction

Nowadays one pay considerable attention to queueing systems reliability. Herewith, the problem of customer reliability becomes actual to a marked degree too. Queueing network with temporarily non-active customers is a model with customers, which are partly unreliable. Non-active customers are in a system queue and do not get service, partly loosing their capacity for service. Customers can pass from non-active state into state, when they can get their service and vice versa. In papers [1], [2] G. Tsitsiashvili and M. Osipova observed an open queueing network with non-active customers and established the form of stationary distribution. We have generalized their result in a case of random distributed service times. We have obtained stationary distribution invariance with reference to service time distribution functional form.

2 Queueing Network Description

An open queueing network with set of systems $J = \{1, 2, \ldots, N\}$ is considered. Customers arrive at the network according to a Poisson processes at rates λ_i, $i \in J$. Non-active customers are in a system queue and can not get service. There are input Poisson flows of signals at rates ν_i and φ_i, $i \in J$. When arriving at the system $i \in J$ the signal at rate ν_i induces an ordinary customer at system, if any, to become non-active. When arriving at the system $i \in J$ the signal at rate

* Corresponding author.

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 26–32, 2013.

φ_i induces an non-active customer, if any, to become an ordinary. Signals do not need service. Let $n_i(t), n_i'(t)$ are numbers of ordinary and non-active customers at system $i \in J$ at time t accordingly. States space for process $z(t) = (n_i(t), n_i'(t))$ is $Z = \{((n_1, n_1'), \ldots, (n_N, n_N'))|\, n_i,\, n_i' \geq 0,\, i \in J\}$. Service times are independent random distributed values with functions of distribution $B_i(n_i + n_i', x_{n_i+n_i'})$ and expected values $1/\mu_i$, $i \in J$. After finishing of service process at system $i \in J$ customer is routed to system $j \in J$ with the probability $p_{i,j}$ and with the probability $p_{i,0}$ is removed from network ($\sum_{j=1}^{N} p_{i,j} + p_{i,0} = 1$), $i \in J$. Let $p_{i,i} = 0$, $i \in J$. The discipline of service is LSFS-PR.

A traffic equations system is:

$$\varepsilon_i = \lambda_i + \sum_{j=1}^{N} \varepsilon_j p_{j,i}, \quad i \in J. \tag{1}$$

One can prove that under certain conditions traffic equations system has unique non-trivial solution.

3 Stationary Distribution Invariance

G. Tsitsiashvili, M. Osipova [1], [2] considered an open queueing network with temporarily non-active customers and exponentially distributed service time. It has been proved that under conditions of ergodicity

$$\varepsilon_i < \mu_i,$$

$$\varepsilon_i \nu_i < \mu_i \varphi_i,\ i = 1, \ldots, N,$$

process $z(t) = (n_i(t), n_i'(t))$ has stationary distribution

$$p(z) = p_1(n_1, n_1') p_2(n_2, n_2') \ldots p_N(n_N, n_N'),\ z \in Z,$$

where

$$p_i(n_i, n_i') = \Big(1 - \frac{\varepsilon_i \nu_i}{\varphi_i \mu_i}\Big)\Big(1 - \frac{\varepsilon_i}{\mu_i}\Big)\Big(\frac{\varepsilon_i \nu_i}{\varphi_i \mu_i}\Big)^{n_i'}\Big(\frac{\varepsilon_i}{\mu_i}\Big)^{n_i},\ i = 1, \ldots, N,$$

ε_i, $i \in J$ – is a traffic equations system solution.

We consider an open queueing network, where service times are independent random distributed values. In this case $z(t)$ is not Markov process. So we introduce Markov process $\zeta(t) = (z(t), \xi(t))$, where $\xi(t) = (\xi_1(t), \ldots, \xi_N(t))$, $\xi_i(t) = (\xi_{i,1}(t), \ldots, \xi_{i,n_i+n_i'}(t))$. Here $\xi_{i,k}(t)$ – rest service time of a customer, which has position k at system i at time t, $i \in J$.

Denote by

$$F(z, x) = F(z, x_{1,1}, \ldots, x_{1,n_1+n_1'}; x_{2,1}, \ldots, x_{2,n_2+n_2'}; \ldots; x_{N,1}, \ldots, x_{N,n_N+n_N'}) =$$

$$= \lim_{t \to \infty} P\{z(t) = z, \xi_{i,1}(t) < x_{i,1}, \ldots, \xi_{i,n_i+n_i'}(t) < x_{i,n_i+n_i'}, i \in J\}.$$

Functions $F(z, x)$ are called stationary functions of $\zeta(t)$ distribution.

Theorem 1. *Under conditions of ergodicity:*

$$\varepsilon_i < \mu_i, \tag{2}$$

$$\varepsilon_i \nu_i < \mu_i \varphi_i,\ i = 1, \ldots, N, \tag{3}$$

stationary functions of $\zeta(t)$ *distribution* $F(z, x)$ *are:*

$$F(z,x) = p_1(n_1, n_1')p_2(n_2, n_2') \ldots p_N(n_N, n_N') \times$$

$$\times \prod_{i=1}^{N} \mu_i^{n_i + n_i'} \prod_{s=1}^{n_i + n_i'} \int_0^{x_{i,s}} (1 - B_i(s,u))du,\ z \in Z, \tag{4}$$

where

$$p_i(n_i, n_i') = \Big(1 - \frac{\varepsilon_i \nu_i}{\varphi_i \mu_i}\Big)\Big(1 - \frac{\varepsilon_i}{\mu_i}\Big)\Big(\frac{\varepsilon_i \nu_i}{\varphi_i \mu_i}\Big)^{n_i'}\Big(\frac{\varepsilon_i}{\mu_i}\Big)^{n_i},\ i = 1, \ldots, N, \tag{5}$$

ε_i, $i \in J$ – *is a traffic equations system solution.*

Proof. Denote by $e_i \in Z$ – a vector, which coordinates equal 0 with the exception of $(n_i, n_i') = (1, 0)$, denote by $e_i' \in Z$ – a vector, which coordinates equal 0 with the exception of $(n_i, n_i') = (0, 1)$, $i \in J$.

Consider process $\zeta(t)$. In the case of exponentially distributed service times a process $z(t)$ was ergodic under conditions (2), (3) [1]. Accordingly a process $\zeta(t)$ is ergodic under conditions (2), (3), because $\zeta(t)$ is obtained from $z(t)$ by continuous components adding.

$\zeta(t)$ condition changes that occur through customers or signals arriving will name spontaneous changes. Suppose that h is small time interval and consider the probability

$$P\{z(t+h) = z, \xi_{i,1}(t+h) < x_{i,1}, \ldots, \xi_{i,n_i+n_i'}(t+h) < x_{i,n_i+n_i'},\ i \in J\}.$$

This event may occur in the following ways:

1. From the moment t during time h there were no spontaneous changes and service at any system was not over. The probability of this event is

$$P\{z(t) = z, \xi_{i,1}(t) < x_{i,1}, \ldots, hI_{n_i>0} \le \xi_{i,n_i+n_i'}(t) < x_{i,n_i+n_i'} + hI_{n_i>0},\ i \in J\} \times$$

$$\times (1 - \sum_{i=1}^{N} (\lambda_i + \nu_i I_{n_i>0} + \varphi_i I_{n_i'>0})h + o(h)).$$

2. During time h a customer has arrived at system $i \in J$. There were no other spontaneous changes. No customer was serviced.

$$P\{z(t) = z - e_i, \xi_{k,1}(t) < x_{k,1}, \ldots, hI_{n_k>0} \le \xi_{k,n_k+n_k'}(t) < x_{k,n_k+n_k'} + hI_{n_k>0},$$

$$k \in J,\ k \neq i,$$

$$\xi_{i,1}(t) < x_{i,1}, \ldots, hI_{n_i>1} \leq \xi_{i,n_i+n'_i-1}(t) < x_{i,n_i+n'_i-1} + hI_{n_i>1}\} \times$$

$$\times(\lambda_i h + o(h))B_i(n_i + n'_i, x_{i,n_i+n'_i} + \theta h)I_{n_i>0},\ 0 < \theta < 1.$$

3. During time h a customer was serviced at system $i \in J$ and was routed to system $j \in J$. There were no spontaneous changes.

$$P\{z(t) =$$

$$= z - e_i + e_j, \xi_{k,1}(t) < x_{k,1}, \ldots, hI_{n_k>0} \leq \xi_{k,n_k+n'_k}(t) < x_{k,n_k+n'_k} + hI_{n_k>0},$$

$$k \in J,\ k \neq i,\ k \neq j,$$

$$\xi_{j,1}(t) < x_{j,1}, \ldots, \xi_{j,n_j+n'_j}(t) < x_{j,n_j+n'_j}, \xi_{j,n_j+n'_j+1}(t) < h,$$

$$\xi_{i,1}(t) < x_{i,1}, \ldots, hI_{n_i>1} \leq \xi_{i,n_i+n'_i-1}(t) < x_{i,n_i+n'_i-1} + hI_{n_i>1}\} \times$$

$$\times B_i(n_i + n'_i, x_{i,n_i+n'_i} + \theta h)p_{j,i}I_{n_i>0},\ 0 < \theta < 1.$$

4. During time h a customer was serviced at system $i \in J$ and was removed from the network. There were no spontaneous changes.

$$P\{z(t) = z+e_i, \xi_{k,1}(t) < x_{k,1}, \ldots, hI_{n_k>0} \leq \xi_{k,n_k+n'_k}(t) < x_{k,n_k+n'_k} + hI_{n_k>0},$$

$$k \in J,\ k \neq i,$$

$$\xi_{i,1}(t) < x_{i,1}, \ldots, \xi_{i,n_i+n'_i+1}(t) < h\}p_{i,0}.$$

5. During time h an informational signal at rate ν_i has arrived at system $i \in J$. There were no other spontaneous changes. No customer was serviced.

$$P\{z(t) =$$

$$= z + e_i - e'_i, \xi_{k,1}(t) < x_{k,1}, \ldots, hI_{n_k>0} \leq \xi_{k,n_k+n'_k}(t) < x_{k,n_k+n'_k} + hI_{n_k>0},$$

$$k \in J,\ k \neq i,$$

$$\xi_{i,1}(t) < x_{i,1}, \ldots, h \leq \xi_{i,n_i+n'_i}(t) < x_{i,n_i+n'_i} + h\}(\nu_i h + o(h))I_{n'_i>0}.$$

6. During time h an informational signal at rate φ_i has arrived at system $i \in J$. There were no other spontaneous changes. No customer was serviced.

$$P\{z(t) =$$

$$= z - e_i + e'_i, \xi_{k,1}(t) < x_{k,1}, \ldots, hI_{n_k>0} \leq \xi_{k,n_k+n'_k}(t) < x_{k,n_k+n'_k} + hI_{n_k>0},$$

$$k \in J,\ k \neq i,$$

$$\xi_{i,1}(t) < x_{i,1}, \ldots, hI_{n_i>1} \leq \xi_{i,n_i+n'_i}(t) < x_{i,n_i+n'_i} + hI_{n_i>1}\}(\varphi_i h + o(h))I_{n_i>0}.$$

7. During time h there were more than two changes of queueing network condition. This probability is $o(h)$.

Therefore

$$P\{z(t+h)=z,\xi_{i,1}(t+h)<x_{i,1},\dots,\xi_{i,n_i+n'_i}(t+h)<x_{i,n_i+n'_i},\ i\in J\}=$$
$$=P\{z(t)=z,\xi_{i,1}(t)<x_{i,1},\dots,hI_{n_i>0}\le\xi_{i,n_i+n'_i}(t)<x_{i,n_i+n'_i}+hI_{n_i>0},\ i\in J\}\times$$
$$\times(1-\sum_{i=1}^{N}(\lambda_i+\nu_iI_{n_i>0}+\varphi_iI_{n'_i>0})h+o(h))+\sum_{i=1}^{N}\Big(P\{z(t)=$$
$$=z-e_i,\xi_{k,1}(t)<x_{k,1},\dots,hI_{n_k>0}\le\xi_{k,n_k+n'_k}(t)<x_{k,n_k+n'_k}+hI_{n_k>0},$$
$$k\in J,\ k\ne i,$$
$$\xi_{i,1}(t)<x_{i,1},\dots,hI_{n_i>1}\le\xi_{i,n_i+n'_i-1}(t)<x_{i,n_i+n'_i-1}+hI_{n_i>1}\}\times$$
$$\times(\lambda_ih+o(h))B_i(n_i+n'_i,x_{i,n_i+n'_i}+\theta h)I_{n_i>0}+$$
$$+\sum_{j=1}^{N}P\{z(t)=z-e_i+e_j,\xi_{k,1}(t)<x_{k,1},\dots,hI_{n_k>0}\le\xi_{k,n_k+n'_k}(t)<x_{k,n_k+n'_k}+hI_{n_k>0},$$
$$k\in J,\ k\ne i,\ k\ne j,$$
$$\xi_{j,1}(t)<x_{j,1},\dots,\xi_{j,n_j+n'_j}(t)<x_{j,n_j+n'_j},\xi_{j,n_j+n'_j+1}(t)<h,$$
$$\xi_{i,1}(t)<x_{i,1},\dots,hI_{n_i>1}\le\xi_{i,n_i+n'_i-1}(t)<x_{i,n_i+n'_i-1}+hI_{n_i>1}\}\times$$
$$\times B_i(n_i+n'_i,x_{i,n_i+n'_i}+\theta h)p_{j,i}I_{n_i>0}+$$
$$+P\{z(t)=z+e_i,\xi_{k,1}(t)<x_{k,1},\dots,hI_{n_k>0}\le\xi_{k,n_k+n'_k}(t)<x_{k,n_k+n'_k}+hI_{n_k>0},$$
$$k\in J,\ k\ne i,$$
$$\xi_{i,1}(t)<x_{i,1},\dots,\xi_{i,n_i+n'_i+1}(t)<h\}p_{i,0}+$$
$$+P\{z(t)=z+e_i-e'_i,\xi_{k,1}(t)<x_{k,1},\dots,hI_{n_k>0}\le\xi_{k,n_k+n'_k}(t)<x_{k,n_k+n'_k}+hI_{n_k>0},$$
$$k\in J,\ k\ne i,$$
$$\xi_{i,1}(t)<x_{i,1},\dots,h\le\xi_{i,n_i+n'_i}(t)<x_{i,n_i+n'_i}+h\}(\nu_ih+o(h))I_{n'_i>0}+$$
$$+P\{z(t)=z-e_i+e'_i,\xi_{k,1}(t)<x_{k,1},\dots,hI_{n_k>0}\le\xi_{k,n_k+n'_k}(t)<x_{k,n_k+n'_k}+hI_{n_k>0},$$
$$k\in J,\ k\ne i,$$
$$\xi_{i,1}(t)<x_{i,1},\dots,hI_{n_i>1}\le\xi_{i,n_i+n'_i}(t)<x_{i,n_i+n'_i}+hI_{n_i>1}\}(\varphi_ih+o(h))I_{n_i>0}+o(h)\Big). \tag{6}$$

Every probability from (6) may be expressed in terms of functions $F_t(z,x)=P\{z(t)=z,\xi_{i,1}(t)<x_{i,1},\dots,\xi_{i,n_i+n'_i}(t)<x_{i,n_i+n'_i},\ i\in J\}$, taking into consideration that

$$P\{z(t)=z,\xi_{i,1}(t)<x_{i,1},\dots,h\le\xi_{i,n_i+n'_i}(t)<x_{i,n_i+n'_i}+h,\ i\in J\}=$$
$$=F_t(z,x_{i,1},\dots,x_{i,n_i+n'_i}+h,\ i\in J)-\sum_{k=1}^{N}F_t(z,x_{i,1},\dots,x_{i,n_i+n'_i}+h,$$
$$i\in J,i\ne k;x_{k,1},\dots,x_{k,n_k+n'_k-1},h)+\dots+F_t(z,x_{i,1},\dots,x_{i,n_i+n'_i-1},h,\ i\in J)$$

and $P\{z(t)=z, \xi_{i,1}(t) < x_{i,1}, \ldots, h \le \xi_{i,n_i+n'_i}(t) < x_{i,n_i+n'_i} + h, i \in J\} =$

$$= F_t(z, x_{i,1}, \ldots, x_{i,n_i+n'_i},\ i \in J) + \sum_{i=1}^{N} \frac{\partial F_t(z, x_{i,1}, \ldots, x_{i,n_i+n'_i},\ i \in J)}{\partial x_{i,n_i+n'_i}} h -$$

$$- \sum_{i=1}^{N} \frac{\partial F_t(z, x_{l,1}, \ldots, x_{l,n_l+n'_l},\ l \in J, l \ne i; x_{i,1}, \ldots, x_{i,n_i+n'_i-1}, 0)}{\partial x_{i,n_i+n'_i}} h + o(h).$$

Letting t tend to infinity, we obtain the following equations system:

$$F(z,x) = F(z,x) + h \sum_{i=1}^{N} \left(\frac{\partial F(z,x)}{\partial x_{i,n_i+n'_i}} - \left(\frac{\partial F(z,x)}{\partial x_{i,n_i+n'_i}} \right)_{x_{i,n_i+n'_i}=0} \right) I_{n_i>0} -$$

$$- \left(\sum_{i=1}^{N} \left(\lambda_i + \nu_i I_{n_i>0} + \varphi_i I_{n'_i>0} \right) h + o(h) \right) F(z,x) +$$

$$+ \sum_{i=1}^{N} F(z - e_i, x) B_i(n_i + n'_i, x_{n_i+n'_i}) (\lambda_i h + o(h)) I_{n_i>0} +$$

$$+ h \sum_{j=1}^{N} \sum_{i=1, i\ne j}^{N} p_{j,i} B_i(n_i + n'_i, x_{i,n_i+n'_i}) \left(\frac{\partial F(z + e_j - e_i, x)}{\partial x_{j,n_j+n'_j+1}} \right)_{x_{j,n_j+n'_j+1}=0} I_{n_i>0} +$$

$$+ h \sum_{i=1}^{N} p_{i,0} \left(\frac{\partial F(z + e_i, x)}{\partial x_{i,n_i+n'_i+1}} \right)_{x_{i,n_i+n'_i+1}=0} + \sum_{i=1}^{N} F(z + e_i - e'_i, x)(\nu_i h + o(h)) I_{n'_i>0} +$$

$$+ \sum_{i=1}^{N} F(z - e_i + e'_i, x)(\varphi_i h + o(h)) I_{n_i>0} + o(h). \quad (7)$$

Subtracting $F(z,x)$ from both sides of (7), dividing both sides of (7) by h and letting h tend to zero, we obtain the following differential equations system:

$$F(z,x) \sum_{i=1}^{N} \left(\lambda_i + \nu_i I_{n_i>0} + \varphi_i I_{n'_i>0} \right) =$$

$$= \sum_{i=1}^{N} \left(\frac{\partial F(z,x)}{\partial x_{i,n_i+n'_i}} - \left(\frac{\partial F(z,x)}{\partial x_{i,n_i+n'_i}} \right)_{x_{i,n_i+n'_i}=0} \right) I_{n_i>0} +$$

$$+ \sum_{i=1}^{N} F(z - e_i, x) B_i(n_i + n'_i, x_{n_i+n'_i}) \lambda_i I_{n_i>0} + \sum_{i=1}^{N} p_{i,0} \left(\frac{\partial F(z + e_i, x)}{\partial x_{i,n_i+n'_i+1}} \right)_{x_{i,n_i+n'_i+1}=0} + \quad (8)$$

$$+ \sum_{j=1}^{N} \sum_{i=1, i\ne j}^{N} p_{j,i} B_i(n_i + n'_i, x_{i,n_i+n'_i}) \left(\frac{\partial F(z + e_j - e_i, x)}{\partial x_{j,n_j+n'_j+1}} \right)_{x_{j,n_j+n'_j+1}=0} I_{n_i>0} +$$

$$+ \sum_{i=1}^{N} F(z + e_i - e'_i, x) \nu_i I_{n'_i>0} + \sum_{i=1}^{N} F(z - e_i + e'_i, x) \varphi_i I_{n_i>0}.$$

Divide (8) into local balance equations:

$$F(z,x)\big(\nu_i I_{n_i>0}+\varphi_i I_{n'_i>0}\big)=F(z+e_i-e'_i,x)\nu_i I_{n'_i>0}+F(z-e_i+e'_i,x)\varphi_i I_{n_i>0}, \tag{9}$$

$$\Big(\Big(\frac{\partial F(z,x)}{\partial x_{i,n_i+n'_i}}\Big)_{x_{i,n_i+n'_i}=0}-\frac{\partial F(z,x)}{\partial x_{i,n_i+n'_i}}\Big)I_{n_i>0}=$$

$$=\sum_{j=1,j\neq i}^{N}p_{j,i}B_i(n_i+n'_i,x_{i,n_i+n'_i})\Big(\frac{\partial F(z+e_j-e_i,x)}{\partial x_{j,n_j+n'_j+1}}\Big)_{x_{j,n_j+n'_j+1}=0}I_{n_i>0}+ \tag{10}$$

$$+F(z-e_i,x)B_i(n_i+n'_i,x_{n_i+n'_i})\lambda_i I_{n_i>0},$$

$$\lambda_i F(z,x)=p_{i,0}\Big(\frac{\partial F(z+e_i,x)}{\partial x_{i,n_i+n'_i+1}}\Big)_{x_{i,n_i+n'_i+1}=0}. \tag{11}$$

Substituting $F(z,x)$, determined by means of (4), (5) into local balance equations (9) - (11), considering traffic equation system (1), we obtain identity. □

Denote by $\{p(z),\ z\in Z\}$ – stationary distribution of process $z(t)$. From the foregoing theorem, considering equality $p(z)=F(z,+\infty)$, we obtain

Corollary. Under conditions of ergodicity (2), (3) process $z(t)$ has stationary distribution

$$p(z)=p_1(n_1,n'_1)p_2(n_2,n'_2)\dots p_N(n_N,n'_N),\ z\in Z,$$

where $p_i(n_i,n'_i)$, $i\in J$ may be found by means of (5).

4 Conclusion

We have considered stationary functioning of an open queueing network with temporarily non-active customers. Expression for stationary distribution has been derived. Stationary distribution invariance with reference to service time distribution functional form has been obtained. Research results have practical importance and may be used for real networks investigation.

References

1. Tsitsiashvili, G.S., Osipova, M.: Distributions in stochastic network models. Nova Publishers (2008)
2. Tsitsiashvili, G.S., Osipova, M.: Queueing models with different schemes of customers transformations. In: Proceedings of the 19th International Conference Mathematical Methods for Increasing Efficiency of Information Telecommunication Networks, pp. 128–133 (2007)
3. Gnedenko, B., Kovalenko, I.: Introduction to queueing theory, Moscow, Nauka (1987)
4. Malinkovsky, Y., Bojarovich, J.: An open queueing network with partly non-active customers. In: Proceedings of the 21st International Conference Modern Probbabilistic Methods for Analysis and Optimization of Information and Telecommunication Networks, pp. 34–37 (2011)

An Open Queueing Network with Temporarily Non-active Customers and Rounds

Julia Bojarovich* and Larisa Marchenko

Gomel State University
juls1982@list.ru, lamarchenko@yandex.ru

Abstract. An open queueing network with partly non-active customers is considered. Non-active customers are in a system queue and do not get service. Customers can pass from non-active state into state, when they can get their service and vice versa. The form of stationary distribution and conditions of stationary distribution existence are obtained.

Keywords: queueing network, temporarily non-active customers, rounds, stationary distribution.

1 Introduction

Nowadays queueing networks with partly non-active customers become actual to a marked degree. Non-active customers are in a system queue and do not get service. We consider network, where customers may partly loose their capacity for service. Customers can pass from non-active condition into condition, when they can get their service and vice versa.

In paper [1] G. Tsitsiashvili and M. Osipova have observed an open queueing network with non-active customers and have established the form of stationary distribution. This paper generalizes results for network from [1]. We consider model with temporarily non-active customers and rounds of queueing systems. We have researched the form of stationary distribution and have established the criterion of stationary distribution existence.

2 An Open Queueing Network with Temporarily Non-active Customers and Rounds

Consider an open queueing network with set of systems $J = \{1, 2, \ldots, N\}$. Customers arrive at the network according to Poisson processes at rates λ_i, $i \in J$. There are input Poisson flows of signals at rates ν_i and φ_i, $i \in J$. When arriving at the system $i \in J$ the signal at rate ν_i induces an ordinary customer at system, if any, to become non-active. When arriving at the system $i \in J$ the signal at rate φ_i induces an non-active customer, if any, to become an ordinary. Non-active customers are in a system queue and can not get service. Signals do not need service.

* Corresponding author.

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 33–36, 2013.

Service times are independent exponentially distributed random values with parameters $\mu_i,\ i \in J$. When arriving at the system i customer queues up to the system with the probability f_i and with the probability $1 - f_i$ the customer goes round the system $i \in J$ (such customer is considered to be served). After finishing of service process at system $i \in J$ customer is routed to system $j \in J$ with the probability $p_{i,j}$ and with the probability $p_{i,0}$ is removed from network $(\sum_{j=1}^{N} p_{i,j} + p_{i,0} = 1), i \in J$. Let $p_{i,i} = 0, i \in J$. Let $n_i(t), n'_i(t)$ are numbers of ordinary and non-active customers at system $i \in J$ at time t accordingly. Consider $X(t) = \Big((n_1(t), n'_1(t)), \ldots, (n_N(t), n'_N(t))\Big)$. $X(t)$ is a continuous-time Markov chain. States space for process $X(t)$ is $Z = \{((n_1, n'_1), \ldots, (n_N, n'_N))|\, n_i,\, n'_i \geq 0,\, i \in J\}$.

A traffic equations system is:

$$\varepsilon_i = \lambda_i + \sum_{j=1}^{N} \varepsilon_j p_{j,i}, \quad i \in J. \tag{1}$$

One can prove that under certain conditions traffic equations system has unique non-trivial solution.

Theorem 1. *Under conditions of ergodicity:*

$$\varepsilon_i f_i < \mu_i, \tag{2}$$

$$\varepsilon_i f_i \nu_i < \mu_i \varphi_i,\ i = 1, \ldots, N, \tag{3}$$

$X(t)$ has stationary distribution:

$$\pi(n, n') = \pi_1(n_1, n'_1)\pi_2(n_2, n'_2) \ldots \pi_N(n_N, n'_N), \tag{4}$$

where

$$\pi_i(n_i, n'_i) = \Big(1 - \frac{\varepsilon_i f_i}{\mu_i}\Big)\Big(1 - \frac{\varepsilon_i f_i \nu_i}{\mu_i \varphi_i}\Big)\Big(\frac{\varepsilon_i f_i}{\mu_i}\Big)^{n_i}\Big(\frac{\varepsilon_i f_i \nu_i}{\mu_i \varphi_i}\Big)^{n'_i}, \tag{5}$$

here $\varepsilon_i,\ i \in J$ – is a traffic equations system solution.

Proof. Consider the following events:

1. A customer sent to the system $i \in J$, will not change the state of the network. The probability of this event denote by ψ_i.
2. A customer sent to the system $i \in J$, will be served by system $j \in J$ first time. The probability of this event denote by $\psi_{i,j}$.
3. A customer served by system $i \in J$, will not change the state of the network. The probability of this event denote by β_i.
4. A customer served by system $i \in J$, will be served by system $j \in J$ first time. The probability of this event denote by $\beta_{i,j}$.

It has been obtained in [5], that

$$\psi_i = (1 - f_i)(p_{i,0} + \sum_{j=1}^{N} \psi_j p_{i,j}); \tag{6}$$

$$\psi_{i,j} = f_i \delta_{i,j} + (1 - f_i) \sum_{k=1}^{N} p_{i,k} \psi_{k,j}; \tag{7}$$

$$\beta_i = p_{i,0} + \sum_{j=1}^{N} p_{i,j} \psi_j; \tag{8}$$

$$\beta_{i,j} = \sum_{k=1}^{N} p_{i,k} \psi_{k,j}. \tag{9}$$

Herewith

$$\psi_i + \sum_{j=1}^{N} \psi_{i,j} = 1; \tag{10}$$

$$\beta_i + \sum_{j=1}^{N} \beta_{i,j} = 1; \tag{11}$$

here $\delta_{i,j}$ – is Kronecker symbol.

It has been proved in [5], that traffic equations system solution satisfies generalized traffic equations system:

$$f_i \varepsilon_i = \sum_{k=1}^{N} \lambda_k \psi_{k,i} + \sum_{j=1}^{N} f_j \varepsilon_j \beta_{j,i}. \tag{12}$$

Intensities of transitions for Markov process $X(t)$ are

$$q(n, n + e_i) = \sum_{j=1}^{N} \lambda_j \psi_{j,i};$$

$$q(n, n - e_i) = \mu_i \beta_i I_{n_i > 0};$$
$$q(n, n + e_i - e_i') = \varphi_i I_{n_i' > 0};$$
$$q(n, n - e_i + e_i') = \nu_i I_{n_i > 0};$$
$$q(n, n - e_i + e_j) = \mu_i \beta_{i,j} I_{n_i > 0}.$$

It is obvious, that under conditions (2) Markov process $X(t)$ is ergodic, therefore unique stationary distribution $\pi(n)$, $n \in Z$ exists.

Global balance equations are:

$$\sum_{i \in J} \Big(\sum_{j \in J} \lambda_j \psi_{j,i} + \mu_i \beta_i I_{n_i > 0} + \varphi_i I_{n_i' > 0} + \nu_i I_{n_i > 0} + \sum_{j \in J} \mu_i \beta_{i,j} I_{n_i > 0} \Big) \pi(n) =$$

$$= \sum_{i \in J} \Big(\pi(n - e_i) \sum_{j=1}^{N} \lambda_j \psi_{i,j} I_{n_i > 0} + \pi(n + e_i) \mu_i \beta_i +$$

$$+ \pi(n - e_i + e_i') \varphi_i I_{n_i > 0} + \pi(n + e_i - e_i') \nu_i I_{n_i' > 0} +$$

$$+ \sum_{j \in J} \pi(n + e_i - e_j) \mu_i \beta_{i,j} I_{n_j > 0} \Big), \; n \in Z.$$

It is easy to show, that with foregoing intensities of transitions Markov process $X(t)$ is reversible.

Substituting $\pi(n)$, determined by means of (4), (5) into global balance equations, considering (6) - (11), traffic equation system (1) and generalized traffic equations system (12), we obtain identity. □

3 Conclusion

We have considered an open queueing network with temporarily non-active customers and rounds. Customers could partly loose their capacity for service. Customers could pass from non-active condition into condition, when they can get their service and vice versa. Conditions of ergodicity have been established. The form of stationary distribution and conditions of stationary distribution existence have been obtained.

References

1. Tsitsiashvili, G.S., Osipova, M.: Distributions in stochastic network models. Nova Publishers (2008)
2. Tsitsiashvili, G.S., Osipova, M.: Queueing models with different schemes of customers transformations. In: Proceedings of the 19th International Conference Mathematical Methods for Increasing Efficiency of Information Telecommunication Networks, pp. 128–133 (2007)
3. Gnedenko, B., Kovalenko, I.: Introduction to queueing theory, Moscow, Nauka (1987)
4. Malinkovsky, Y., Bojarovich, J.: An open queueing network with partly non-active customers. In: Proceedings of the 21st International Conference Modern Probbabilistic Methods for Analysis and Optimization of Information and Telecommunication Networks, pp. 34–37 (2011)
5. Malinkovsky, Y., Evdokimovich, V.: Queueing Networks with Dynamic Routing and Dynamic Stochastic Bypass of Nodes. Problems of Information Transmission 37, 236–247 (2001)

Analysis of $MAP/PH/c$ Retrial Queue with Phase Type Retrials – Simulation Approach

Srinivas R. Chakravarthy

Department of Industrial and Manufacturing Engineering
Kettering University, Flint, MI-48504, USA
schakrav@kettering.edu

Abstract. In this paper we study a multi-server retrial queueing model in which customers arrive according to a Markovian arrival process (MAP) and the service times are assumed to be of phase type (PH-type). An arriving customer finding all servers busy will enter into a (retrial) orbit of infinite size. The customers in orbit will try to capture a free server after a random amount of time which is assumed to be of PH-type. Thus, every customer in the orbit has his/her own phase type distribution before attempting to get into service. Due to the complexity of the model and lack of attention to such models in the literature, we study this via simulation. After validating our simulated results against known results (both exact and approximation) for some special cases, we illustrate how one can underestimate or overestimate some key system performance measures by incorrectly assuming the retrial times to be exponential.

1 Introduction and Model Description

Retrial queueing models play an important role in practice. The literature on retrial queues is extensive and covers a wide spectrum of models [2, 7]. With the exception of a few papers [1, 6, 8, 17–20] the models studied in the literature assume the retrial times to be exponential. The few papers in which non-exponential retrial times are assumed the authors propose a variety of approximations to compute selected system performance measures for very restrictive class of models such as single server or two-state phase type distribution or $M/M/c$ type queues or assume that not all customers in orbit attempt to capture a free server but only the customer at the head of the queue or discuss only the stability condition of the queue with no qualitative discussion on the role played by non-exponential retrial times. These are mainly due to the explosive nature of the state space that is required to keep track of the system which is not needed in the case of Poisson/exponential type retrial queues. Thus, there is a huge void in the literature on retrial queues and this paper is an attempt to fill this gap. The approach we take to address this issue is via simulation for two reasons. First, we want to get a feel for how some key system performance measures behave in multi-server retrial models with non-exponential retrials. Secondly, the simulated results may be helpful for any future study on such general retrial models with approxiation/truncation procedures. The results of our on-going research

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 37–49, 2013.

including the use of truncation/approximation procedures on this general retrial model using matrix-analytic methods will be reported elsewhere.

The retrial model under study in this paper is as follows. Customers arrive according to a Markovian arrival process (MAP) with representation (D_0, D_1) of order m to a multi-server queueing system. A brief description of MAP is given below. There are c homogeneous servers who offer services and the service times are assumed to be of phase type PH-type with representation $(\boldsymbol{\alpha}, T)$ of dimension n_1. Recall that a PH-distribution is obtained as the time until absorption in a finite state Markov chain with one absorption state. It is characterized by an initial probability vector ($\boldsymbol{\alpha}$) and a square matrix (T) governing the transitions to various transient states. PH-distributions are defined for both discrete and continuous time. For details on PH-distributions and their properties, we refer the reader to [13, 14, 16]. An arriving customer finding all servers busy will enter into an orbit of infinite size. These customers will independently try to capture a free server after a random amount of time that is assumed to be of phase type. Note that we assume that every customer will have his/her own phase type distribution that will be started when entering the retrial buffer as well as when the attempt to capture a free server is unsuccessful. The common PH-distribution for the retrial times has representation $(\boldsymbol{\beta}, S)$ of dimension n_2.

Now we will briefly describe the versatile point process introduced by Neuts [12]. A MAP is a tractable class of Markov renewal processes. It should be noted that by appropriately choosing the parameters of the MAP the underlying arrival process can be made as a renewal process. The MAP is a rich class of point processes that includes many well-known processes such as Poisson, PH-renewal processes, and Markov-modulated Poisson process. One of the most significant features of the MAP is the underlying Markovian structure and fits ideally in the context of matrix-analytic solutions to stochastic models. Matrix-analytic methods were first introduced and studied by Neuts [13]. The idea of the MAP is to significantly generalize the Poisson processes and still keep the tractability for modelling purposes. Furthermore, MAP is a convenient tool to model both renewal and non-renewal arrivals. While MAP is defined for both discrete and continuous times, here we will need only the continuous time case.

The MAP in continuous time is described as follows. Let the underlying Markov chain be irreducible and let Q^* be the generator of this Markov chain. At the end of a sojourn time in state i, that is exponentially distributed with parameter λ_i, one of the following two events could occur: with probability $p_{ij}^{(1)}$ the transition corresponds to an arrival and the underlying Markov chain is in state j with $1 \leq i, j \leq m$; with probability $p_{ij}^{(0)}$ the transition corresponds to no arrival and the state of the Markov chain is j, $j \neq i$. Note that the Markov chain can go from state i to state i only through an arrival. Define matrices $D_0 = (d_{ij}^{(0)})$ and $D_1 = (d_{ij}^{(1)})$ such that $d_{ii}^{(0)} = -\lambda_i$, $1 \leq i \leq m$, $d_{ij}^{(0)} = \lambda_i p_{ij}^{(0)}$, for $j \neq i$ and $d_{ij}^{(1)} = \lambda_i p_{ij}^{(1)}$, $1 \leq i, j \leq m$. By assuming D_0 to be a nonsingular matrix, the interarrival times will be finite with probability one and the arrival process does

not terminate. Hence, we see that D_0 is a stable matrix. The generator Q^* is then given by $Q^* = D_0 + D_1$.

Thus, D_0 governs the transitions corresponding to no arrival and D_1 governs those corresponding to an arrival. It can be shown that MAP is equivalent to Neuts' versatile Markovian point process. The point process described by the MAP is a special class of semi-Markov processes. For further details on MAP and their usefulness in stochastic modelling, we refer to ([10], [14], [15]) and for a review and recent work on MAP we refer the reader to ([3–5]).

Let $\boldsymbol{\eta}$ be the stationary probability vector of the Markov process with generator Q^*. That is, $\boldsymbol{\eta}$ is the unique (positive) probability vector satisfying $\boldsymbol{\eta}Q^* = \mathbf{0}, \boldsymbol{\eta}\boldsymbol{e} = 1$, where $\boldsymbol{e}$ is a column vector of 1's of appropriate dimension. We denote the average arrival rate, the average service rate, and the average retrial rate by, respectively, λ, μ, and θ. These are given by $\lambda = \boldsymbol{\eta}D_1\boldsymbol{e},\ \mu = [\boldsymbol{\alpha}(-T)^{-1}\boldsymbol{e}]^{-1},\ \theta = [\boldsymbol{\beta}(-S)^{-1}\boldsymbol{e}]^{-1}$.

The rest of the paper is organized as follows. In Section 2 we give a Markov process description of the model under study. Some key system performance measures used in this study are listed in Section 3. The simulated model is validated in Section 4. The roles of retrial distribution in the context of $M/PH/c$ and $MAP/PH/c$ with $PH-$ retrials are discussed in Sections 5 and 6, respectively. Some concluding remarks are given in Section 7.

2 Markov Process Description

The model outlined in Section 1 can be studied as a Markov process by keeping track of (a) the number, $K_i(t)$, of retrial customers waiting in phase $i, 1 \le i \le n_2$, at time t; (b) the number, $L_j(t)$, of servers busy serving in phase $j, 1 \le j \le n_1$, at time t; and (c) the phase, $J(t)$, of the arrival process at time t. Note that the retrial orbit being empty is indicated by taking $K_i(t) = 0, 1 \le i \le n_2$. The number of free servers at time t is given by $c - \sum_{j=1}^{n_1} L_j(t)$. The process $\{(K_1(t), \cdots, K_{n_2}(t), L_1(t), \cdots, L_{n_2}(t), J(t)) : t \ge 0\}$ is a continuous-time Markov chain with state space given by

$$\Omega = \{(k_1, \cdots, k_{n_2}, l_1, \cdots, l_{n_1}, r) : k_i \ge 0, 1 \le i \le n_2, l_j \ge 0, 1 \le j \le n_1,$$

$$0 \le l_1 + \cdots + l_{n_1} \le c, 1 \le r \le m\}.$$

The generator of this Markov process can be set up with the help of Kronecker products and sums of matrices [11]. However, it is clear that the steady-state analysis requires some form of approximation or truncation due to many (sub)states that grow without bound. The accuracy of the approximation or truncation depends on the degree to which these are carried out. Our focus in this paper is not in providing an approximation or truncation or a combination of both in performing the steady-state analysis. These are currently work-in-process and the results will be reported elsewhere. Instead, our goal is to see how the type of distributional assumption affects some selected system performance measures through simulation. Further, this simulated results can be used to compare any

approximation/truncation methods possibly proposed in the future. Thus, the rest of the paper is based on simulating the retrial model described in Section 1 with the help of ARENA [9].

3 Selected System Performance Measures

In this section we will list a number of key system performance measures for our illustration.

1. The probability, $PBLK$, that an arriving customer finds all servers to be busy.
2. The probability, $PESO$, that an arriving customer enters into service with at least one customer in the orbit.
3. The fraction, $FRSF$, of customers successfully capturing a free server at the time of a retry.
4. The mean, $MWTS$, waiting time in the system of a customer.
5. The mean, $MNIO$, number of customers in the orbit.
6. The mean, $MWTO$, waiting time in the orbit of a customer (given that a customer enters into the orbit). Note that $MWTS = \frac{1}{\mu} + (PBLK)(MWTO)$.

4 Validation of the Simulated Model

Before we proceed to discuss the simulated results, it is important to validate our simulated model by comparing our results with the published results in the literature. As pointed earlier only few papers deal with non-exponential retrial times for restricted class of models. Thus, we validate our model by comparing our simulated results with the ones for which results (either exact or approximations) are reported.

4.1 $M/M/5$ with PH_2 Retrial Times [18]

In [18], the author studied an M/M/c type retrial queueing model with two-state PH-distributions for the retrial times using level-dependent quasi-birth-and-death (QBD) process. Since the generator of the QBD is of infinite size with entries depending on the level (which is the number of customers in the orbit), the author truncated the generator so as to arrive at the steady-state results. The author reported the results for M/M/5 model with three two-state PH-distributions: (a) Erlang of order 2 labeled as E_2; (b) hyperexponential with squared coefficient variation (SCV set at 2; and (c) hyperexponential with SCV set at 10. Specifically, the hyperexponential considered has the mixing probabilities $(p, 1-p)$ and the rates in these two states are γ_1 and γ_2, where

$$p = 0.5[1 + \sqrt{(SCV-1)/(SCV+1)}], \quad \gamma_1 = 2\theta p, \quad \gamma_2 = 2\theta(1-p).$$

We reproduce a part Table 1 in [18]) with our notations in Table 1 below for comparison purposes. We analyzed the same model using our simulated model

and the results are summarized in Table 2. In Table 3 we display the error percentage which is calculated as (Shin approx - Simulated)/Shin approx.

By looking at these tables we notice that our simulated results are very close to the approximated results presented in [18].

Table 1. Shin's approximation (see Table 1 in [18])

		Retrial distribution					
		E_2		$HE_2(SCV=2)$		$HE_2(SCV=10)$	
ρ	θ^{-1}	$P(block)$	$E(NIO)$	$P(block)$	$E(NIO)$	$P(block)$	$E(NIO)$
0.3	0.1	0.0191	0.0108	0.0188	0.0121	0.0187	0.0150
	1	0.0163	0.0300	0.0166	0.0352	0.0169	0.0399
	5	0.0152	0.1179	0.0155	0.1275	0.0156	0.1327
	10	0.0150	0.2305	0.0152	0.2416	0.0153	0.2468
	20	0.0150	0.4574	0.0151	0.4695	0.0151	0.4747
0.5	0.1	0.1218	0.1561	0.1196	0.1691	0.1190	0.2018
	1	0.1003	0.3644	0.1016	0.4221	0.1045	0.4895
	5	0.0910	1.3010	0.0928	1.4150	0.0941	1.4920
	10	0.0897	2.5020	0.0909	2.6340	0.0917	2.7130
	20	0.0891	4.9230	0.0898	5.0690	0.0903	5.1480
0.8	0.1	0.5248	2.5190	0.5192	2.6190	0.5171	2.8730
	1	0.4545	5.0030	0.4572	5.4660	0.4646	6.2040
	5	0.4245	16.1500	0.4276	17.0000	0.4315	17.9900
	10	0.4198	30.2700	0.4217	31.2200	0.4241	32.2700
	20	0.4174	58.6100	0.4185	59.6200	0.4198	60.7000

Table 2. Simulated results for $M/M/5$

		Retrial distribution					
		E_2		$HE_2(SCV=2)$		$HE_2(SCV=10)$	
ρ	θ^{-1}	$P(block)$	$E(NIO)$	$P(block)$	$E(NIO)$	$P(block)$	$E(NIO)$
0.3	0.1	0.0191	0.0108	0.0187	0.0120	0.0189	0.0150
	1	0.0163	0.0296	0.0165	0.0345	0.0168	0.0389
	5	0.0153	0.1193	0.0154	0.1288	0.0154	0.1307
	10	0.0150	0.2286	0.0152	0.2405	0.0153	0.2473
	20	0.0148	0.4521	0.0152	0.4729	0.0151	0.4707
0.5	0.1	0.1221	0.1562	0.1202	0.1713	0.1187	0.2011
	1	0.1003	0.3623	0.1020	0.4230	0.1045	0.4895
	5	0.0906	1.2937	0.0932	1.4224	0.0943	1.5015
	10	0.0894	2.4933	0.0912	2.6391	0.0924	2.7602
	20	0.0892	4.9187	0.0898	5.0671	0.0908	5.1584
0.8	0.1	0.5239	2.5045	0.5217	2.6267	0.5180	2.8862
	1	0.4526	4.9391	0.4575	5.4802	0.4630	6.1722
	5	0.4225	15.9597	0.4282	16.9751	0.4323	18.0360
	10	0.4191	30.2014	0.4237	31.4826	0.4245	32.3222
	20	0.4164	58.3282	0.4173	59.2781	0.4189	60.4343

Table 3. Error percentage for $M/M/5$

		Retrial distribution					
		E_2		$HE_2(SCV=2)$		$HE_2(SCV=10)$	
ρ	θ^{-1}	$P(block)$	$E(NIO)$	$P(block)$	$E(NIO)$	$P(block)$	$E(NIO)$
0.3	0.1	0.00%	0.00%	0.53%	0.83%	-1.07%	0.00%
	1	0.00%	1.33%	0.60%	1.99%	0.59%	2.51%
	5	-0.66%	-1.19%	0.65%	-1.02%	1.28%	1.51%
	10	0.00%	0.82%	0.00%	0.46%	0.00%	-0.20%
	20	1.33%	1.16%	-0.66%	-0.72%	0.00%	0.84%
0.5	0.1	-0.25%	-0.06%	-0.50%	-1.30%	0.25%	0.35%
	1	0.00%	0.58%	-0.39%	-0.21%	0.00%	0.00%
	5	0.44%	0.56%	-0.43%	-0.52%	-0.21%	-0.64%
	10	0.33%	0.35%	-0.33%	-0.19%	-0.76%	-1.74%
	20	-0.11%	0.09%	0.00%	0.04%	-0.55%	-0.20%
0.8	0.1	0.17%	0.58%	-0.48%	-0.29%	-0.17%	-0.46%
	1	0.42%	1.28%	-0.07%	-0.26%	0.34%	0.51%
	5	0.47%	1.18%	-0.14%	0.15%	-0.19%	-0.26%
	10	0.17%	0.23%	-0.47%	-0.84%	-0.09%	-0.16%
	20	0.24%	0.48%	0.29%	0.57%	0.21%	0.44%

4.2 $M/M/5$ with Mixture of Erlang Retrial Times [19]

In [19], the authors studied $M/M/c$ retrial queues with PH-distribution for the retrial times through approximating the steady-state probabilities by assuming that the service facility behaving like a birth-and-death process in which the rates are independent of the retrials. They compare their approximations to the special cases considered in [18] and also to the simulated results for their specialized models. They look at four retrial distributions: Erlang of order 4 (E_4), mixture of two Erlangs (MER_3), and two distributions that are mixtures of generalized Erlang and Erlang ($CE_{3,1}$). All these are four distributions are normalized so as to have the same mean retrial times. We refer the reader to the paper [19] for the form of these distributions. We reproduce a part of that table in Table 4 below for comparison purposes. We analyzed the same model using our simulated model and the results are summarized in Table 4. Our simulated results are given within parentheses in Table 4. In Table 5 we display the error percentage which is calculated as [(SM approx - Simulated)/SM approx].

It should be pointed out that the results reported in Shin and Moon are approximations and as mentioned in [19], the approximation for the case when the retrial time distribution is given by $CE_{3,1}(0.185487)$ is worse. Thus, we notice a higher percentage for the error for that case. This could be due to the approximation rather than due to our simulated results.

Table 4. Shin's approximation and simulated results for $E(NIO)$ for $M/M/5$

		Retrial distribution			
ρ	θ^{-1}	E_4	MER_3	$CE_{3,1}(0.007773)$	$CE_{3,1}(0.185487)$
0.3	10.0	0.1598 (0.1526)	0.1629 (0.154)	0.1811 (0.1728)	0.2328 (0.22)
	1.0	0.3634 (0.3519)	0.373 (0.3627)	0.4296 (0.4229)	0.7133 (0.7171)
	0.2	1.294 (1.2975)	1.304 (1.2938)	1.413 (1.3987)	1.96 (2.0572)
	0.1	2.501 (2.4976)	2.507 (2.4882)	2.631 (2.6215)	3.267 (3.4217)
	0.1	4.931 (4.9072)	4.933 (4.9247)	5.064 (5.0522)	5.762 (5.9654)
0.5	10.0	2.594 (2.5062)	2.622 (2.5133)	2.785 (2.6991)	3.317 (3.1043)
	1.0	4.842 (4.9176)	4.922 (4.9549)	5.439 (5.525)	8.1 (8.1949)
	0.2	15.98 (16.0332)	16.01 (16.0437)	16.79 (17.0079)	21 (23.0209)
	0.1	30.15 (30.2656)	30.16 (30.2923)	30.97 (31.2837)	35.47 (38.1137)
	0.1	58.52 (58.555)	58.53 (58.4076)	59.35 (59.1616)	63.99 (67.3217)

Table 5. Error percentage for $E(NIO)$ for $M/M/5$

		Retrial distribution			
ρ	θ^{-1}	E_4	MER_3	$CE_{3,1}(0.007773)$	$CE_{3,1}(0.185487)$
0.3	10.0	4.51%	5.46%	4.58%	5.50%
	1.0	3.16%	2.76%	1.56%	-0.53%
	0.2	-0.27%	0.78%	1.01%	-4.96%
	0.1	0.14%	0.75%	0.36%	-4.74%
	0.1	0.48%	0.17%	0.23%	-3.53%
0.5	10.0	3.38%	4.15%	3.08%	6.41%
	1.0	-1.56%	-0.67%	-1.58%	-1.17%
	0.2	-0.33%	-0.21%	-1.30%	-9.62%
	0.1	-0.38%	-0.44%	-1.01%	-7.45%
	0.1	-0.06%	0.21%	0.32%	-5.21%

5 Role of Retrial Distribution in $M/PH/c$ with PH Retrial Times

In this section we will look at the role of the retrial time distribution on the selected system performance measures under different scenarios for $M/PH/c$ type queueing model with PH retrial times. The reason for looking at this model first before studying the most general one is to highlight how some system performance measures are sensitive to the type of retrial and or service time assumed even for this simplest retrial model. For all cases considered we fix the arrival and retrial rates to be 1 (i.e., $\lambda = \theta = 1$) and vary other parameters as illustrated in the discussions below. We simulated the model using 10 replications and for 100,000 units (which in our case is minutes) for each replicate.

First we look at the case of exponential services since $M/M/c$ type retrial models have been very widely studied in the literature. We vary c from 1 to 4 and vary μ so as to achieve a given value for $\rho = \frac{\lambda}{c\mu}$. Specifically we consider

$\rho = 0.1, 0.2, 0.3, 0.5$, and 0.9. For retrial times we look at the following three PH-distributions: (a) Erlang of order 5 ($ERLR$); (b) Exponential ($EXPR$); and (c) hyperexponential ($HEXR$) with mixing probabilities 0.9 and 0.1 with rates, respectively, 10 and 1/9.1. The graphs of the measures: (i) FRSF; (ii) MWTO; (iii) MWTS; and (iv) PESO are plotted in Figure 1 for various scenarios. It should be pointed out that after conducting a statistical analysis on the output based on 10 replications, the measures that are found to be significant are plotted. Thus, for example, we did not plot the measure, PBLK, since we did not find any significant (at 5% level) differences among the various retrial distributions used with regard to this measure.

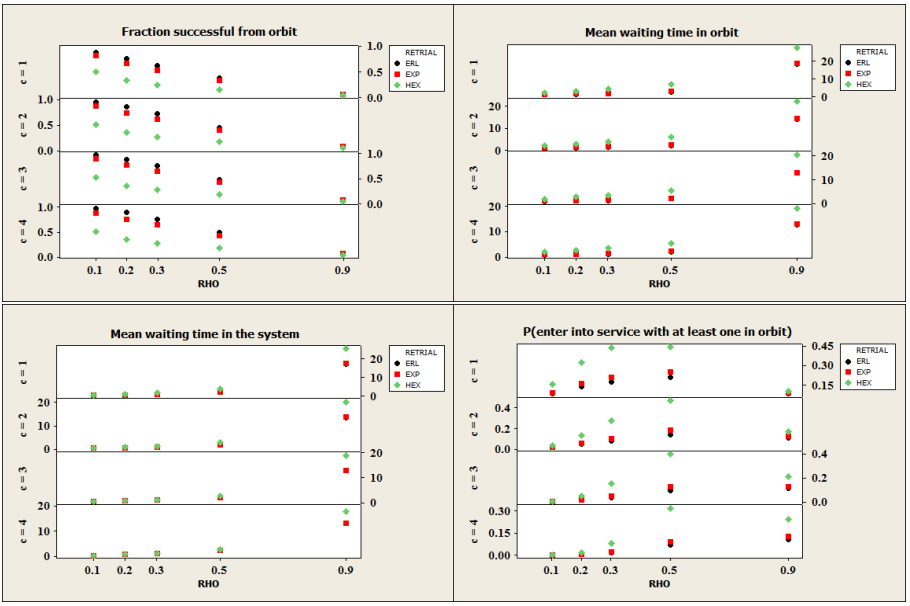


Fig. 1. System measures under various scenarios

A quick look at Figure 1, one will notice the significant role played by the type of retrial distribution used. The significance of the role of retrial distribution depends not only the type of system measure but also on the traffic intensity. For example, in the case of the fraction of orbiting customers successfully reaching an idle server, the significance of the type of retrial time distribution is seen in the case of low to moderate values of ρ. However, when looking at the mean waiting time in the system or the mean waiting time in orbit, the significance of the type of retrial time distribution is noticed for reasonably large values of ρ. Thus, one cannot make a statement that the significance of the type of retrial distribution is only in the case of low to moderate or only for large values of ρ. In conclusion, assuming exponential retrial times (when in practice it is not the case) will lead to incorrect decisions.

We noticed a similar behavior with regard to Erlang and hyperexponential services and due to space restriction we did not display the graphs here. However, it should be pointed out that these measures vary significantly when the retrial times are varied from Erlang to exponential to hyperexponential. This indicates that both the variability in the services as well as the retrial times play a significant role and should not be overlooked.

6 Role of Retrial Distribution in $MAP/PH/c$ with PH Retrial Times

In this section we will continue our discussion of the retrial model under study with a primary focus on the role of the retrial time variability on selected system performance measures under different scenarios for $MAP/PH/c$ type queueing model with PH retrial times. For the arrival process, we consider the following five sets of values for D_0 and D_1.

1. Erlang ($ERLA$):

$$D_0 = \begin{pmatrix} -5 & 5 & 0 & 0 & 0 \\ 0 & -5 & 5 & 0 & 0 \\ 0 & 0 & -5 & 5 & 0 \\ 0 & 0 & 0 & -5 & 5 \\ 0 & 0 & 0 & 0 & -5 \end{pmatrix}, D_1 = \begin{pmatrix} 0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0 \\ 5\,0\,0\,0\,0 \end{pmatrix}$$

2. Exponential ($EXPA$):

$$D_0 = \begin{pmatrix} -1 \end{pmatrix}, D_1 = \begin{pmatrix} 1 \end{pmatrix}$$

3. Hyperexponential ($HEXA$):

$$D_0 = \begin{pmatrix} -10 & 0 \\ 0 & -\frac{10}{91} \end{pmatrix}, D_1 = \begin{pmatrix} 9 & 1 \\ \frac{9}{91} & \frac{1}{91} \end{pmatrix}$$

4. MAP with negative correlation ($MNCA$):

$$D_0 = \begin{pmatrix} -1.1 & 1.1 & 0 \\ 0 & -1.1 & 0 \\ 0 & 0 & -5.5 \end{pmatrix}, D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0.055\,0 & 1.045 \\ 5.225\,0 & 0.275 \end{pmatrix}$$

5. MAP with positive correlation ($MPCA$):

$$D_0 = \begin{pmatrix} -1.1 & 1.1 & 0 \\ 0 & -1.1 & 0 \\ 0 & 0 & -5.5 \end{pmatrix}, D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 1.045\,0 & 0.055 \\ 0.275\,0 & 5.225 \end{pmatrix}$$

All these five MAP processes are normalized so as to have a specified arrival rate. However, these are qualitatively different in that they have different variance and

correlation structure. The first three arrival processes, namely $ERLA$, $EXPA$, and $HEXA$, correspond to renewal processes and so the correlation is 0. The arrival process labeled $MNCA$ has correlated arrivals with correlation between two successive inter-arrival times given by -0.3984 and the arrivals corresponding to the processes labelled $MPCA$ has a positive correlation with values 0.3984. The ratio of the standard deviations of the inter-arrival times of these five arrival processes with respect to $ERLA$ are, respectively, 1, 2.2361, 8.8261, 2.7499, and 2.7499.

For the service times $((\boldsymbol{\alpha}, T))$ as well as for retrial times $((\boldsymbol{\beta}, S))$, we consider the following three $PH-$distributions.

$$\textbf{A. Erlang (ERLS/ERLR)}: \ \boldsymbol{\alpha} = \boldsymbol{\beta} = (1, 0), \ \ T = S = \begin{pmatrix} -2 & 2 \\ 0 & -2 \end{pmatrix}.$$

$$\textbf{B. Exponential (EXPS/EXPR)}: \ \boldsymbol{\alpha} = \boldsymbol{\beta} = 1, \ \ T = S = \begin{pmatrix} -1 \end{pmatrix}.$$

$$\textbf{C. Hyperexponential (HEXS/HEXR)}: \ \boldsymbol{\alpha} = \boldsymbol{\beta} = (0.9, 0.1), \ \ T = S = \begin{pmatrix} -10 & 0 \\ 0 & -d \end{pmatrix}.$$

For all cases considered we again fix the arrival and retrial rates to be 1 (i.e., $\lambda = \theta = 1$) and vary other parameters as illustrated in the discussions below, and accordingly the PH-representations will be normalized except in the case of hyperexponential in which case d will be chosen to have the specific mean. We simulated the model using 10 replications and for 100,000 units (which in our case is minutes) for each replicate.

In Table 6, we display the significance (at 5% level based on the analysis of variance) of each of the measures. The notations used in the table are as follows: the symbol "S" for the significance of the role of the service times and "R" for the significance of the role of the retrial times. From this table it is obvious how the system performance measures for various combinations of the arrival process, service times, c, and ρ, are sensitive to the type of retrial distribution assumed. For example, in Figure 2, we display under various scenarios, the ratios of the five measures: $PBLK, PESO, FRSF, MTWS$, and $MWTO$. These ratios are calculated as follows. Since in the literature exponential distribution is the most commonly assumed one for retrial times, we use this as the base and compute the ratios, labeled $R1$ and $R2$, of the selected measures when the retrial times are Erlang and hyperexponential, respectively. For example, when a particular measure is computed for the model $MAP/PH/c$ with Erlang retrials and say this measure has a value a and for the corresponding retrial model but with exponential retrials the measure has a value b, the ratio $\frac{a}{b}$ will have a label $R1$. Note that whenever these ratios are closer to 1, using exponential retrials instead of Erlang or hyperexponential will have no adverse effect. However, it is when these ratios are far away from 1, then one needs to use nonexponential retrial times. From Figure 2 we notice that the ratio for $PBLK$ is close to 1 for all the scenarios under consideration; however, in the case of other measures the ratios are far away from 1 under many different scenarios.

As mentioned earlier, one should not ignore the variability in the services as well as the retrial times when using such retrial models in practice.

Table 6. Significance based on ANOVA for $MAP/PH/1$

		$c=1$			$c=2$			$c=3$			$c=4$		
		ρ			ρ			ρ			ρ		
Measures	MAP	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
PBLK	ERLA	S	S			S	S,R	S	S	S,R	S	S	S,R
	EXPA				S,R	S	S	S,R	S	S		S	S
	HEXA	S,R	S,R	S,R	S,R	S,R	S,R	S,R	S,R	S,R	S,R	S,R	S
	MNCA	S,R	S	S	S	S		S	S	S		S	S,R
	MPCA	R	S,R		S,R	S,R	S	S,R	S		S,R	S	S
PNDW1	ERLA	S	R	S,R		S,R			R	S,R	S	R	S,R
	EXPA	R	S,R		R	S,R	S,R	R	S,R	S,R	R	R	S,R
	HEXA	R	S	S		S,R	S,R	S,R	S,R	S,R	S,R	S,R	S,R
	MNCA	R	R		R	S,R	S,R	R	R	S,R		R	S,R
	MPCA	R	R	R	S,R	S,R	S,R	S,R	R	S,R	S,R	R	R
FRSF	ERLA	R	S,R	S	S	S	S	S	S	S	S	S	S
	EXPA	R	S,R	S	S,R	S	S	S,R	S	S	S,R	S	S
	HEXA	S,R	S,R	S,R	S,R	S,R	S	S,R	S,R	S	S,R	S,R	S
	MNCA	R	S,R	S	S,R	S	S	S,R	S	S	S,R	S	S
	MPCA	R	S,R	S	S,R	S,R	S	S,R	S,R	S	S,R	S,R	S
MWTS	ERLA	S	S	S,R		S	S,R		S	S,R	S	S	S
	EXPA	R	S,R	S,R	R	S,R	S,R	R	S,R	S		S,R	S
	HEXA	R	S,R	S	S,R	S,R	S,R	S,R	R	S	S,R	R	S
	MNCA	R	S,R	S,R	R	S,R	S,R		S,R	S,R		S,R	S,R
	MPCA	R	S,R		R	S,R	S	R	S,R	S	S,R	S,R	S
MNIO	ERLA	S	S	S,R		S	S,R	S	S	S,R	S	S	S
	EXPA	R	S,R	S,R	R	S,R	S,R	R	S,R	S	R	S,R	S
	HEXA	R	S,R	S	S,R	S,R	S,R	S,R	R	S	S,R	S	S
	MNCA	R	S,R	S,R	R	S,R	S,R		S,R	S,R		S,R	S,R
	MPCA	R	S,R	S	R	S,R	S	S,R	S,R	S	S,R	S,R	S
MWTO	ERLA		S	S,R		S	S,R		S,R	S,R	S	S,R	S,R
	EXPA	R	S,R	S,R	R	S,R	S,R	R	S,R	S	R	S,R	S
	HEXA	R	S,R	S	S,R	S,R	S,R	S,R	S,R	S	S	S,R	S
	MNCA	R	S,R	S,R	R	S,R	S,R	R	S,R	S,R	R	S,R	S,R
	MPCA	R	S,R	S	R	S,R	S	S,R	S,R	S	S,R	S,R	S

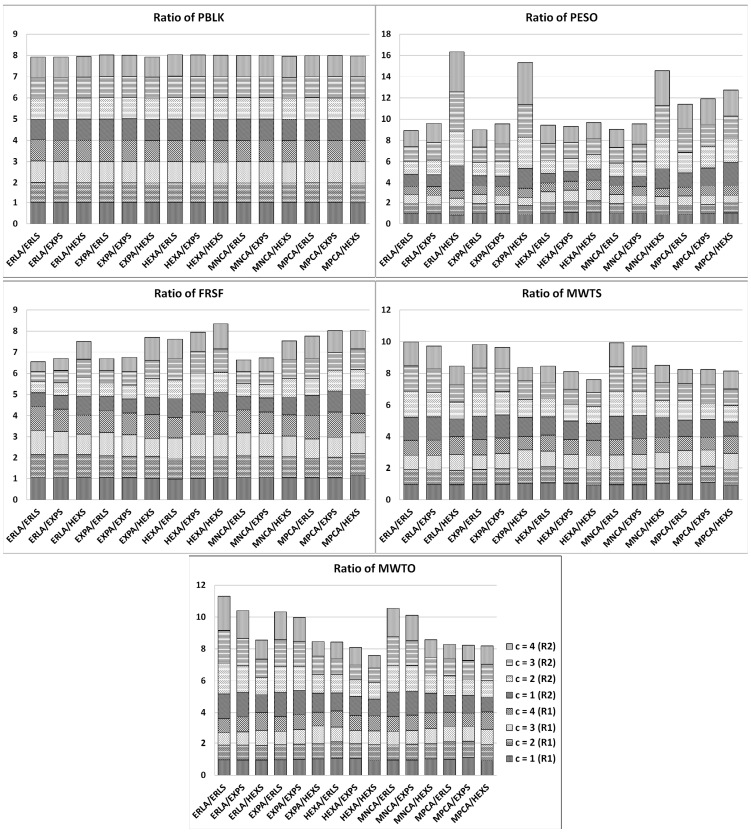


Fig. 2. Ratios of the measures: R1: ERLR/EXPR; R2: HEXR/ERLR

7 Concluding Remarks

In this paper we analyzed a multi-server retrial queueing system in which customers arrive according to a Markovian arrival process. Assuming the service and retrial times to be of phase type, we investigated the impact of the type of distribution assumed for the retrials through simulation as most work in the literature concentrate on exponential retrial times. We showed how one can underestimate or overestimate key system measures by incorrectly assuming the retrial times to be exponential. We hope that this study will help researchers pursuing the study of this class models using some kind of approxiation/truncation to compare their results with the simulated ones. The results of our on-going research using matrix-analytic methods will be reported elsewhere.

References

1. Artalejo, J.R., Gomez-Corral, A.: Modelling communication systems with phase type service and retrial times. IEEE Communications Letters 11, 955–957 (2007)
2. Artalejo, J.R., Gomez-Corral, A.: Retrial Queueing Systems, A Computational Approach. Springer, Heidelberg (2008)
3. Artalejo, J.R., Gomez-Correl, A., He, Q.M.: Markovian arrivals in stochastic modelling: a survey and some new results. SORT 34(2), 101–144 (2010)
4. Chakravarthy, S.R.: The batch Markovian arrival process: A review and future work. In: Krishnamoorthy, A., et al. (eds.) Advances in Probability Theory and Stochastic Processes, pp. 21–39. Notable Publications Inc., NJ (2001)
5. Chakravarthy, S.R.: Markovian Arrival Processes. Wiley Encyclopedia of Operations Research and Management Science (June 15, 2010)
6. Diamond, J.E., Alfa, A.S.: Approximation method for M/PH/1 retrial queues with phase type inter-retrial times. European Journal of Operational Research 113, 620–631 (1999)
7. Falin, G.I., Templeton, J.G.C.: Retrial Queues. Chapman and Hall, London (1997)
8. He, Q.M., Li, H., Zhao, Y.Q.: Ergodicity of the $BMAP/PH/s/s+K$ retrial queue with PH-retrial times. Queueing Systems 35, 323–347 (2000)
9. Kelton, W.D., Sadowski, R.P., Swets, N.B.: Simulation with ARENA, 5th edn. McGraw-Hill, New York (2010)
10. Lucantoni, D.M.: New results on the single server queue with a batch Markovian arrival process. Stochastic Models 7, 1–46 (1991)
11. Marcus, M., Minc, H.: A Survey of Matrix Theory and Matrix Inequalities. Allyn and Bacon, Boston (1964)
12. Neuts, M.F.: A versatile Markovian point process. J. Appl. Prob. 16, 764–779 (1979)
13. Neuts, M.F.: Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. The Johns Hopkins University Press, Baltimore (1994 version is Dover Edition)
14. Neuts, M.F.: Structured Stochastic Matrices of $M/G/1$ type and their Applications. Marcel Dekker, NY (1989)
15. Neuts, M.F.: Models based on the Markovian arrival process. IEICE Transactions on Communications E75B, 1255–1265 (1992)
16. Neuts, M.F.: Algorithmic Probability: A collection of problems. Chapman and Hall, NY (1995)
17. Senthilkumar, M., Sohraby, K., Kim, K.: On a multiserver retrial queue with phase type retrial time. In: Nadarajan, R., et al. (eds.) Mathematical and Computational Models, pp. 65–78. Narosa Publishing House, New Delhi (2012)
18. Shin, Y.W.: Algorithmic solutions for $M/M/c$ retrial queue with PH2 retrial time. Journal of Applied Mathematics and Informatics 29, 803–811 (2011)
19. Shin, Y.W., Moon, D.H.: Approximation of $M/M/c$ retrial queue with PH-retrial times. European Journal of Operational Research 213, 205–209 (2011)
20. Yang, T., Posner, M.J.M., Templeton, J.G.C., Li, H.: An approximation method for the M/G/1 retrial queues with general retrial times. European Journal of Operational Research 76, 552–562 (1994)

Fluid Flow Analysis of RED Algorithm with Modified Weighted Moving Average

Joanna Domańska[2], Adam Domański[1], and Tadeusz Czachórski[2]

[1] Institute of Informatics
Silesian Technical University
Akademicka 16, 44–100 Gliwice, Poland
adamd@polsl.pl
[2] Institute of Theoretical and Applied Informatics
Polish Academy of Sciences
Baltycka 5, 44–100 Gliwice, Poland
{joanna,tadek}@iitis.gliwice.pl

Abstract. We study with the use of fluid flow approximation the impact of a modified weighted moving average on the performance of RED mechanism. A model of TCP/UDP connection with RED implemented in an intermediate IP router is used, the weighted moving average is determined on the basis of a difference (recursive) equation. The fluid flow approximation technique is applied to model the interactions between the set of TCP/UDP flows and RED mechanism.

1 Introduction

The rapid growth of the Internet has imposed many new challenges, one of them is the necessity to control continuously growing traffic which has a complex statistical nature. Delays and congestions are supervised by TCP protocol, however, more advanced congestion control mechanisms, such as RED – Random Early Detection (or Drop) [9] or other Active Queue Management (AQM) schemes, are widely used.

The RED mechanism is conceptually very simple but interaction between RED and TCP is rather complex [12]. A lot of analytical models of RED in IP routers was already studied in open-loop scenario [7], [8]. In this paper we try to analyse the RED/TCP interaction with the use of fluid flow modelling methodology based on mean value analysis [11], [16], [19]. The model enables not only the steady-state analysis, but also allows us to obtain the transient behaviour when a set of TCP flows start or end transmission. This nonlinear dynamic model of TCP was conceived to analyse and understand various network congestion scenarios [16]. It was shown that it is able to reproduce the dynamics of TCP flows [21].

Several extensions of classical RED mechanism were investigated in the literature and some of them were based on modifications of the drop probability function [22], [1], [2]. The authors already investigated the influence of weighted moving average on the performance of the RED mechanism in open loop scenario

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 50–58, 2013.

[3] and proposed there an approach based on estimation of weighted moving average by high order difference equations. Here we consider a closed-loop model and use the fluid flow approximation to analyse the influence of our RED modification in the TCP/UDP environment.

The rest of this article is organized as follows. The section 2 describes the fluid flow model of router supporting RED queue management with TCP/UDP flows. The section 3 displays the obtained results and is followed by conclusions in section 4.

2 The Fluid Flow Model of AQM Router

This section presents a fluid flow model of the AQM router supporting TCP/UDP flows. The model presented in [16] demonstrates TCP protocol dynamics and allows to obtain the average value of key network variables. The model is described by the following nonlinear differential equations [10]:

$$W'(t) = \frac{1}{R(t)} - \frac{W(t)W(t-R(t))}{2R(t-R(t))}p(t-R(t)) \tag{1}$$

$$q'(t) = \frac{W(t)}{R(t)}N(t) - C \tag{2}$$

where:

W – expected TCP sending window size (packets),
q – expected queue length (packets),
R – round-trip time $= q/C + T_p$ (secs),
C – link capacity (packets/sec),
T_p – propagation delay (secs),
N – number of TCP sessions,
p – packet drop probability.

The maximum values of q and W (queue length and congestion window size) depend on the buffer capacity and maximum window size. The dropping probability p depends on the queue algorithm.

The traffic composed of TCP and UDP streams was considered in [20]. In this model a single router supports N sessions and each session is assumed to be either a TCP or UDP session. Each TCP stream is a TCP-Reno connection and each UDP sender is a CBR source. The rate of UDP sessions is denoted by λ. Fluid-flow equations of TCP and UDP mixed traffic become:

$$W'(t) = \frac{1}{R'(t)} - \frac{W(t)W(t-R(t))}{2R'(t-R(t))}p(t-R'(t)) \tag{3}$$

$$q'(t) = \frac{W(t)}{R(t)}N_\gamma(t) - (C-\lambda) \tag{4}$$

where R' = round-trip time $= q/(C-\lambda) + T_p$ (secs)

The RED algorithm was proposed by IETF to improve the transmission through IP routers. It was first described by Sally Floyd and Van Jacobson in [9]. The idea of RED mechanism is based on a drop function giving probability that a packet is rejected. The argument *avg* of this function is a weighted moving average queue length working as a low-pass filter and calculated at the arrival of each packet using the following formula:

$$x_i = (1-\alpha)x_{i-1} + \alpha q_{inst} \tag{5}$$

where q_{inst} is the current queue length and α is a parameter.

Our approach to determine the weighted moving average queue is based on a difference equation (a recursive equation).

Let $A(n)$ denote the weighted moving average length at the n-th moment of time and may be expressed using the difference equation as follows:

$$\begin{aligned} A(n) = a_1 A(n-1) + a_2 A(n-2) + \ldots + \\ + a_k A(n-k) + b_0 Q(n) + \\ + b_1 Q(n-1) + \ldots + b_m Q(n-m) \end{aligned} \tag{6}$$

where $a_j = const$ for $j = 1, \ldots, k$, $b_i = const$ for $i = 0, \ldots, m$, $A(l)$ is the weighted moving average queue length at the l-th moment of time, $Q(l)$ is the current length of the packet queue at the l-th moment.

Constraint conditions for a_j and b_i coefficients are:

$$\sum_{j=1}^{k} a_j + \sum_{i=0}^{m} b_i = 1 \wedge a_j \geq 0 \wedge b_i \geq 0. \tag{7}$$

The classical RED approach (were the weighted moving average queue length is given by eq. 5) satisfies the equation of the model given by eq. 1 when only a_1 and b_0 are significant coefficients. Only one parameter (a_1) should be determined in the classical RED approach (because $b_0 = 1 - a_1$). This was named 1-dimesional model, $[a_1]$, $(k = 1, m = 0)$.

We propose to take into account 4 significant parameters (a_1, a_2, b_0, b_1), so we consider 3-dimensional model $[a_1, a_2, b_1]$ $(k = 2, m = 1)$ were $b_0 = 1 - a_1 - a_2 - b_1$. Based on (1) we can calculate the weighted moving average queue length as:

$$\begin{aligned} A(n) = a_1 A(n-1) + a_2 A(n-2) + \\ + (1 - a_1 - a_2 - b_1) Q(n) + b_1 Q(n-1) \end{aligned} \tag{8}$$

In particular, for selected values of a_2 and b_1 the proposed model $[a_1, a_2, b_1] = [a_1, 0, 0]$ becomes the classical RED model, i.e. $[a1]$.

3 Results

During the tests we assumed the following parameters for AQM buffer:

- $Min_{th} = 10$,
- $Max_{th} = 15$,
- buffer size (measured in packets) = 20,

Table 1. Mean queue length

Type of flows	Number of flows	weighted moving average	Mean queue length
TCP	1	normal	8.335941
TCP	1	modified ver. 1	8.828522
TCP	1	modified ver. 2	9.771673
TCP	2	normal	8.333797
TCP	2	modified ver. 1	7.968082
TCP	2	modified ver. 2	9.557045
TCP	10	normal	10.298166
TCP	10	modified ver. 1	10.210228
TCP	10	modified ver. 2	10.075144
TCP+UDP	1	normal	8.880204
TCP+UDP	1	modified ver. 1	9.040307
TCP+UDP	1	modified ver. 2	9.969054
TCP+UDP	2	normal	9.184550
TCP+UDP	2	modified ver. 1	8.650753
TCP+UDP	2	modified ver. 2	9.527762
TCP+UDP	10	normal	10.411907
TCP+UDP	10	modified ver. 1	10.410298
TCP+UDP	10	modified ver. 2	10.233717

and the parameters of TCP connection:

- transmission capacity of AQM router: $C = 0.075$,
- propagation delay for i-th flow: $T_{p_i} = 2$,
- initial congestion window size for i-th flow (measured in packets): $W_i = 1$.

All computations were made with the use of PyLab (Python numeric computation environment) [18] – a combination of Python, NumPy, SciPy, Matplotlib, and IPython. Table 1 presents overall results: the weighted moving average column presents the kind of function used during calculation of the moving average queue length. The *normal* position presents the results of the standard RED function, $a1 = 0.007$, $a2 = 0.0$, $b1 = 0.0$. The *modified* position denotes our approach; we used two sets of parameters [3], ,,ver. 1": $a1 = 0.004$, $a2 = 0.008$, $b1 = 0.0$ and ,,ver. 2": $a1 = 0.08$, $a2 = 0.0014$, $b1 = 0.001$.

The curves in figures present transient system behaviour, the time axis is drawn in seconds. Figures 1 (a), (b) and (c) present one TCP flow. The use of modified weighted moving average, despite the fact that increases the mean queue length (which seems to be adverse), results in smaller changes of TCP window and thus smoother transmissions.

The situation is similar in figure 2. A thing to notice is that the queue fills up faster, resulting in a much earlier reduction of congestion window. For the case of 10 TCP flows (figures 3 (a), (b) and (c)), a stable state of the network is reached after the the initial large number of losses. For all TCP flows the congestion window oscillates around the minimum value.

When we look at the figures 4 (a), (b) and (c) we can see that the introduction of UDP traffic increases the mean queue length for all three solutions. The introduction of UDP traffic increases also the fluctuation of the TCP window. It can be seen clearly when we compare figures 1 (a) and 4 (a). The use of modified weighted moving average causes less fluctuation of TCP window than for the classical RED. In this case one can see an advantage of the solution with the parameters "ver. 2" (mean queue length increases the least). For a large number of TCP+UDP flows (figures 5 (a), (b) and (c)) - the solution with the parameters "ver. 2" gives the best results (the last column of table 1). This is in accordance with the results obtained in [3]. In that study we examined the traffic excluding the impact of TCP and this solution was also the best.

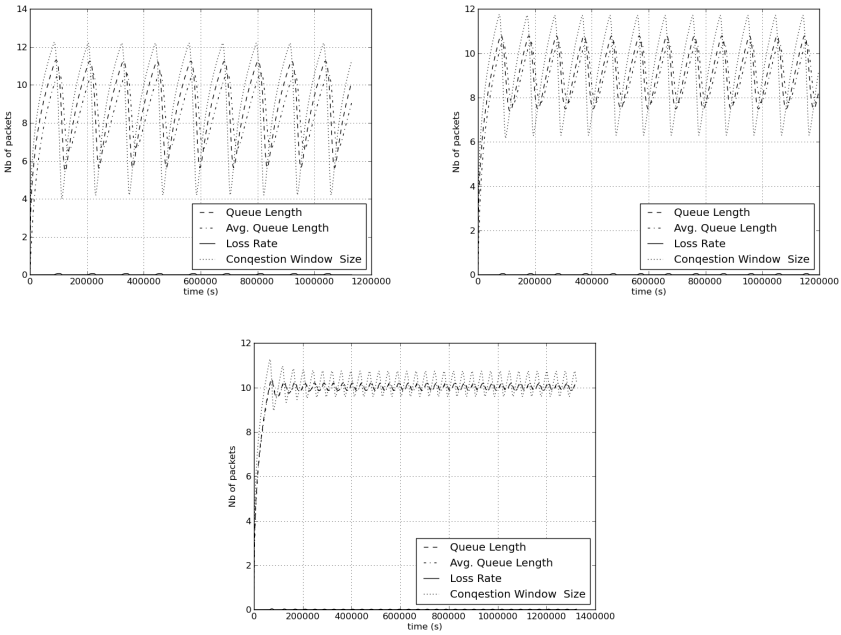


Fig. 1. RED queue, one TCP flow: (a) standard weighted moving average ($a1 = 0.007$, $a2 = 0.0$, $b1 = 0.0$), (b) modified weighted moving average ($a1 = 0.004$, $a2 = 0.008$, $b1 = 0.0$), (c) modified weighted moving average ($a1 = 0.08$, $a2 = 0.0014$, $b1 = 0.001$)

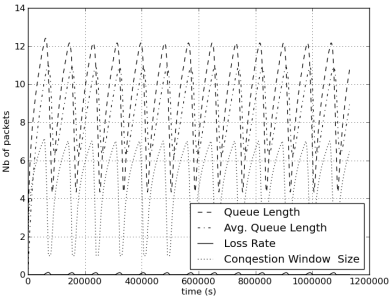


Fig. 2. RED queue, two TCP flows - standard weighted moving average ($a1 = 0.007$, $a2 = 0.0$, $b1 = 0.0$)

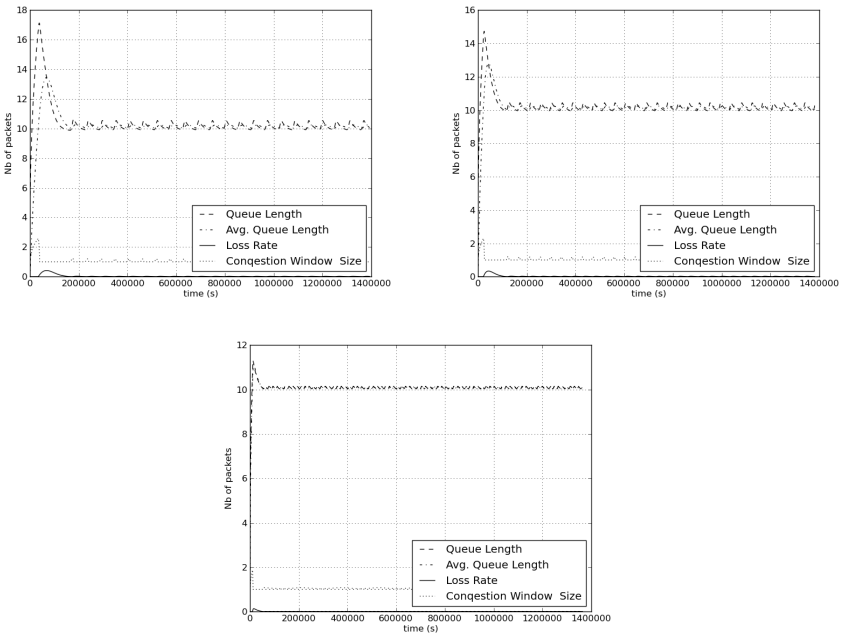


Fig. 3. RED queue, ten TCP flows: (a) standard weighted moving average ($a1 = 0.007$, $a2 = 0.0$, $b1 = 0.0$), (b) modified weighted moving average ($a1 = 0.004$, $a2 = 0.008$, $b1 = 0.0$), (c) modified weighted moving average ($a1 = 0.08$, $a2 = 0.0014$, $b1 = 0.001$)

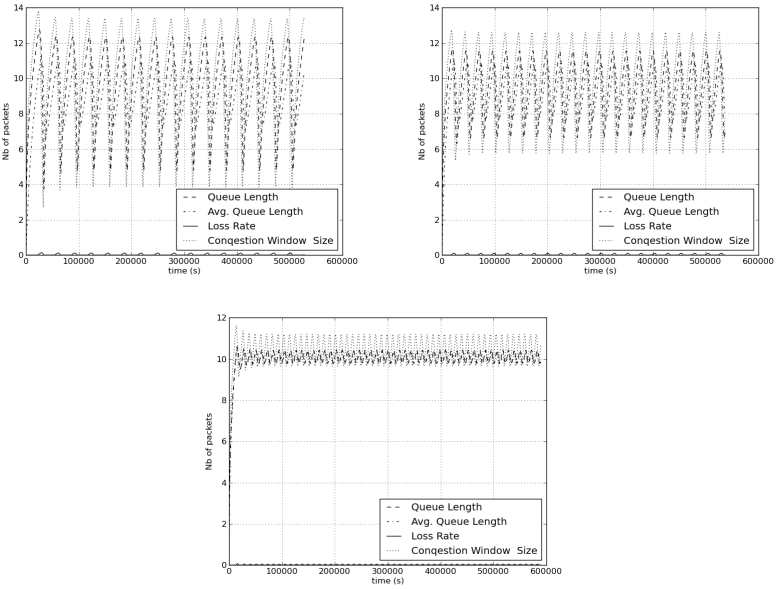


Fig. 4. RED queue, one TCP flow with UDP: (a) standard weighted moving average ($a1 = 0.007$, $a2 = 0.0$, $b1 = 0.0$), (b) modified weighted moving average ($a1 = 0.004$, $a2 = 0.008$, $b1 = 0.0$), (c) modified weighted moving average ($a1 = 0.08$, $a2 = 0.0014$, $b1 = 0.001$)

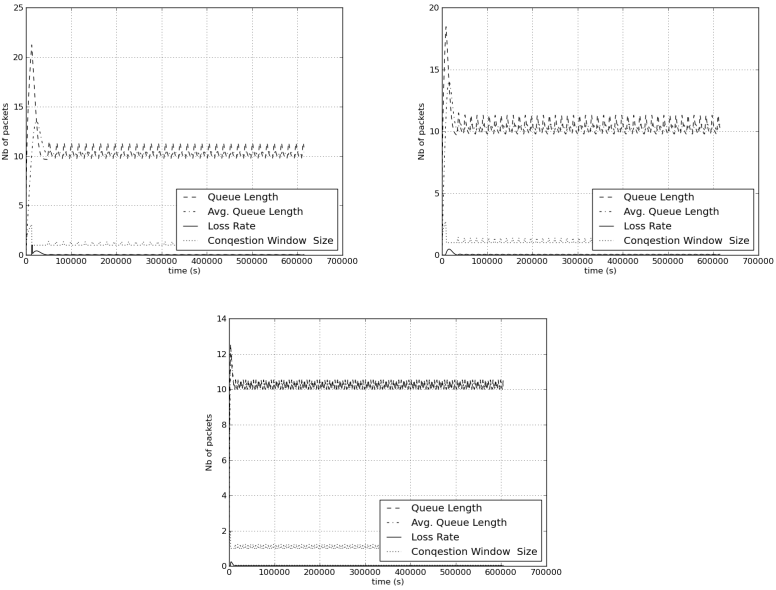


Fig. 5. RED queue, ten TCP flows with UDP: (a) standard weighted moving average ($a1 = 0.007$, $a2 = 0.0$, $b1 = 0.0$), (b) modified weighted moving average ($a1 = 0.004$, $a2 = 0.008$, $b1 = 0.0$), (c) modified weighted moving average ($a1 = 0.08$, $a2 = 0.0014$, $b1 = 0.001$)

4 Conclusions

The presented approach is based on a dynamical discrete model to define the average packet queue length and makes use of linear difference equations. The paramters used in these equations were received earlier in the open-loop scenario [3]. Numerical examples show that the classical RED displays the best performance in the case of small traffic. If the load and the share of UDP traffic flows increase, our modified RED algorithm performs better that the classical RED. This is in accordance with the results obtained by us previously in the open-loop scenario. Taking into account the increasing share of UDP traffic in the internet traffic, the use of our modification may assure better smoothness of transmissions (less fluctuation in congestion window).

Our study has also shown that the selection of the optimum parameters of the modified RED mechanism (the weight parameter of the moving average) depends on the type of traffic (the load and the TCP to UDP ratio).

Our future works will concentrate on creating an adaptive mechanism which will be change a way of counting the weighted moving average depending on the number of TCP and UDP flows passing through the router.

Acknowledgements. This research was partially financed by Polish Ministry of Science and Higher Education project no. N N516479640.

References

1. Augustyn, D.R., Domański, A., Domańska, J.: Active Queue Management with non linear packet dropping function. In: 6th International Conference on Performance Modelling and Evaluation of Heterogeneous Networks, HET-NETs (2010)
2. Augustyn, D.R., Domański, A., Domańska, J.: A Choice of Optimal Packet Dropping Function for Active Queue Management. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) CN 2010. CCIS, vol. 79, pp. 199–206. Springer, Heidelberg (2010)
3. Domańska, J., Domański, A., Augustyn, D.R.: The Impact of the Modified Weighted Moving Average on the Performance of the RED Mechanism. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) CN 2011. CCIS, vol. 160, pp. 37–44. Springer, Heidelberg (2011)
4. Chang Feng, W., Kandlur, D., Saha, D.: Adaptive packet marking for maintaining end to end throughput in a differentiated service internet. IEEE/ACM Transactions on Networking 7(5) (1999)
5. Chen, J., Paganini, F., Wang, R., Sanadidi, M.Y., Gerla, M.: Fluid-flow Analysis of TCP Westwood with RED. In: GLOBECOM (2004)
6. Czachórski, T., Grochla, K., Pekergin, F.: Stability and Dynamics of TCP-NCR(DCR) Protocol in Presence of UDP Flows. In: García-Vidal, J., Cerdà-Alabern, L. (eds.) Euro-NGI 2007. LNCS, vol. 4396, pp. 241–254. Springer, Heidelberg (2007)
7. Domańska, J., Domański, A., Czachórski, T.: The drop-from-front strategy in AQM. In: Koucheryavy, Y., Harju, J., Sayenko, A. (eds.) NEW2AN 2007. LNCS, vol. 4712, pp. 61–72. Springer, Heidelberg (2007)

8. Domańska, J., Domański, A., Czachórski, T.: Implementation of modified AQM mechanisms in IP routers. Journal of Communications Software and Systems 4(1) (March 2008)
9. Floyd, S., Jacobson, V.: Random early detection gateways for congestion avoidance. IEEE/ACM Transactions on Networking 1(4) (1993)
10. Hollot, C.V., Misra, V., Towsley, D.: A control theoretic analysis of RED. In: IEEE/INFOCOM (2001)
11. Hollot, C.V., Misra, V., Towsley, D., Gong, W.-B.: On Designing Improved Controllers for AQM Routers Supporting TCP Flows. In: IEEE INFOCOM (2002)
12. Hollot, C.V., Misra, V., Towsley, D., Gong, W.-B.: Analysis and design of controllers for AQM Routers Supporting TCP Flows. IEEE Trans. Automatic Control 47(6) (2002)
13. Kiddle, C., Simmonds, R., Williamson, C., Unger, B.: Hybrid packet/fluid flow network simulation. In: Parallel and Distributed Simulation (2003)
14. Liu, C., Jain, R.: Improving explicit congestion notification with the mark-front strategy. Computer Networks 35(2-3) (2000)
15. May, M., Diot, C., Lyles, B., Bolot, J.: Influence of active queue management parameters on aggregate traffic performance. Technical report, Research Report, Institut de Recherche en Informatique et en Automatique (2000)
16. Misra, V., Gong, W.-B., Towsley, D.: Fluid-based Analysis of a Network of AQM Routers Supporting TCP Flows with an Application to RED. In: ACM SIGCOMM (2000)
17. Mao, P., Xiao, Y., Hu, S., Kim, K.: Stable parameter settings for PI router mixing TCP and UDP traffic. In: IEEE 10th International Conference on Signal Processing, ICSP (2010)
18. http://www.scipy.org
19. Rahme, S., Labit, Y., Gouaisbaut, F.: An unknown input sliding observer for anomaly detection in TCP/IP networks. In: Ultra Modern Telecommunications & Workshops (2009)
20. Wang, L., Li, Z., Chen, Y.-P., Xue, K.: Fluid-based stability analysis of mixed TCP and UDP traffic under RED. In: 10th IEEE International Conference on Engineering of Complex Computer Systems (2005)
21. Yung, T.K., Martin, J., Takai, M., Bagrodia, R.: Integration of fluid-based analytical model with Packet-Level Simulation for Analysis of Computer Networks. In: SPIE (2001)
22. Zhou, K., Yeung, K.L., Li, V.O.K.: Nonlinear RED: Asimple yet efficient active queue management scheme. Computer Networks 50 (2006)

A Tandem Queueing System with Batch Session Arrivals

Sergey Dudin* and Olga Dudina

Belarusian State University, 4, Nezavisimosti Ave.,
Minsk, 220030, Belarus
dudin@madrid.com, dudina_olga@email.com

Abstract. We consider a tandem queueing system with session arrivals. Session means a group of customers which should be sequentially processed in the system. In contrast to the standard batch arrival when a whole group of customers arrives into the system at one epoch, we assume that the customers of an accepted session arrive one by one in exponentially distributed times. Generation of sessions at the first stage is described by a Batch Markov Arrival Process ($BMAP$). At the first stage of tandem, it is determined whether a session has the access to the second stage. After the first stage the session moves to the second stage or leaves the system. At the second stage having a finite buffer the customers from sessions are serviced. A session consists of a random number of customers. This number is geometrically distributed and is not known at a session arrival epoch. The number of sessions, which can be admitted into the second stage simultaneously, is subject to control. An accepted session can be lost, with a given probability, in the case of any customer from this session rejection.

Keywords: tandem system, batch Markovian arrival process, session admission control, performance modeling.

1 Introduction

Queueing theory is widely used for modelling and performance evaluation of modern telecommunications networks. Typically, an user of telecommunication system can generate not a single request but a group of requests. That is the reason why a Batch Markovian Arrival Process ($BMAP$) as an arrival process is assumed when the queueing system that modelles a real telecommunication system is considered. The $BMAP$ was introduced by D. Lucantoni in [1]. In [1], a single-server queueing system with the $BMAP$ arrival process, the general service time distribution and an infinite buffer is analyzed. In [2], a departure process of $BMAP/G/1$ is analysed. In [3, 4], $BMAP/G/1$ queue with controlled service intensity is investigated. $BMAP/G/1$ queue with generalized vacations is considered in [5] and $BMAP/G/1$ with disasters is considered in [6].

* Corresponding author.

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 59–68, 2013.

$BMAP/G/1$ cyclic polling models are investigated in [7]. $BMAP/G/1$ retrial queueing system is considered in [8, 9].

Single-server queues with $BMAP$ arrivals, semi-Markovian service process and an infinite buffer are analysed, e.g., in [10–12].

If a queueing system with a finite buffer and $BMAP$ arrivals is considered, it is assumed that at a batch arrival epoch all requests of the batch arrive into the system simultaneously and decision whether or not the batch should be admitted into the system is based on comparison of the batch size and the available capacity of the system, see, e.g., [13], [14].

Very general model of the $BMAP/SM/1/N$ type with discipline of partial admission is investigated in [15], and the $BMAP/G/1/N$ type with disciplines of complete rejection and complete admission is investigated in [16]. Numerically stable algorithms, which are taking into account special structure of transition probability matrix and are suitable even if the buffer capacity N is equal to several thousands, are presented there.

Tandem queueing systems with $BMAP$ arrivals are considered, e.g., in [17–20].

The queues with $BMAP$ arrival process are well suited for modeling the real systems in which the requests can arrive simultaneously. However, in many nowadays communication networks, IP networks in particular, customers can arrive in groups, but the arrival of customers from a group is not simultaneous. To distinguish the standard batches from the group with non-simultaneous customers arrivals the latter ones are called sessions.

Session arrivals are typical for multiple access telecommunication system which resources are shared by a set of users. An user establishes a session (sends the first request) when it enters the system. If this user's request is admitted to the system, the session is considered as established. Once the user has established the session, he(she) can generate the sequence of requests. Belonging of the requests to established sessions is determined by means of IP address. Note that the number of requests at a session is random and unknown at the session arrival epoch. If the arrival request belongs to existed session, it is accepted to the system. If the request belongs to a new session (the first request of the session), the buffer and channel capacity is still available, and the number of session in the system is non critical, the session and request are admitted into the system and the session count is increased. Otherwise, the session and its first request are rejected. When the requests from admitted session do not arrive to the system during a certain time interval, the session is assumed to be finished and the session count is decreased by one.

Due to the requests from a session arrive to the system non-simultaneously and the number of requests in a session is unknown at the session arrival epoch it is impossible to make a decision to accept or not the arriving session to the system based on comparison the session size with the available capacity of the system. Under consideration queues with session arrivals, it is assumed that the number of sessions is restricted by means of so called tokens. The number of

tokens, which defines the maximal number of flows that can be admitted into the system simultaneously, is very important control parameter.

In paper [21], a novel finite capacity queueing model of $M/M/N/R$ type with request arrivals in sessions is investigated. In paper [22], the $MAP/PH/1/N$ queueing system with session arrivals is investigated. It was assumed in [21] and [22], that the sessions arrival is regulated by means of tokens. The pool of tokens consists of K tokens and a new session is admitted to the system only if there is an available token and the buffer is not full at a session arrival epoch. Otherwise, the session leaves the system forever.

In paper [23], the mechanism of requests arrival within a session is significantly generalized comparing to the model considered in [22] by suggesting that the customers from the admitted session can arrive in groups. Session arrivals are directed by a MAP (Markovian Arrival Process) and customers' arrivals in session are directed by the $BMAP$ in [23].

In presented paper the tandem queueing system with $BMAP$ arrivals of session is investigated. At the first stage of the system it is determined whether an arriving session has the access to the system. After the first stage the session moves to the second one if it has the access or leaves the system. At the second stage admitted sessions are serviced.

The paper is organized as follows. In section 2, the mathematical model is described. The stationary distribution of system states is analyzed in section 3. The expressions for the main system performance measures are given in section 4. Section 5 concludes the paper.

2 Mathematical Model

The system consists of two stages. The first stage is a single server queueing system with a finite buffer of capacity $R, 1 \leq R < \infty$.

The customers arrive to the system in sessions. Groups of sessions arrive at the first stage according to the Batch Markov Arrival Process ($BMAP$). Sessions arrival in the $BMAP$ is directed by an irreducible continuous time Markov chain ν_t, $t \geq 0$, with the finite state space $\{0, 1, \ldots, W\}$. The sojourn time of the Markov chain ν_t in the state ν has an exponential distribution with the parameter λ_ν, $\nu = \overline{0, W}$. After this sojourn time expires, with probability $p_l(\nu, \nu')$ the process ν_t transits to the state ν', and l, $l \geq 0$, sessions arrive to the system.

The intensities of jumps from one state into another, which are accompanied by an arrival of l sessions, are combined into the square matrices $D_l, l \geq 0$, of size $\bar{W} = W + 1$. The matrix generating function of these matrices is $D(z) = \sum_{l=0}^{\infty} D_l z^l, |z| \leq 1$.

The (ν, ν')th entry of the matrix D_l has form

$$(D_l)_{\nu,\nu'} = \lambda_\nu p_l(\nu, \nu'),\ \nu,\ \nu' = \overline{0, W},\ l \geq 1,$$

$$(D_0)_{\nu,\nu'} = \begin{cases} \lambda_\nu p_0(\nu, \nu'), & \nu \neq \nu',\ \nu,\ \nu' = \overline{0, W}; \\ -\lambda_\nu, & \nu = \nu',\ \nu = \overline{0, W}. \end{cases}$$

The matrix $D(1)$ is the infinitesimal generator of the process ν_t, $t \geq 0$. The stationary distribution vector $\boldsymbol{\chi}$ of this process satisfies the equations $\boldsymbol{\chi} D(1) = \mathbf{0}$, $\boldsymbol{\chi}\mathbf{e} = 1$. Here and in the sequel $\mathbf{0}$ is a zero row vector and $\mathbf{e}$ denotes unit column vector.

The average intensity λ (fundamental rate) of the sessions arrivals is defined as $\lambda = \boldsymbol{\chi} D'(z)|_{z=1}\mathbf{e}$. The intensity λ_b of group session arrivals is defined as $\lambda_b = \boldsymbol{\chi}(-D_0)\mathbf{e}$. The coefficient of variation c_{var} of intervals between group session arrivals is defined by $c_{var}^2 = 2\lambda_b \boldsymbol{\chi}(-D_0)^{-1}\mathbf{e} - 1$. The coefficient of correlation c_{cor} of the successive intervals between group session arrivals is given by $c_{cor} = (\lambda_b \boldsymbol{\chi}(-D_0)^{-1}(D(1) - D_0)(-D_0)^{-1}\mathbf{e} - 1)/c_{var}^2$.

The service time of a session at the first stage is exponentially distributed with the parameter η.

If at the arrival epoch of a batch of sessions the size of the batch does not exceed the number of available waiting places, the whole group is admitted to the system. Otherwise, the sessions, for which there is no available place in the buffer, leave the system forever. This means that we assume so called partial sessions admission discipline. Complete rejection and complete admission disciplines need separate treatment.

After service at the first stage a session leaves the system forever with probability q, $0 \leq q \leq 1$, or proceeds to the second stage with complementary probability.

The second stage consists of N identical independent servers and a finite buffer of capacity $M, 1 \leq M < \infty$.

We assume that admission of sessions (they are called also flows, connections, sessions, exchanges, windows, etc. in different real-life applications) to the second stage is restricted by means of tokens. The total number of available tokens is assumed to K, $K \geq 1$.

If there is no available token at a session arrival epoch at the second stage or the buffer at the second stage is full, the session is rejected, and leaves the system forever. If the number of available tokens at the session arrival epoch at the second stage is positive and the buffer is not full, this session is admitted into the second stage and the number of available tokens decreases by one. We assume that the first request of a session arrives at the session arrival epoch and if it meets a free server at the second stage, it occupies the server and is processed. If all servers are busy, the customer moves to a buffer and later it is picked up for the service according to the First Came - First Served discipline. After admission of the session at the second stage, the next customer of this session should arrive directly into the second stage in a random interval length which is exponentially distributed with the parameter γ.

If there is an available server at the second stage, the customer is admitted, otherwise, it is rejected and leaves the system forever. If the customer from admitted session is rejected, this session leaves the system forever with probability p, $0 \leq p \leq 1$, and releases the token. The rejection of customer does not affect on the future behavior of the session with complementary probability $1 - p$.

The number of customers in the session has geometrical distribution with parameter θ, i.e., probability that the flow consists of k customers is equal to $\theta^{k-1}(1-\theta),\ k \geq 1$. If the random time since arrival of the previous customer of a session expires and a new customer does not arrive, it means that the arrival of the session is finished. The token, which was obtained by this flow upon arrival, is returned into the pool of available tokens. The customers of this session, which stay in the buffer of the second stage at the epoch of returning the token, should be completely processed by the second stage.

The service time of a customer at the second stage is exponentially distributed with the parameter μ.

3 The Process of System States

Let $i_t,\ i_t = \overline{0, R+1}$, be the number of sessions at the first stage, $n_t,\ n_t = \overline{0, N+M}$, be the number of customers at the second stage, $k_t,\ k_t = \overline{0, K}$, be the number of sessions having token for admission to the system, $\nu_t,\ \nu_t = \overline{0, W}$, be the state of the directing process of the $BMAP$ arrival process at the epoch $t,\ t \geq 0$.

It is obvious that the four-dimensional process $\xi_t = \{i_t, n_t, k_t, \nu_t\},\ t \geq 0$, is the irreducible regular continuous time Markov chain.

Let us enumerate the states of this Markov chain in lexicographic order and refer to (i, n) as macro-state consisting of $\bar{K} = \bar{W}(K+1)$ states $(i, n, k, \nu),\ k = \overline{0, K},\ \nu = \overline{0, W}$.

Introduce the following notation:

- I_m is an identity matrix of size m, O_m is a zero matrix of size $m \times m$;
- $\gamma^- = \gamma(1-\theta),\ \gamma^+ = \gamma\theta$;
- $\otimes$ and $\oplus$ are symbols of Kronecker's sum and product respectively, see, e.g., [24];
- $\tilde{C} = \mathrm{diag}\{0, 1, \ldots, K\}$, $C = \tilde{C} \otimes I_{\bar{W}}$;
- E^- is the square matrix of size $K+1$ with all zero entries except entries $(E^-)_{i,i-1},\ i = \overline{1, K}$, which are equal to 1;
- $E_l^+,\ l = N+M, K$ is the square matrix of size $l+1$, with all zero entries except entries $(E_l^+)_{i,i+1},\ i = \overline{0, l-1}$, $(E_l^+)_{l,l} = 1$, which are equal to 1;
- $A = (-\gamma\tilde{C} + \gamma^-\tilde{C}E^-) \otimes I_{\bar{W}}$;
- $\delta_{i,j}$ is Kronecker delta, $\delta_{i,j}$ is equal to 1 if $i = j$ and equal to 0 otherwise.

Let Q be the generator of the Markov chain $\xi_t,\ t \geq 0$, with blocks $Q_{i,j}$ consisting of intensities $(Q_{i,j})_{n,n'}$ of this chain transitions from the macro-state (i, n) into the macro-state $(j, n'),\ n, n' = \overline{0, N+M}$. The diagonal entries of the matrix $Q_{i,i}$ are negative and the modulus of the diagonal entry of $(Q_{i,i})_{n,n}$ defines the total intensity of leaving the corresponding state (i, n, k, ν) of the Markov chain. The block $Q_{i,j},\ i, j = \overline{0, R+1}$, has dimention $\bar{M} \times \bar{M}$, where $\bar{M} = \bar{K}(N+M+1)$.

Lemma 1. *The generator Q of the Markov chain $\xi_t,\ t \geq 0$, has the following block structure*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & Q_{0,2} & \dots & Q_{0,R} & Q_{0,R+1} \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & \dots & Q_{1,R} & Q_{1,R+1} \\ O & Q_{2,1} & Q_{2,2} & \dots & Q_{2,R} & Q_{2,R+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \dots & Q_{R,R} & Q_{R,R+1} \\ O & O & O & \dots & Q_{R+1,R} & Q_{R+1,R+1} \end{pmatrix},$$

where non-zero blocks $Q_{i,j}$ are defined by

$$Q_{i,i} = \begin{pmatrix} C_{0,0}^{(i)} & C_{0,1} & O & \dots & O & O \\ C_{1,0} & C_{1,1}^{(i)} & C_{1,2} & \dots & O & O \\ O & C_{2,1} & C_{2,2}^{(i)} & \dots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \dots & C_{N+M-1,N+M-1}^{(i)} & C_{N+M-1,N+M} \\ O & O & O & \dots & C_{N+M,N+M-1} & C_{N+M,N+M}^{(i)} \end{pmatrix}, \ i = \overline{0, R+1},$$

$$Q_{i,i+l} = \begin{cases} I_{(N+M+1)(K+1)} \otimes D_l, & 0 < l < R-i+1, \\ I_{(N+M+1)(K+1)} \otimes \sum\limits_{j=l}^{\infty} D_j, & l = R-i+1 \end{cases}, \ i = \overline{0, R}.$$

$$Q_{i,i-1} = Q^- = \eta(qI_{(N+M+1)(K+1)\bar{W}} + (1-q)E_{N+M}^+ \otimes E_K^+ \otimes I_{\bar{W}}), i = \overline{1, R+1}.$$

Here

- $C_{n,n}^{(i)} = A - (1-\delta_{i,0})\eta I_{\bar{K}} + I_{K+1} \otimes [D_0 + \delta_{i,R+1}(D(1) - D_0)] - \min\{n, N\}\mu I_{\bar{K}},\ n = \overline{0, N+M-1},\ i = \overline{0, R+1}$;
- $C_{N+M,N+M}^{(i)} = A - (1-\delta_{i,0})\eta I_{\bar{K}} + I_{K+1} \otimes [D_0 + \delta_{i,R+1}(D(1) - D_0)] - N\mu I_{\bar{K}} + \gamma^+(1-p)C + \gamma^+ p(\tilde{C}E^-) \otimes I_{\bar{W}},\ i = \overline{0, R+1}$;
- $C_{n,n+1} = \gamma^+ C,\ n = \overline{0, N+M-1}$,
- $C_{n,n-1} = \min\{n, N\}\mu I_{\bar{K}},\ n = \overline{1, N+M}$.

Proof of the lemma consists of analysis of the Markov chain $\xi_t,\ t \geq 0$, transitions during the infinitesimal interval of time and further combining corresponding transition intensities into the matrix blocks. Value γ^- is the intensity of tokens releasing due to the finish of the session arrival, γ^+ is the intensity of new customers arrival in the session.

Since the four-dimensional Markov chain $\xi_t = \{i_t, n_t, k_t, \nu_t\},\ t \geq 0$, is the irreducible and regular and has the finite state space, the following limits (stationary probabilities) exist:

$$\pi(i, n, k, \nu) = \lim_{t \to \infty} P\{i_t = i,\ n_t = n,\ k_t = k,\ \nu_t = \nu\},$$

$$i = \overline{0, R+1},\ n = \overline{0, N+M},\ k = \overline{0, K},\ \nu = \overline{0, W}.$$

Let us combine these probabilities into the row-vectors

$$\boldsymbol{\pi}(i,n,k) = (\pi(i,n,k,0), \pi(i,n,k,1), \ldots, \pi(i,n,k,W)),\ k = \overline{0,K},$$

$$\boldsymbol{\pi}(i,n) = (\boldsymbol{\pi}(i,n,0), \boldsymbol{\pi}(i,n,1), \ldots, \boldsymbol{\pi}(i,n,K)),\ n = \overline{0,N+M},$$

$$\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i,0), \boldsymbol{\pi}(i,1), \ldots, \boldsymbol{\pi}(i,N+M)),\ i = \overline{0,R+1}.$$

It is well known that the vector $(\boldsymbol{\pi}_0, \ldots, \boldsymbol{\pi}_{R+1})$ is the unique solution to the following system of linear algebraic equations:

$$(\boldsymbol{\pi}_0, \ldots, \boldsymbol{\pi}_{R+1})Q = \mathbf{0},\ \ (\boldsymbol{\pi}_0, \ldots, \boldsymbol{\pi}_{R+1})\mathbf{e} = 1.$$

This system can be solved on computer directly ("by brute force"). Alternatively, the following numerically stable algorithm for solving this system, which takes into account the special structure of the generator Q, can be applied.

Step 1. Compute the matrices $P_{i,j}$ recurrently:

$$P_{i,R+1} = -Q_{i,R+1}(Q_{R+1,R+1})^{-1},\ i = \overline{0,R},$$

$$P_{i,j} = -(Q_{i,j} + P_{i,j+1}Q^{-})(Q_{j,j} + P_{j,j+1}Q^{-})^{-1},\ i = \overline{0,j-1},\ j = R, R-1, \ldots, 1.$$

Step 2. Calculate the matrices $\Phi_j, j = \overline{0,R+1}$:

$$\Phi_0 = I,\ \Phi_j = \sum_{i=0}^{j-1} \Phi_i P_{i,j},\ j = \overline{1,R+1}.$$

Step 3. Calculate the vector $\boldsymbol{\pi}_0$ as the unique solution to the following system of linear algebraic equations:

$$\boldsymbol{\pi}_0(Q_{0,0} + P_{0,1}Q^{-}) = \boldsymbol{\pi}_0,\ \ \boldsymbol{\pi}_0 \sum_{j=0}^{R+1} \Phi_j \mathbf{e} = 1.$$

Step 4. Calculate the vectors $\boldsymbol{\pi}_j$: $\boldsymbol{\pi}_j = \boldsymbol{\pi}_0\Phi_j,\ j = \overline{1,R+1}$.

4 Performance Measures

As soon as the vectors $\boldsymbol{\pi}_i,\ i = \overline{0,R+1}$, have been calculated, we are able to find various performance measures of the system under consideration.

The average number of sessions at the first stage is calculated as

$$L^{(1)} = \sum_{i=1}^{R+1} i\boldsymbol{\pi}_i \mathbf{e}.$$

The average number of customers at the second stage is calculated as

$$L^{(2)} = \sum_{i=0}^{R+1} \sum_{n=1}^{N+M} n\boldsymbol{\pi}(i,n)\mathbf{e}.$$

The average number of sessions in the buffer at the first stage is calculated as

$$N_{buffer}^{(1)} = \sum_{i=2}^{R+1}(i-1)\boldsymbol{\pi}_i\mathbf{e}.$$

The average number of customers in the buffer at the second stage is calculated as

$$N_{buffer}^{(2)} = \sum_{i=0}^{R+1}\sum_{n=N+1}^{N+M}(n-N)\boldsymbol{\pi}(i,n)\mathbf{e}.$$

The average number of busy servers at the second stage is calculated as

$$N_{server} = \sum_{i=0}^{R+1}(\sum_{n=1}^{N} n\boldsymbol{\pi}(i,n)\mathbf{e} + N\sum_{n=N+1}^{N+M}\boldsymbol{\pi}(i,n)\mathbf{e}).$$

The intensity of flow of sessions, which get the service at the first stage, is calculated as

$$\lambda_{out}^{(1)} = \eta(1-\boldsymbol{\pi}_0\mathbf{e}).$$

The intensity of flow of customers, which get the service in the system, is calculated as

$$\lambda_{out}^{(2)} = \mu N_{server}.$$

The loss probability of whole group of sessions at the entrance to the first stage due to buffer overflow is calculated as

$$P^{(ent-loss)} = \lambda_b^{-1}\boldsymbol{\pi}_{R+1}(\mathbf{e}\otimes\sum_{k=1}^{\infty}D_k\mathbf{e}).$$

The average number of sessions at the second stage is computed as

$$B = \sum_{i=0}^{R+1}\sum_{n=0}^{N+M}\sum_{k=1}^{K}k\boldsymbol{\pi}(i,n,k)\mathbf{e}.$$

The loss probability of arbitrary session at the first stage is calculated as

$$P_1^{(session-loss)} = \lambda^{-1}\sum_{i=0}^{R+1}\boldsymbol{\pi}_i(\mathbf{e}\otimes\sum_{k=R-i+2}^{\infty}(i+k-R-1)D_k\mathbf{e}).$$

The probability $P_s^{(loss)}$ of an arbitrary session rejection upon arrival at the second stage is computed by

$$P_2^{(session-loss)} = \frac{\eta}{\lambda_{out}^{(1)}}\sum_{i=1}^{R+1}\left(\sum_{n=0}^{N+M-1}\boldsymbol{\pi}(i,n,K) + \sum_{k=0}^{K}\boldsymbol{\pi}(i,N+M,k)\right)\mathbf{e}.$$

The probability $P_c^{(loss)}$ of an arbitrary customer from admitted session rejection is computed by

$$P_c^{(loss)} = \sum_{i=0}^{R+1} \frac{\sum_{k=1}^{K} k\gamma^{+}\boldsymbol{\pi}(i, N+M, k)\mathbf{e}}{\sum_{k=1}^{K}\sum_{n=0}^{N+M} k\gamma^{+}\boldsymbol{\pi}(i, n, k)\mathbf{e}}.$$

5 Conclusion

A tandem queueing system with batch session arrivals is investigated. The system underlying process is constructed. The stable algorithm for calculation of the stationary distribution of system states is presented. The key system performance measures are computed.

Acknowledgments. This research was supported by Belarusian Republican Foundation for Fundamental Research (grant No. F11M-003).

References

1. Lucantoni, D.: New results on the single server queue with a batch Markovian arrival process. Communication in Statistics-Stochastic Models 7, 1–46 (1991)
2. Ferng, H.W., Chang, J.F.: Departure processes of $BMAP/G/1$ queues. Queueing Systems 39, 100–135 (2001)
3. Dudin, A.: Optimal multithreshold control for a $BMAP/G/1$ queue with N service modes. Queueing Systems 30, 273–287 (1998)
4. Dudin, A.N., Nishimura, S.: Optimal Control for a $BMAP/G/1$ Queue with Two Service Modes. Mathematical Problems in Engineering 5, 255–273 (1999)
5. Chang, S.H., Takine, T., Chae, K.C., Lee, H.W.: A unified queue length formula for $BMAP/G/1$ queue with generalized vacations. Stochastic Models 18, 369–386 (2002)
6. Shin, Y.W.: $BMAP/G/1$ queue with correlated arrivals of customers and disasters. Operations Research Letters 32, 364–373 (2004)
7. Saffer, Z., Telek, M.: Unified analysis of $BMAP/G/1$ cyclic polling models. Queueing Systems 64, 69–102 (2010)
8. Dudin, A., Klimenok, V.: Queueing system $BMAP/G/1$ with repeated calls. Mathematical and Computer Modelling 30, 115–128 (1999)
9. Dudin, A.N., Krishnamoorthy, A., Joshua, V.C., Tsarenkov, G.V.: Analysis of the $BMAP/G/1$ retrial system with search of customers from the orbit. European Journal of Operational Research 157, 169–179 (2004)
10. Dudin, A., Semenova, O.: A stable algorithm for stationary distribution calculation for a $BMAP/SM/1$ queueing system with Markovian arrival input of disasters. Journal of Applied Probability 41, 547–556 (2004)
11. Semenova, O.V.: Optimal control for a $BMAP/SM/1$ queue with map-input of disasters and two operation modes. RAIRO - Operations Research 38, 153–171 (2004)

12. Choi, B.D., Chung, Y.H., Dudin, A.N.: $BMAP/SM/1$ retrial queue with controllable operation modes. European Journal of Operational Research 131, 16–30 (2001)
13. Blondia, C.: The $N/G/1$ finite capacity queue. Communications in Statistics - Stochastic Models 5, 273–274 (1989)
14. Dudin, A.N., Nishimura, S.: Optimal hysteretic control for a $BMAP/SM/1/N$ queue with two operation modes. Mathematical Problems in Engineering 5, 397–420 (2000)
15. Dudin, A.N., Klimenok, V.I., Tsarenkov, G.V.: Characteristics calculation for single-server queue with the $BMAP$ input, SM service and finite buffer. Automation and Remote Control 63, 1285–1297 (2002)
16. Dudin, A.N., Shaban, A.A., Klimenok, V.I.: Analysis of a $BMAP/G/1/N$ queue. International Journal of Simulation: Systems, Science and Technology 6, 13–23 (2005)
17. Kim, C.S., Klimenok, V., Tsarenkov, G., Breuer, L., Dudin, A.: The $BMAP/G/1 \to \bullet/PH/1/M$ tandem queue with feedback and losses. Performance Evaluation 64, 802–818 (2007)
18. Klimenok, V., Breuer, L., Tsarenkov, G., Dudin, A.: The $BMAP/G/1/N \to \bullet/PH/1/M$ tandem queue with losses. Performance Evaluation 61, 17–40 (2005)
19. Kim, C.S., Klimenok, V., Taramin, O.: A tandem retrial queueing system with two Markovian flows and reservation of channels. Computers and Operations Research 37, 1238–1246 (2010)
20. Kim, C., Dudin, A., Klimenok, V., Taramin, O.: A tandem $BMAP/G/1 \to \bullet/M/N/0$ queue with group occupation of servers at the second station. Mathematical Problems in Engineering 2012, art. no. 324604 (2012)
21. Lee, M.H., Dudin, S., Klimenok, V.: Queueing Model with Time-Phased Batch Arrivals. In: Mason, L.G., Drwiega, T., Yan, J. (eds.) ITC 2007. LNCS, vol. 4516, pp. 719–730. Springer, Heidelberg (2007)
22. Kim, C.S., Dudin, S.A., Klimenok, V.I.: The $MAP/PH/1/N$ queue with time phased arrivals as model for traffic control in telecommunication networks. Performance Evaluation 66, 564–579 (2009)
23. Kim, C.S., Dudin, A., Dudin, S., Klimenok, V.: A Queueing System with Batch Arrival of Customers in Sessions. Computers and Industrial Engeneering 62, 890–897 (2012)
24. Graham, A.: Kronecker Products and Matrix Calculus with Applications. Ellis Horwood, Chichester (1981)

Socio-behavioral Scheduling of Time-Frequency Resources for Modern Mobile Operators

Alexander Dudin[1,⋆], Evgeny Osipov[2], Sergey Dudin[1], and Olov Schelén[2]

[1] Belarusian State University, 4, Nezavisimosti Ave., Minsk, 220030, Belarus
[2] Lulea University of Technology, Lulea, Sweden
dudin@bsu.by, dudin@madrid.com,
{evgeny.osipov,olov.schelen}@ltu.se

Abstract. This article presents a mathematical foundation for scheduling of batch data produced by mobile end users over the time-frequency resources provided by modern mobile operators. We model the mobile user behavior by Batch Markovian Arrival Process, where a state corresponds to a specific user data activity (i.e. sending a photo, writing a blog message, answering an e-mail etc). The state transition is marked by issuing a batch of data of the size typical to the activity. To model the changes of user behavior caused by the environment, we introduce a random environment which affects the intensities of transitions between states (i.e., the probabilities of the user data activities). The model can be used for calculating probability of packet loss and probability of exceeding the arbitrarily fixed value by the sojourn time of a packet in the system conditional that the packet arrives to the system at moments when the random environment has a given state. This allows to compute the realistic values of these probabilities and can help to properly fix their values that can be guaranteed, depending on the state of the random environment, by a service provider.

Keywords: batch Markovian arrival process, random environment, phase type service time distribution.

1 Introduction

According to the vision of leading vendors of equipment for cellular networks[1] it is expected that the traffic demand in *Long Term Evolution* (LTE) networks will increase thousand times by the year 2020. By the same time it is expected that the total number of diverse radio communicating devices will reach fifty billion. Besides this visionary drastic increase of mobile terminals, the mobile operators experience major difficulties in providing quality of service in densely populated areas already at present days. According to the recent studies by

⋆ Corresponding author.

[1] "More than 50 billion connected devices", Ericsson white paper 284 23-3149 Uen, February 2011. [Online] Available: http://www.ericsson.com/res/docs/whitepapers/wp-50-billions.pdf

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 69–82, 2013.

Actix [2] *three quarters* of subscribers in transport hubs, central business districts, tourist centers and the locations for major conventions and sporting events are not getting satisfactory quality during peak times. The drop in the data rate in worst cases could be as bad as 95 percent. One of the major reasons for this is user behavior agnostic way of distributing time-frequency resources. Historically and due to a tough competition for customers, the mobile operators normally offer data services as relatively cheep flat-rate-priced data pipes without accounting for traffic patterns generated by diverse multimedia applications.

This article presents a mathematical foundation for socio-behavioral scheduling of time-frequency resources for modern mobile operators. We model the mobile user behavior by Batch Markovian Arrival Process (BMAP), where a state corresponds to a specific user data activity (e.g., sending a photo, writing a blog message, answering an e-mail) and the state transition is marked by issuing a batch of data of the size typical to the activity.

To capture the fact that the user behavior changes based on the external environment we introduce a random environment (RE). The intensities of transitions between states (i.e., the probabilities of the user data activities) are governed by a process describing a random environment in which the user is presently located (e.g., a soccer game, busy hour business activities). The random environment can be further detailed to include specific events in the current environment (e.g., a goal being made at a soccer match).

The user activities impose a varying load on the LTE base stations both up link and down link. In this paper we focus on scheduling *up link* transmissions in an LTE-A cell and concentrate on the proof of the properties of the suggested mathematical model. The effect of the user activities is than that batches of data are queued in the buffers of the wireless devices. The data of a queue is served by a base-station and herein each base station is modeled as a number of servers, where each server handles resource in part of the frequency domain.

The model can be used to calculate critical performance metrics of the system such as the data/packet loss probability when using finite buffers, the mean amount of data queued, the mean amount of served data (i.e., number of busy servers and idle servers), the probability of exceeding delay limits by the sojourn time of a packet, etc.

An analogous queueing model was considered in [1]. In this article we use another multi-dimensional Markov chain for description of the system behavior what allows to compute performance measures of the system for much larger number of servers. We focus on two key performance measures of the system separately for data (customers) arriving to the system at the time periods when the RE has given states and discuss the issues of application of the model for managing operation of the system with different requirements to the quality of the service under different states of the RE.

[2] "Actix finds 75 % of subscribers cannot get the data speeds they want at peak times", [Online.] Available: http://www.actix.com/sites/www.actix.com/files/Actix_Hotspots_Study_Findings.pdf

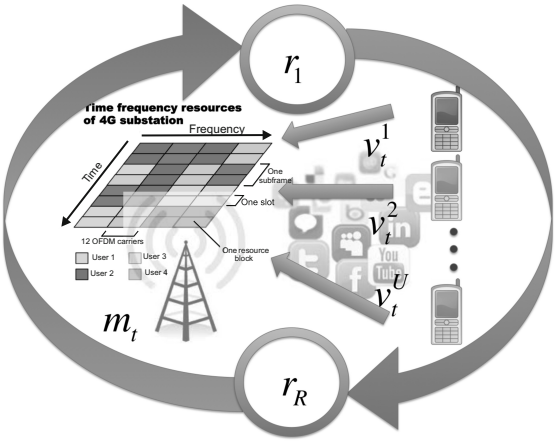


Fig. 1. Socio-behavioral approach to uplink resource scheduling

The article is structured as follows. In section 2 we overview major mechanisms of data transmission in LTE-Advanced wireless networks. The mathematical model is described in section 3. The process of the system states as a multi-dimensional continuous time Markov chain and its generator are analyzed in section 4. Expressions for some key performance measures of the system are given in section 5. Sojourn time distributions of an arbitrary customer and of an arbitrary customer arriving under the fixed value of the random environment are obtained in terms of Laplace-Stieltjes transform in section 6. Expressions for two initial moments of these distributions and approximate formulas for distribution functions are given there. Section 7 concludes the paper.

2 LTE Scheduling Principle, Assumptions and Model Notations

In fourth generation wireless networks also known under name LTE [2] a combination of *frequency and time division* multiple access also known as *Orthogonal Frequency Division Multiplexing* is used to communicate data between a base station and a mobile terminal. Essentially, this means that the time line for all mobile terminals and the base station is divided into fixed length time intervals. During each interval multiple terminals may concurrently transmit data to the base station using a unique set of frequencies assigned by the base station without disturbing each other. In the frequency domain, depending on the strength of the signal, different modulation techniques can be used allowing for different data transmission rates. A scheduler of the time-frequency resources is a key element which to a large extent determines the overall system performance.

According to the LTE specification the minimum scheduled to a terminal quantum of spectrum is 180 kHz. This portion of the spectrum assigned to a particular terminal during 0.5 ms is referred as to one *resource block*. The scheduling unit in

time is equal to 1 ms, in terminology of LTE referred as to a *subframe*. While the details of scheduler's implementation is left outside the standard and is vendor-specific, in principle, a scheduled terminal can be assigned an arbitrary combination (but not more than 110) of resource blocks in each 1ms subframe. In the case of a good channel quality between a terminal and the base station this corresponds to the base station's grant to transmit from 50 to 5000 bytes of user data roughly.

The overall approach to the modeling is graphically illustrated in Figure 1. We consider the case where the system in question changes it's behavior in reaction to a random environment process r_t, $t \geq 0$, which is assumed to be an irreducible continuous time Markov chain with the state space $\{1, \ldots, R\}$, $R \geq 2$, and the infinitesimal generator Q. An example of such process would be scoring events at a soccer game, business activities at busy hours, arrival, departure of transport at a transport hub etc. Different states of the random environment affect the overall user behavior and the service times at the base station as described below.

Since implementation of the base station's scheduler is vendor-specific, for the sake of modeling we describe an abstract generic scheduler as a set of N identical *virtual servers*. Each server is responsible for policing of user's traffic in time within 180kHz chunk of bandwidth according to the defined below *service rate*. In order to account for the base station scheduler's ability to assign multiple resource blocks within the 1 ms subframe we model the content of the transmit buffer in *each* user terminal as a batch of customers and allow customers from one batch to use more than one server simultaneously.

A *batch* is all the data generated instantaneously by a terminal as a reaction to the user activities. A batch consists of a number of customers. The customers belonging to the same batch could be of different sizes, which are the multiples of a number of bits that could be transmitted up-link with a selected modulation technique over a time interval of 1 millisecond. In the real system one customer corresponds to the data from one batch scheduled on an arbitrary 180 kHz portion of bandwidth. Note that the batches of data *produced* by the user activities are queued consecutively in the wireless device and that the sizes of the *served* customers are independent, i.e., the scheduler can at each epoch serve any size customer from the queue.

By modeling the uplink scheduler this way and analyzing the performance results for different distribution of customer sizes within batches would allow to derive rules for real scheduler at the base station which would deliver the desired per-user performance. This type of the result interpretation is, however, outside the scope for this article and will be reported elsewhere.

3 The Mathematical Model

Let N be the number of identical independent servers. The service time in a server time is interpreted as the time until the irreducible continuous time Markov chain m_t, $t \geq 0$, with the state space $\{1, \ldots, M+1\}$ reaches the absorbing state $M+1$. Under the fixed value r of the random environment,

transitions of the chain $m_t,\ t \geq 0$, within the state space $\{1, \ldots, M\}$ are defined by an irreducible sub-generator $S^{(r)}$ while the intensities of transition into the absorbing state are defined by the vector $\boldsymbol{S}_0^{(r)} = -S^{(r)}\mathbf{e}$. Here $\mathbf{e}$ is a column vector consisting of ones. At the service beginning epoch, the state of the process $m_t,\ t \geq 0$, is chosen according to the probabilistic row vector $\boldsymbol{\beta}^{(r)},\ r = \overline{1, R}$. It is assumed that the state of the process $m_t,\ t \geq 0$, is not changed at the epoch of the process $r_t,\ t \geq 0$, transitions. Just the exponentially distributed sojourn time of the process $m_t,\ t \geq 0$, in the current state is re-started with a new intensity defined by the sub-generator corresponding to the new state of the random environment $r_t,\ t \geq 0$.

The input flow into the system is produced by U terminals each modeled as a continuous time Batch Markovian Arrival Process (BMAP), see, e.g. [3]. So, superposition of these flows is also the BMAP. For modeling we consider a set of distinct user events to be fixed and finite. These events constitute the states of BMAP. At transitions between states the terminal produces batches of data (in bits) corresponding to defined user events. The intensity of transitions depends on the external environment process as introduced below. More formally, the arrival of customers is directed by the process $\nu_t,\ t \geq 0$, (the underlying process) with the state space $\{0, 1, \ldots, W\}$. Under the fixed state r of the RE, this process behaves as an irreducible continuous time Markov chain. Transitions of the chain $\nu_t,\ t \geq 0$, which are accompanied by arrival of k-size batch, are described by the matrices $D_k^{(r)},\ k \geq 0,\ r = \overline{1, R}$, with the generating function $D^{(r)}(z) = \sum_{k=0}^{\infty} D_k{}^{(r)} z^k,\ |z| \leq 1$. The matrix $D^{(r)}(1)$ is an irreducible generator for all $r = \overline{1, R}$. Under the fixed state r of the random environment, the average intensity λ_r (fundamental rate) of the BMAP is defined as $\lambda_r = \boldsymbol{\theta}^{(r)}(D^{(r)}(z))'|_{z=1}\mathbf{e}$, and the intensity $\lambda_r^{(b)}$ of batch arrivals is defined as $\lambda_r^{(b)} = \boldsymbol{\theta}^{(r)}(-D_0^{(r)})\mathbf{e}$. Here the row vector $\boldsymbol{\theta}^{(r)}$ is the solution to the equations $\boldsymbol{\theta}^{(r)}D^{(r)}(1) = \mathbf{0}$, $\boldsymbol{\theta}^{(r)}\mathbf{e} = 1$, $\mathbf{e}$ is a column vector of appropriate size consisting of 1's. The variation coefficient $c_{var}^{(r)}$ of intervals between batch arrivals is given by

$$(c_{var}^{(r)})^2 = 2\lambda_r^{(b)}\boldsymbol{\theta}^{(r)}(-D_0^{(r)})^{-1}\mathbf{e} - 1$$

while the correlation coefficient $c_{cor}^{(r)}$ of intervals between successive batch arrivals is calculated as

$$c_{cor}^{(r)} = (\lambda_r^{(b)}\boldsymbol{\theta}^{(r)}(-D_0^{(r)})^{-1}(D^{(r)}(1) - D_0^{(r)})(-D_0^{(r)})^{-1}\mathbf{e} - 1)/(c_{var}^{(r)})^2.$$

The necessary condition for a terminal to begin the transmission is to obtain a *scheduling grant* in response to the scheduling request message submitted to the base station prior to the scheduled period. For the purpose of modeling we assume that the system has $L,\ 0 \leq L < \infty$, *virtual* waiting positions. If the system has all servers being busy at a batch arrival epoch, the batch occupies the waiting position. Due to a possibility of the batch arrivals, it can occur that there are free servers in the system at an arrival epoch, however the number of these positions is less than the number of the customers in an arriving batch. In such

situation the acceptance of the customers to the system is realized according to the partial admission discipline (only a part of the batch corresponding to the number of free servers is allowed to enter the system while the rest of the batch is lost). The disciplines of complete rejection and complete admission are considered analogously, details are omitted here.

At the epochs of the process $r_t,\ t \geq 0$, transitions, the state of the process $\nu_t,\ t \geq 0$, does not change, but the intensities of its transitions are immediately changed. This corresponds to a situation where, for example, the intensity of blog messages, photo or vide clips uploads increases after the scoring event in a soccer game.

Our aim is calculation of the stationary state distribution and main performance measures of the described queueing model with further use for optimization of parameters of quality of service in the system.

4 Process of the System States

It is easy to see that operation of the considered queueing model can be described in terms of the regular irreducible continuous-time Markov chain

$$\xi_t = \{n_t, r_t, \nu_t, m_t^{(1)}, \ldots, m_t^{(M)}\},\ t \geq 0,$$

where

- n_t is the number of customers in the system, $n_t = \overline{0, N+L}$;
- r_t is the state of the random environment, $r_t = \overline{1, R}$;
- ν_t is the state of the BMAP underlying process, $\nu_t = \overline{0, W}$;
- $m_t^{(l)}$ be the number of servers at the phase l of service, $l = \overline{1, M}$, $m_t^{(l)} = \overline{0, \min\{n_t, N\}}$, $\sum\limits_{l=1}^{M} m_t^{(l)} = \min\{n_t, N\}$, at the epoch $t,\ t \geq 0$.

Note that the use of the components $m_t^{(l)}$ for description of service processes in multi-server queues stems from the works by Ramaswami and Lucantoni, see [5,6].

Let us enumerate the states of the chain $\xi_t,\ t \geq 0$, in the lexicographic order of components (r_t, ν_t) and the reverse lexicographic order of components $(m_t^{(1)}, \ldots, m_t^{(M)})$ and form the row vectors $\mathbf{p}_n$ of probabilities corresponding to the state n of the first component of the process $\xi_t,\ t \geq 0$. Denote also $\mathbf{p} = (\mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{N+L})$.

It is well known that the vector $\mathbf{p}$ satisfies the system of the linear algebraic equations (so called equilibrium equations or Chapman-Kolmogorov equations) of the form:

$$\mathbf{p}A = \mathbf{0},\ \mathbf{p}\mathbf{e} = 1, \tag{1}$$

where A is the infinitesimal generator of the Markov chain $\xi_t,\ t \geq 0$.

To present explicit expression for generator A, let us introduce the following notation:

- $\mathbf{e}_n$ ($\mathbf{0}_n$) is a column (row) vector of size n, consisting of 1's (0's). Suffix may be omitted if the dimension of the vector is clear from context;
- $K_n = \binom{n+M-1}{M-1}$, $n = \overline{0, N}$;
- I (O) is an identity (zero) matrix of appropriate dimension (when needed the dimension of this matrix is identified with a suffix);
- $\mathrm{diag}\{a_l,\ l = \overline{1, L}\}$ is a diagonal matrix with diagonal entries or blocks a_l;
- $\otimes$ and $\oplus$ are symbols of the Kronecker product and sum of matrices;
- $D(z) = \sum\limits_{k=0}^{\infty} \mathrm{diag}\{D_k^{(r)},\ r = \overline{1, R}\} z^k$;
- $H_k^{(n)} = \mathrm{diag}\{D_k^{(r)} \otimes I_{K_n},\ r = \overline{1, R}\},\ n = \overline{0, N},\ k \geq 0$;
- $H^{(n)}(z) = \sum\limits_{k=0}^{\infty} H_k^{(n)} z^k,\ n = \overline{0, N}$;
- $B_l^{(n)} = \mathrm{diag}\{I_{\bar{W}} \otimes P_{n,n+l}(\boldsymbol{\beta}^{(r)}),\ r = \overline{1, R}\},\ n = \overline{0, N-1},\ l = \overline{1, N-n},\ \bar{W} = W + 1$;
- $\hat{A}^{(n)} = \mathrm{diag}\{I_{\bar{W}} \otimes \tilde{A}_n(N, S^{(r)}),\ r = \overline{1, R}\},\ n = \overline{1, N}$;
- $\Delta^{(n)} = -\mathrm{diag}\{I_{\bar{W}} \otimes \mathrm{diag}\{\tilde{A}_n(N, S^{(r)})\mathbf{e} + L_{N-n}(N, \tilde{S}^{(r)})\mathbf{e}\},\ r = \overline{1, R}\},\ n = \overline{1, N}$;
- $\tilde{S}^{(r)} = \begin{pmatrix} 0 & \mathbf{0} \\ \boldsymbol{S}_0^{(r)} & S^{(r)} \end{pmatrix},\ r = \overline{1, R}$;
- $L^{(n)} = \mathrm{diag}\{I_{\bar{W}} \otimes L_{N-n}(N, \tilde{S}^{(r)}),\ r = \overline{1, R}\},\ n = \overline{1, N}$;
- $\bar{L} = \mathrm{diag}\{I_{\bar{W}} \otimes L_0(N, \tilde{S}^{(r)}) P_{N-1}(\boldsymbol{\beta}^{(r)}),\ r = \overline{1, R}\}$;
- $C^{(n)} = Q \otimes I_{\bar{W}} \otimes I_{K_n} + H_0^{(n)} + \hat{A}^{(n)} + \Delta^{(n)},\ n = \overline{0, N}$;
- $\bar{C} = Q \otimes I_{\bar{W}} \otimes I_{K_N} + \sum\limits_{k=0}^{\infty} H_k^{(N)} + \hat{A}^{(N)} + \Delta^{(N)}$;
- $P_{n,l}(\boldsymbol{\beta}^{(r)}) = P_n(\boldsymbol{\beta}^{(r)}) \times \cdots \times P_{l-1}(\boldsymbol{\beta}^{(r)}),\ n = \overline{0, N-1},\ l = \overline{n+1, N}$;
- $E_k^{(n)} = H_k^{(n)} B_{\min\{\mathrm{N-n,k}\}}^{(n)},\ n = \overline{0, N-1},\ k = \overline{1, N+L-1}$;
- $\hat{E}_k^{(n)} = \sum\limits_{l=k}^{\infty} E_l^{(n)},\ n = \overline{0, N-1},\ k \geq 0$;
- $\hat{H}_k = \sum\limits_{l=k}^{\infty} H_l^{(N)},\ k = \overline{1, L}$.

The algorithms for calculating the matrices $P_n(\boldsymbol{\beta}^{(r)})$, $n = \overline{0, N-1}$, $\tilde{A}_n(N, S^{(r)})$, $n = \overline{0, N}$, and $L_{N-n}(N, \tilde{S}^{(r)})$, $n = \overline{0, N}$, $r = \overline{1, R}$, are presented in Appendix of paper [7].

Lemma 1. Infinitesimal generator A of the Markov chain $\xi_t,\ t \geq 0$, has the following block structure:

$$A = (A_{n,n'})_{n,n'=\overline{0,N+L}} =$$

$$= \begin{pmatrix} C^{(0)} & E_1^{(0)} & \cdots & E_{N-1}^{(0)} & E_N^{(0)} & E_{N+1}^{(0)} & \cdots & E_{N+L-1}^{(0)} & \hat{E}_{N+L}^{(0)} \\ L^{(1)} & C^{(1)} & \cdots & E_{N-2}^{(1)} & E_{N-1}^{(1)} & E_N^{(1)} & \cdots & E_{N+L-2}^{(1)} & \hat{E}_{N+L-1}^{(1)} \\ O & L^{(2)} & \cdots & E_{N-3}^{(2)} & E_{N-2}^{(2)} & E_{N-1}^{(2)} & \cdots & E_{N+L-3}^{(2)} & \hat{E}_{N+L-2}^{(2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & \cdots & C^{(N-1)} & E_1^{(N-1)} & E_2^{(N-1)} & \cdots & E_L^{(N-1)} & \hat{E}_{L+1}^{(N-1)} \\ O & O & \cdots & L^{(N)} & C^{(N)} & H_1^{(N)} & \cdots & H_{L-1}^{(N)} & \hat{H}_L \\ O & O & \cdots & O & \bar{L} & C^{(N)} & \cdots & H_{L-2}^{(N)} & \hat{H}_{L-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & \cdots & O & O & O & \cdots & C^{(N)} & \hat{H}_1 \\ O & O & \cdots & O & O & O & \cdots & \bar{L} & \bar{C} \end{pmatrix}.$$

Proof of the Lemma follows from analysis of Markov chain $\xi_t, t \geq 0$, transitions during an infinitesimal interval. Block entries of the generator have the following meaning.

The non-diagonal entries of the matrices $C^{(n)}$, $n = \overline{0, N}$, and $\bar{C}$ define intensity of transition of the components $\{r_t, \nu_t, m_t^{(1)}, \ldots, m_t^{(M)}\}$ of the Markov chain $\xi_t, t \geq 0$, which do not lead to the change of the number n of customers in the system. The diagonal entries of the matrices $C^{(n)}$, $n = \overline{0, N}$, and $\bar{C}$ are negative and define, up to the sign, intensity of leaving the corresponding states of the Markov chain $\xi_t,\ t \geq 0$.

The entries of the matrices $E_k^{(n)}$ define intensity of transitions of the components $\{r_t, \nu_t, m_t^{(1)}, \ldots, m_t^{(M)}\}$ of the Markov chain $\xi_t,\ t \geq 0$, which are accompanied by arrival of k customers conditioned on the fact that the number of busy servers at arrival epoch is n.

The entries of the matrices $\hat{E}_k^{(n)}$ define intensity of transitions of the components $\{r_t, \nu_t, m_t^{(1)}, \ldots, m_t^{(M)}\}$ of the Markov chain $\xi_t,\ t \geq 0$, which are accompanied by arrival of more then $k-1$ customers conditioned on the fact that the number of busy servers at arrival epoch is n.

The entries of the matrices $H_k^{(N)}, k = \overline{1, L-1}$, $(\hat{H}_l,\ l = \overline{1, L},)$ define the intensity of transitions which are accompanied by arrival of k (not less than l) customers at the epoch when all servers are busy.

The entries of the matrices $L^{(n)}$, $n = \overline{1, N}$, define intensity of transitions, which are accompanied by a departure of a customer, conditioned on the fact that the number of busy servers is n and there is no customers in the buffer.

The entries of the matrix $\bar{L}$ define intensity of transitions, which are accompanied by a departure of a customer, conditioned on the fact that there are customers in the buffer.

To solve system (1) with the matrix A defined by Lemma 1, we use the effective numerically stable procedure developed in [8] that exploits the special structure of the matrix A (it is upper block Hessenbergian) and probabilistic meaning of the unknown vector $\mathbf{p}$.

5 Performance Measures

Having the probability vector $\mathbf{p}$ been computed, we are able to calculate performance measures of the considered model. The main performance measure in the case of a finite buffer is the probability P_{loss} that an arbitrary customer will be lost due to the buffer overflow (the loss probability).

The loss probability P_{loss} is calculated as follows

$$P_{loss} = 1 - \frac{1}{\lambda} \sum_{n=0}^{N+L-1} \sum_{k=0}^{N+L-n} (k+n-N-L)\mathbf{p}_n H_k^{(\min\{n,N\})}\mathbf{e}, \tag{2}$$

where

$$\lambda = \tilde{\boldsymbol{\theta}} D(z)|'_{z=1}\mathbf{e}, \tag{3}$$

and $\tilde{\boldsymbol{\theta}}$ is the unique solution to the following system of linear algebraic equations

$$\tilde{\boldsymbol{\theta}}(Q \otimes I_{\bar{W}} + D(1)) = \mathbf{0},\ \tilde{\boldsymbol{\theta}}\mathbf{e} = 1.$$

The loss probability of an arbitrary customer that arrives to the system when the random environment is staying in the state r is calculated by formula

$$P_{loss}^{(r)} = 1 - \frac{1}{\lambda^{(r)}} \sum_{n=0}^{N+L-1} \sum_{k=0}^{N+L-n} (k+n-N-L)\tilde{\mathbf{p}}_n^{(r)} H_k^{(\min\{n,N\})}\mathbf{e} \tag{4}$$

where

$$\lambda^{(r)} = \sum_{n=0}^{L+N} \sum_{k=1}^{\infty} k\tilde{\mathbf{p}}_n^{(r)} H_k^{(\min\{n,N\})}\mathbf{e},\ r = \overline{1,R},$$

$$\tilde{\mathbf{p}}_n^{(r)} = \mathbf{p}_n(\mathbf{e}^{(r)} \otimes I_{\bar{W}} \otimes I_{K_{\min\{n,N\}}}),$$

and $\mathbf{e}^{(r)}$, $r = \overline{1,R}$, is a column vector of size R with all zero entries except the entry r which is equal to one.

Other important system performance measures are calculated as

- The mean number of customers in the system

$$L_{system} = \sum_{n=1}^{N+L} n\mathbf{p}_n\mathbf{e};$$

- The mean number of busy servers

$$N_{busy} = \sum_{n=1}^{L+N} \min\{n,N\}\mathbf{p}_n\mathbf{e};$$

- The mean number of idle servers

$$N_{idle} = N - N_{busy};$$

- The probability P_{imm} that an arbitrary customer will enter the service immediately upon arrival (without visiting a buffer)

$$P_{imm} = \lambda^{-1} \sum_{n=0}^{N-1} \sum_{k=0}^{N-n} (k+n-N)\mathbf{p}_n H_k^{(n)}\mathbf{e}.$$

6 Sojourn Time Distribution

Let $v(s)$, $Re\ s > 0$, be the Laplace-Stieltjes transform of the sojourn time distribution and $\bar{v}$ be the mean sojourn time of the arbitrary customer in the system.

Theorem 1. *The Laplace-Stieltjes transform $v(s)$ of the sojourn time distribution of an arbitrary customer in the system is calculated as follows*

$$v(s) = P_{loss} + \frac{1}{\lambda}\{\sum_{n=0}^{N-1}\sum_{k=1}^{\infty}\min\{k, N-n\}\mathbf{p}_n H_k^{(n)}(I_R \otimes \mathbf{e}_{\bar{W}K_n}) + \tag{5}$$

$$+ \sum_{n=0}^{N+L-1}\{\sum_{k=\max\{1,N-n+1\}}^{N+L-n}\mathbf{p}_n H_k^{(\min\{n,N\})}B^{(\max\{0,N-n\})}\times$$

$$\times \sum_{l=\max\{1,N-n+1\}}^{k}(F(s))^{n-N+l}(I_R \otimes \mathbf{e}_{K_N}) +$$

$$+ \sum_{k=N+L-n+1}^{\infty}\mathbf{p}_n H_k^{(\min\{n,N\})}B^{(\max\{0,N-n\})}\times$$

$$\times \sum_{l=\max\{1,N-n+1\}}^{N+L-n}(F(s))^{n-N+l}(I_R \otimes \mathbf{e}_{K_n})\}\mathcal{H}(s)\mathbf{e}_R,$$

where

$$\mathcal{H}(s) = \mathrm{diag}\{\boldsymbol{\beta}^{(r)},\ r = \overline{1,R}\}(sI - (Q \otimes I_M + \mathcal{S}))^{-1}\mathrm{diag}\{\mathbf{S}_0^{(r)},\ r = \overline{1,R}\},$$

$$F(s) = (sI - (Q \otimes I_{K_N} + \mathrm{diag}\{\tilde{A}_N(N, S^{(r)}) -$$

$$-\mathrm{diag}\{\tilde{A}_N(N, S^{(r)})\mathbf{e} + L_0(N, \tilde{S}^{(r)})\mathbf{e}\},\ r = \overline{1,R}\}))^{-1}\times$$

$$\times\mathrm{diag}\{L_0(N, \tilde{S}^{(r)})P_{N-1}(\boldsymbol{\beta}^{(r)}),\ r = \overline{1,R}\},$$

$$B^{(n)} = \mathrm{diag}\{\mathbf{e}_{\bar{W}} \otimes P_{N-n,N}(\boldsymbol{\beta}^{(r)}),\ r = \overline{1,R}\},\ n = \overline{1,N},$$

$$B^{(0)} = \mathrm{diag}\{\mathbf{e}_{\bar{W}} \otimes I_{K_N},\ r = \overline{1,R}\}$$

$$\mathcal{S} = \mathrm{diag}\{S^{(r)},\ r = \overline{1,R}\}.$$

Remark 1. The Laplace-Stieltjes transform $v^{(r)}(s)$ of the sojourn time distribution of an arbitrary customer, which arrives to the system when the state of random environment is r, can be calculated by means of formula (5) via replacing the values P_{loss}, $\mathbf{p}_n$, and λ by the quantities $P_{loss}^{(r)}$, $\mathbf{p}_n^{(r)}$, and $\lambda^{(r)}$, correspondingly.

Theorem 2. *The mean sojourn time v_1 of an arbitrary customer in the system is calculated by*

$$v_1 = -\frac{1}{\lambda}\Bigg\{\sum_{n=0}^{N-1}\sum_{k=1}^{\infty}\min\{k,N-n\}\mathbf{p}_n H_k^{(n)}(I_R\otimes\mathbf{e}_{\bar{W}K_n})\mathcal{H}'(0)\mathbf{e}_R+ \tag{6}$$

$$+\sum_{n=0}^{N+L-1}\Bigg[\sum_{k=\max\{1,N-n+1\}}^{N+L-n}\mathbf{p}_n H_k^{(\min\{n,N\})}B^{(\max\{0,N-n\})}\times$$

$$\times\sum_{l=\max\{1,N-n+1\}}^{k}\sum_{m=0}^{n+l-N-1}(F(0))^m F'(0)(F(0))^{n+l-N-1-m}(I_R\otimes\mathbf{e}_{K_N})+$$

$$+\sum_{k=N+L-n+1}^{\infty}\mathbf{p}_n H_k^{(\min\{n,N\})}B^{(\max\{0,N-n\})}\times$$

$$\times\sum_{l=\max\{1,N-n+1\}}^{N+L-n}\sum_{m=0}^{n-N+l-1}(F(0))^m F'(0)(F(0))^{n+l-N-1-m}(I_R\otimes\mathbf{e}_{K_N})\Bigg]\mathbf{e}_R+$$

$$+\sum_{n=0}^{N+L-1}\Bigg\{\sum_{k=\max\{1,N-n+1\}}^{N+L-n}\mathbf{p}_n H_k^{(\min\{n,N\})}B^{(\max\{0,N-n\})}\times$$

$$\times\sum_{l=\max\{1,N-n+1\}}^{k}(F(0))^{n-N+l}(I_R\otimes\mathbf{e}_{K_N})+$$

$$+\sum_{k=N+L-n+1}^{\infty}\mathbf{p}_n H_k^{(\min\{n,N\})}B^{(\max\{0,N-n\})}\times$$

$$\times\sum_{l=\max\{1,N-n+1\}}^{N+L-n}(F(0))^{n-N+l}(I_R\otimes\mathbf{e}_{K_N})\Bigg\}\mathcal{H}'(0)\mathbf{e}_R\Bigg\},$$

where

$$F'(0) = -[Q\otimes I_{K_N} + \mathrm{diag}\{\tilde{A}_N(N,S^{(r)}) - \mathrm{diag}\{\tilde{A}_N(N,S^{(r)})\mathbf{e}+$$

$$+L_0(N,\tilde{S}^{(r)})\mathbf{e}\},\ r=\overline{1,R}\}]^{-2}\mathrm{diag}\{L_0(N,\tilde{S}^{(r)})P_{N-1}(\boldsymbol{\beta}^{(r)}),\ r=\overline{1,R}\},$$

$$\mathcal{H}'(0) = -\mathrm{diag}\{\boldsymbol{\beta}^{(r)},\ r=\overline{1,R}\}[Q\otimes I_M+\mathcal{S}]^{-2}\mathrm{diag}\{\mathbf{S}_0^{(r)},\ r=\overline{1,R}\}.$$

To get v_1 we differentiate (5) at the point $s=0$ and use the formula $v_1 = -v'(0)$.

Remark 2. The average sojourn time $v_1^{(r)}$ of an arbitrary customer, which arrives to the system when the state of random environment is r, can be calculated by formula (6) where notations $\mathbf{p}_n$ and λ are replaced with $\mathbf{p}_n^{(r)}$ and $\lambda^{(r)}$, correspondingly.

Calculation of higher moments of the sojourn time distribution is based on formula

$$v_m = (-1)^m \frac{d^m v(s)}{ds^m}|_{s=0},\ m \geq 1, \tag{7}$$

for an arbitrary customer and formula

$$v_m^{(r)} = (-1)^m \frac{d^m v^{(r)}(s)}{ds^m}|_{s=0} \tag{8}$$

for an arbitrary customer which arrives to the system when the state of random environment is r.

Using formulas (7) and (8), the second moments v_2 and $v_2^{(r)}$ of the sojourn time distribution can be found taking into account that

$$\mathcal{H}''(0) = 2\mathrm{diag}\{\boldsymbol{\beta}^{(r)},\ r = \overline{1,R}\}[Q \otimes I_M + \mathcal{S}]^{-3}\mathrm{diag}\{\mathbf{S}_0^{(r)},\ r = \overline{1,R}\},$$

$$F^{''}(0) = 2[Q \otimes I_{K_N} + \mathrm{diag}\{\tilde{A}_N(N, S^{(r)}) - \mathrm{diag}\{\tilde{A}_N(N, S^{(r)})\mathbf{e}+$$

$$+L_0(N, \tilde{S}^{(r)})\mathbf{e}\},\ r = \overline{1,R}\}]^{-3}\mathrm{diag}\{L_0(N, \tilde{S}^{(r)})P_{N-1}(\boldsymbol{\beta}^{(r)}),\ r = \overline{1,R}\},$$

$$[F^k(s)]'|_{s=0} = \sum_{m=0}^{k-1}(F(0))^m F'(0)(F(0))^{k-m-1},$$

$$[F^k(s)]''|_{s=0} = \sum_{m=0}^{k-1}\Bigg[\sum_{j=0}^{m-1}(F(0))^j F'(0)(F(0))^{m-1-j}F'(0)(F(0))^{k-m-1}+$$

$$+(F(0))^m F''(0)(F(0))^{k-m-1}+(F(0))^m F'(0)\sum_{u=0}^{k-m-2}(F(0))^u F'(0)(F(0))^{k-m-2-u}\Bigg].$$

Having known two first moments $v_1^{(r)}$ and $v_2^{(r)}$ of sojourn time distribution of a customer arrived at the moment when the state of random environment is r one can approximate the density of this distribution, e.g., by function

$$v^{(r)}(t) = q^{(r)}\frac{\mu_1^{(r)}(\mu_1^{(r)}t)^{k_1^{(r)}-1}}{(k_1^{(r)}-1)!}e^{-\mu_1^{(r)}t} + (1-q^{(r)})\frac{\mu_2^{(r)}(\mu_2^{(r)}t)^{k_2^{(r)}-1}}{(k_2^{(r)}-1)!}e^{-\mu_2^{(r)}t}$$

where unknown parameters $q^{(r)}$, $\mu_i^{(r)}$, $0 \leq q^{(r)} \leq 1$, $\mu_i^{(r)} \geq 0$, and positive integers $k_i^{(r)}, i = 1, 2$, should be found from the relations

$$v_1^{(r)} = q^{(r)}\frac{k_1^{(r)}}{\mu_1^{(r)}} + (1-q^{(r)})\frac{k_2^{(r)}}{\mu_2^{(r)}},$$

$$v_2^{(r)} = q^{(r)}\frac{k_1^{(r)}(k_1^{(r)}-1)}{(\mu_1^{(r)})^2} + (1-q^{(r)})\frac{k_2^{(r)}(k_2^{(r)}-1)}{(\mu_2^{(r)})^2}.$$

Let now $T^{(r)}$ be the guaranteed by an operator value of sojourn time of an arbitrary customer that arrived at the moment when the state of random environment is r, $r = \overline{1, R}$. Assume that, according to a service level agreement, actual sojourn time of an arbitrary customer that arrived at the moment when the state of random environment is r may be greater than $T^{(r)}$ only with small probability $\epsilon^{(r)}$, $r = \overline{1, R}$. Equation, which matches the values of $T^{(r)}$ and $\epsilon^{(r)}$, has the following form:

$$\int\limits_{T^{(r)}}^{\infty} \left(q^{(r)} \frac{\mu_1^{(r)} (\mu_1^{(r)} t)^{k_1^{(r)}-1}}{(k_1^{(r)}-1)!} e^{-\mu_1^{(r)} t} + (1-q^{(r)}) \frac{\mu_2^{(r)} (\mu_2^{(r)} t)^{k_2^{(r)}-1}}{(k_2^{(r)}-1)!} e^{-\mu_2^{(r)} t} \right) dt = \epsilon^{(r)}.$$

This equation along with formulas for loss probabilities $P_{loss}^{(r)}$ can be considered as a mathematical background for fixing guaranteed level of service to customers arrived to the system under various states of the random environment.

7 Conclusion

The $BMAP/PH/N/L$ system operating in random environment as a model of service providing by mobile operator in varying external conditions is investigated. Variation of conditions may randomly occur at moments when some events having social importance or events relating to sport competitions or disasters occur. The joint stationary distribution of the number of the customers in the system, the state of the random environment, and the states of the underlying processes of arrival and service processes is calculated. The analytic formulas for some performance measures of the system are derived. Laplace-Stieltjes transform of sojourn time distribution is derived and the mean sojourn time is calculated for an arbitrary customer and an arbitrary customer that arrived at the moment when the state of random environment is r, $r = \overline{1, R}$. Formulas matching admissible probabilities of excess of a given level of sojourn time in the system with different levels are presented. The obtained results can be used as a base for computing reasonable parameters of the system operation which should be fixed when service level agreement between an user and operator is prepared. In contrast to existing methods, our methodology takes into account heterogeneous character of information flows and their burstyness and also pre-assumes that quality of service may temporarily become worse due to some external events. Different values of indicators of quality of service may be fixed for normal and abnormal situations.

Aspects relating to possible retrials of customers who did not get immediate access to the system upon arrivals can be treated by analogy with [9].

Acknowledgments. This work was supported by COST IC0906 WiNeMO (Wireless Networking for Moving Objects).

References

1. Kim, C., Dudin, A., Klimenok, V., Khramova, V.: Performance analysis of multi-server queueing system operation under control of a random environment. Trends in Telecommunications Technologies, 315–344 (2010)
2. Dahlman, E., Parkvall, S., Sköld, J.: 4G: LTE/LTE-Advanced for Mobile Broadband. Elsevier (2011) ISBN: 978-0-12-385489-6,
3. Lucantoni, D.M.: New results on the single server queue with a batch Markovian arrival process. Comm. Statist.-Stochastic Models 7, 1–46 (1991)
4. Neuts, M.F.: Matrix-geometric solutions in stochastic models. The Johns Hopkins University Press (1981)
5. Ramaswami, V.: Independent Markov processes in parallel. Comm. Statist.-Stochastic Models 1, 419–432 (1985)
6. Ramaswami, V., Lucantoni, D.: Algorithms for the multi-server queue with phase-type service. Comm. Statist.-Stochastic Models 1, 393–417 (1985)
7. Kim, Ch., Dudin, S., Taramin, O., Baek, J.: Queueing system MAP/PH/N/N + R with impatient heterogeneous customers as a model of call center. Applied Mathematical Modelling (2012) dx.doi.org/10.1016/j.apm.2012.03.021
8. Kim, C., Klimenok, V., Orlovsky, D., Dudin, A.: Lack of invariant property of the Erlang loss model in case of MAP input. Queueing Systems 49, 187–213 (2005)
9. Kim, C., Klimenok, V., Mushko, V., Dudin, A.: The $BMAP/PH/N$ retrial queueing system operating in Markovian random environment. Computers and Operations Research 37, 1228–1237 (2010)

Queueing System $MAP/M/N/N+K$ Operating in Random Environment as a Model of Call Center

Olga Dudina* and Sergey Dudin

Belarusian State University, 4, Nezavisimosti Ave., Minsk, 220030, Belarus
dudina_olga@email.com, dudin@madrid.com

Abstract. A multi-server queueing system with a Markovian Arrival Process (MAP), a finite buffer and impatient customers operating in random environment as a model of a call center is investigated. The service time of a customer by a server has an exponential distribution. If all servers are busy at a customer arrival epoch, the customer may leave the system forever or move to the buffer with probability that depends on the number of customers in the buffer. During a waiting period, a customer can be impatient and can leave the system without the service. System parameters depend on the state of the random environment. An efficient algorithm for calculating the stationary probabilities of system states is proposed. Some key performance measures are calculated. The Laplace-Stieltjes transforms of the sojourn and waiting time distributions are derived.

Keywords: call center, Markovian arrival process, random environment, impatient customer.

1 Introduction

Most major companies use call centers for interaction with their customers. According to the latest research, almost all large companies have at least one call center. Since call centers are at the front line of customer service, the companies that value their customers have to provide good call center performance. To describe the operation and improve the performance of call centers queueing theory is used. Adequate mathematical modeling the call centers can substantially increase their economic efficiency and improves the quality of the customers' service. For the references and the present state-of-art in investigation of call centers the reader is referred to the survey [1], the papers [2], [3] and the references therein.

The models of call centers in the overwhelming majority of existing papers assume that the arrival flow of customers is described by a stationary Poisson arrival process. This assumption greatly simplifies the study of systems, but at

* Corresponding author.

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 83–92, 2013.

the same time reduces the adequacy of the model, because arrival flows in most of modern telecommunication networks are correlated.

The model of call center with Markovian Arrival Process (MAP) is considered in [4]. MAP is useful for modelling the arrival flows that do not possess the properties of stationarity, memory less and ordinarity.

However, even the consideration of MAP arrival flow not always well suits for modelling the real arrival flow. MAP arrival flow takes into account an effect of correlation and inter-arrival times variation but the MAP process is fixed in a border of considered model. At the same time the arrival flow of customers can change it characteristics (average arrival rate, coefficient of correlation, variation coefficient, etc.) depending on some random factors. Moreover, the random factors can also impact on other system parameters such as the intensity of impatience, service time distribution, etc. To take into account the influence of random factors on the system parameters the queueing systems operating in *random environment* are considered. Under consideration queues in random environment it is assumed that there is a finite state stochastic process independent on queueing system called as random environment. Under the fixed state of the random environment the queueing system operates as a classical queueing system. However the system parameters (arrival process, service time distribution, intensity of impatience, etc.) are immediately changed with change the state of random environment. For the references and the present state-of-art in investigation of the queueing systems operating in random environment the reader is referred to the papers [5], [6] and the references therein.

To the best of our knowledge, the models of call centers as the queueing system operating in random environment are not considered in literature previously despite on their practical importance. In this paper, we deal with a multi-server queueing system with the MAP process, a finite buffer and impatient customers operating in random environment. We calculate the stationary distribution of the system states and derive the Laplace-Stieltjes transform of the sojourn and waiting time distributions. The formulas for some system performance measures are obtained.

2 Mathematical Model

We consider a queueing system to model a call center with N operators (servers) and a finite waiting space (buffer) of capacity K. The behavior of the system depends on the state of random environment. Random environment is given by the stochastic process $r_t,\ t \geq 0$, which is an irreducible continuous time Markov chain with the state space $\{1, 2, \ldots, R\}$ and the infinitesimal generator H.

The customers arrive to the system according to MAP. That means the following. The arrival of customers is directed by the stochastic process $\nu_t,\ t \geq 0$, with the state space $\{0, 1, \ldots, W\}$. Under the fixed state r of random environment this process is an irreducible continuous time Markov chain. The sojourn time of this chain in the state ν is exponentially distributed with the positive finite parameter $\lambda_\nu^{(r)}$. When the sojourn time in the state ν expires, with

probability $p_0^{(r)}(\nu,\nu')$ the process ν_t jumps to the state ν' without generation of customers, $\nu,\nu'=\overline{0,W},\ \nu\neq\nu'$, $r=\overline{1,R}$ and with probability $p_1^{(r)}(\nu,\nu')$ the process ν_t jumps to the state ν' with generation of a customer, $\nu,\nu'=\overline{0,W}$, $r=\overline{1,R}$.

The behavior of the MAP is completely characterized by the matrices $D_0^{(r)}$, $D_1^{(r)}$ defined by entries

$$(D_0^{(r)})_{\nu,\nu}=-\lambda_\nu^{(r)},\nu=\overline{0,W},\ (D_0^{(r)})_{\nu,\nu'}=\lambda_\nu^{(r)}p_0^{(r)}(\nu,\nu'),\ \nu,\nu'=\overline{0,W},\nu\neq\nu',$$

$$(D_1^{(r)})_{\nu,\nu'}=\lambda_\nu^{(r)}p_1^{(r)}(\nu,\nu'),\nu,\nu'=\overline{0,W},\ r=\overline{1,R}.$$

The matrix $D^{(r)}(1)=D_0^{(r)}+D_1^{(r)}$ represents the generator of the process $\nu_t, t\geq 0$, under the fixed $r=\overline{1,R}$.

The average arrival rate $\lambda^{(r)}$ under the fixed state r of random environment is given as

$$\lambda^{(r)}=\boldsymbol{\theta}^{(r)}D_1^{(r)}\mathbf{e}$$

where $\boldsymbol{\theta}^{(r)}$ is the invariant vector of a stationary distribution of the Markov chain ν_t, $t\geq 0$, under the fixed state r. The vector $\boldsymbol{\theta}^{(r)}$ is the unique solution to the system $\boldsymbol{\theta}^{(r)}D^{(r)}(1)=\mathbf{0}$, $\boldsymbol{\theta}^{(r)}\mathbf{e}=1$. Here $\mathbf{e}$ is a column vector of appropriate size consisting of 1's and $\mathbf{0}$ is a row vector of appropriate size consisting of zeroes.

The squared coefficient of variation $c_{var}^{(r)}$ of intervals between successive arrivals is given as

$$c_{var}^{(r)}=2\lambda^{(r)}\boldsymbol{\theta}^{(r)}(-D_0^{(r)})^{-1}\mathbf{e}-1.$$

The coefficient of correlation $c_{cor}^{(r)}$ of two successive intervals between arrivals is given as

$$c_{cor}^{(r)}=(\lambda^{(r)}\boldsymbol{\theta}^{(r)}(-D_0^{(r)})^{-1}D_1^{(r)}(-D_0^{(r)})^{-1}\mathbf{e}-1)/c_{var}^{(r)}.$$

More information about the MAP and related research is given, e.g., in [7], [8].

At the epochs of transitions of process r_t, $t\geq 0$, the states of the process ν_t, $t\geq 0$, do not change, only the intensities of transition of this process change.

We also suggest that the call center queue is "visible" (see, e.g., [9]), which means the following. An arriving customer, who cannot enter into service immediately, is informed about the queue length. The customer then decides either to leave the system immediately due to the length of the queue is inadmissible or join the queue.

So, in mathematical model we assume that if at an arbitrary customer arrival epoch there is a free server, a customer is admitted to the system and occupies the free server.

If at a customer arrival epoch all servers are busy and i, $i=\overline{0,K-1}$, customers are presenting in the buffer then this customer leaves the system with probability $q_i^{(r)}$ under the fixed state r of random environment or moves to the buffer with supplementary probability.

If at an arbitrary customer arrival epoch the buffer is full, the customer leaves the system forever.

The customers can be impatient, i.e., under the fixed state r of random environment the customer leaves the system after arrival in random time, which is exponentially distributed with the parameter $\alpha^{(r)}$, $0 < \alpha^{(r)} < \infty$.

The service time (talk time and after talk work time) of a customer by each server has an exponential distribution. Under the fixed state of random environment r the service time is exponentially distributed with the parameter μ_r, $r = \overline{1, R}$.

3 The Process of System States

Let i_t be the number of customers in the system, $i_t = \overline{0, N+K}$, r_t be the state of random environment, $r_t = \overline{1, R}$, and ν_t be the state of the directing process of the MAP, $\nu_t = \overline{0, W}$.

So, the behavior of the system under consideration can be described in terms of the regular irreducible continuous-time Markov chain

$$\xi_t = \{i_t, r_t, \nu_t\},\ t \geq 0.$$

Since the Markov chain ξ_t is regular irreducible and has a finite state space, then for any choice of the system parameters there exist stationary probabilities of the system states which are defined as follows:

$$\pi(i, r, \nu) = \lim_{t \to \infty} P\{i_t = i,\ r_t = r,\ \nu_t = \nu\},\ i = \overline{0, N+K},\ r = \overline{1, R},\ \nu = \overline{0, W}.$$

Then let us form the row vectors $\boldsymbol{\pi}_i$:

$$\boldsymbol{\pi}(i, r) = (\pi(i, r, 0), \pi(i, r, 1), \ldots, \pi(i, r, W)),\ r = \overline{1, R},$$

$$\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i, 1), \boldsymbol{\pi}(i, 2), \ldots, \boldsymbol{\pi}(i, R)),\ i = \overline{0, N+K}.$$

It is well-known that the probability vectors $\boldsymbol{\pi}_i$, $i = \overline{0, N+K}$, satisfy the following system of linear algebraic equations:

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{N+K})Q = \mathbf{0},\ \ (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{N+K})\mathbf{e} = 1 \tag{1}$$

where Q is the infinitesimal generator of the Markov chain $\xi_t,\ t \geq 0$.

Lemma 1. *The infinitesimal generator Q of the Markov chain $\xi_t,\ t \geq 0$, has the block-three-diagonal structure:*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & \ldots & O & O \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & \ldots & O & O \\ O & Q_{2,1} & Q_{2,2} & \ldots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \ldots & Q_{N+K-1,N+R-1} & Q_{N+K-1,N+K} \\ O & O & O & \ldots & Q_{N+K,N+K-1} & Q_{N+K,N+K} \end{pmatrix}.$$

The non-zero blocks $Q_{i,j}$, $i, j \geq 0$, *have the following form:*

$$Q_{i,i} = \tilde{D}_0 + \delta_{i,N}\bar{Q}_{i-N}\tilde{D}_1 + [H - iA] \otimes I_{\bar{W}},\ i = \overline{0, N},$$

$$Q_{i,i} = \tilde{D}_0 + \bar{Q}_{i-N}\tilde{D}_1 + [H - NA - (i-N)E] \otimes I_{\bar{W}},\ i = \overline{N+1, N+K-1},$$

$$Q_{N+K,N+K} = \tilde{D}_0 + \tilde{D}_1 + [H - NA - KE] \otimes I_{\bar{W}},$$

$$Q_{i,i-1} = iA \otimes I_{\bar{W}},\ i = \overline{1, N},$$

$$Q_{i,i-1} = [NA + (i-N)E] \otimes I_{\bar{W}},\ i = \overline{N+1, N+K},$$

$$Q_{i,i+1} = \tilde{D}_1,\ i = \overline{0, N-1},$$

$$Q_{i,i+1} = (I - \bar{Q}_{i-N})\tilde{D}_1,\ i = \overline{N, N+K-1},$$

where

- I *is an identity matrix,* O *is a zero matrix of appropriate dimension;*
- $\otimes$ *is the symbol Kronecker's product, see, e.g., [10];*
- $\bar{W} = W + 1$;
- $\tilde{D}_l = \mathrm{diag}\{D_l^{(r)},\ r = \overline{1, R}\},\ l = 0, 1$;
- $A = \mathrm{diag}\{\mu_r,\ r = \overline{1, R}\}$;
- $\bar{Q}_k = \mathrm{diag}\{q_k^{(r)},\ r = \overline{1, R}\} \otimes I_{\bar{W}},\ k = \overline{0, K-1}$;
- $E = \mathrm{diag}\{\alpha^{(r)},\ r = \overline{1, R}\}$.

The proof of Lemma 1 is implemented by means of the analysis of all transitions of the Markov chain $\xi_t, t \geq 0$, during the interval of an infinitesimal length and rewriting intensities of these transitions into the block matrix form.

If the dimension of the system (1) is small, it can be easily solved on a computer by standard methods. Otherwise, to solve this system the following numerically stable algorithm can be used.

Theorem 1. *The vectors* $\boldsymbol{\pi}_i$, $i = \overline{0, N+K}$, *are given as follows*

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_{i-1}T_{i-1} = \boldsymbol{\pi}_0 F_i, i = \overline{1, N+K},$$

where the matrices F_i *are calculated using the recurrent formulas:*

$$F_0 = I,\ F_i = F_{i-1}T_{i-1},\ i = \overline{1, N+K},$$

the matrices $T_i, i = \overline{0, N+K-1}$, *are calculated using the backward recursion*

$$T_i = -Q_{i,i+1}(Q_{i+1,i+1} + T_{i+1}Q_{i+2,i+1})^{-1},\ \ i = N+K-2, N+K-3, \ldots, 0,$$

under the initial condition

$$T_{N+K-1} = -Q_{N+K-1,N+K}(Q_{N+K,N+K})^{-1},$$

the vector $\boldsymbol{\pi}_0$ *is the unique solution to the system*

$$\boldsymbol{\pi}_0(Q_{0,0} + T_0 Q_{1,0}) = \mathbf{0}, \quad \boldsymbol{\pi}_0 \sum_{l=0}^{N+K} F_l \mathbf{e} = 1.$$

The numerical stability of the proposed algorithm follows from the fact, that all inverted matrices computed in this algorithm are irreducible sub-generators. So, as it is well known from the matrix theory, these inverse matrices exist and are non-negative.

4 Performance Measures

As soon as the vectors $\boldsymbol{\pi}_i,\ i = \overline{0, N+K}$, have been calculated, we are able to find various performance measures of the system (call center) under consideration.

The stationary distribution of the number of customers in the system is given as

$$\lim_{t\to\infty} P\{i_t = i\} = \boldsymbol{\pi}_i \mathbf{e},\ i = \overline{0, N+K}.$$

The average number of customers in the system is calculated as

$$\tilde{L} = \sum_{i=1}^{N+K} i\boldsymbol{\pi}_i \mathbf{e}.$$

The average number of customers in the buffer is given as

$$N^{buffer} = \sum_{i=N+1}^{N+K} (i-N)\boldsymbol{\pi}_i \mathbf{e}.$$

The average number of busy servers is computed by

$$N^{server} = \sum_{i=1}^{N+K} \min\{i, N\}\boldsymbol{\pi}_i \mathbf{e}.$$

The loss probability of an arbitrary customer at the entrance to the call center due to buffer overflow is given as

$$P^{ent-loss} = \lambda^{-1}\boldsymbol{\pi}_{N+K}\tilde{D}_1\mathbf{e}$$

where the average arrival rate λ is calculated as follows

$$\lambda = \boldsymbol{\theta}\tilde{D}_1\mathbf{e},$$

and the vector $\boldsymbol{\theta}$ is the unique solution to the following system

$$\boldsymbol{\theta}(H \otimes I_{\bar{W}} + \tilde{D}_0 + \tilde{D}_1) = \mathbf{0},\ \boldsymbol{\theta}\mathbf{e} = 1.$$

The probability $P^{esc-loss}$ that an arbitrary customer arrives when all servers are busy, buffer is not full, and the customer does not join the buffer and leaves the system is given as

$$P^{esc-loss} = \lambda^{-1} \sum_{i=N}^{N+K-1} \boldsymbol{\pi}_i \bar{Q}_{i-N}\tilde{D}_1\mathbf{e}.$$

The intensity of flow of customers, which get the service in the system, is calculated as

$$\lambda^{out} = \sum_{i=1}^{N+K} \min\{i, N\}\boldsymbol{\pi}_i(A \otimes I_{\bar{W}})\mathbf{e}.$$

The loss probability of an arbitrary customer is calculated as

$$P^{loss} = 1 - \frac{\lambda^{out}}{\lambda}.$$

The probability $P^{imp-loss}$ that an arbitrary customer after arrival will go to the buffer and leave it due to impatience is computed by

$$P^{imp-loss} = P^{loss} - P^{ent-loss} - P^{esc-loss}.$$

5 Distribution of Sojourn Time of an Arbitrary Customer in the System

Let $V(x)$ be the distribution function of the sojourn time of an arbitrary customer in the system and $v(s) = \int\limits_0^\infty e^{-sx} dV(x)$, Re $s > 0$, be its Laplace-Stieltjes transform (LST).

Let us tag an arbitrary customer and keep track of its staying in the system. We will derive the expression for the $LST\, v(s)$ by means of the method of collective marks (method of additional event, method of catastrophes) for references, see, e.g., [11], [12]. To this end, we interpret the variable s as the intensity of some imaginary stationary Poisson flow of catastrophes. So, $v(s)$ has the meaning of the probability that no catastrophe arrives during the sojourn time of the tagged customer.

Let $y(s, r)$ be the probability that a catastrophe will not arrive during the rest of the tagged customer's service time in the system conditioned on the fact that at the given moment the state of the random environment is r, $r = \overline{1, R}$.

The probabilities $y(s, r)$ can be found from the following system of linear algebraic equations:

$$y(s,r) = (\mu_r - H_{r,r} + s)^{-1}(\mu_r + \sum_{r'=1,\, r' \neq r}^{R} H_{r,r'} y(s,r')),\ r = \overline{1,R}. \qquad (2)$$

Let us form the vector

$$\mathbf{y}(s) = (y(s,1), \ldots, y(s,R))^T,$$

and rewrite system (2) into the matrix form as

$$(-A + H - sI)\mathbf{y}(s) = -A\mathbf{e}.$$

Note, that the matrix $-A+H-sI$ is subgenerator, so the matrix $(-A+H-sI)^{-1}$ exists and

$$\mathbf{y}(s) = (A - H + sI)^{-1} A\mathbf{e}.$$

Let $w(s, l, r)$ be the probability that a catastrophe will not arrive during the rest of the tagged customer's sojourn time in the system conditioned on the fact that

at the given moment the tagged customer has the position l, $l = \overline{1, K}$, in the buffer, and the state of the random environment is r, $r = \overline{1, R}$.

The probabilities $w(s, l, r)$, $l = \overline{1, K}, r = \overline{1, R}$, can be found from the system of linear algebraic equations:

$$w(s,l,r) = (s+l\alpha^{(r)}+N\mu_r-H_{r,r})^{-1}\Bigg(\delta_{l,1}N\mu_r y(s,r)+(1-\delta_{l,1})N\mu_r w(s,l-1,r)+$$

$$+\sum_{r'=1,\, r'\neq r}^{R} H_{r,r'}w(s,l,r') + (l-1)\alpha^{(r)}w(s,l-1,r) + \alpha^{(r)}\Bigg). \tag{3}$$

To find the solution to system (3), let us introduce the column vectors

$$\mathbf{w}(s,l) = (w(s,l,1),\dots,w(s,l,R))^T, \mathbf{w}(s) = ((\mathbf{w}(s,1))^T,\dots,(\mathbf{w}(s,K))^T)^T,$$

and rewrite system (3) into the matrix form as

$$(-sI - NA - lE + H)\mathbf{w}(s,l) + \delta_{l,1}NA\mathbf{y}(s)+$$

$$+((1-\delta_{l,1})NA + (l-1)E)\mathbf{w}(s,l-1) + E\mathbf{e} = \mathbf{0}^T,\ l = \overline{1, K},$$

and then

$$(-sI-NI_K\otimes A-C\otimes E+I_K\otimes H+NE^-\otimes A+(C-I_K)E^-\otimes E)\mathbf{w}(s)+\mathbf{a}(s) = \mathbf{0}^T, \tag{4}$$

where

$$C = \mathrm{diag}\{1,\dots,K\},$$

E^- is a square matrix of size K with all zero entries except the entries $(E^-)_{i,i-1}$, $i = \overline{1, K-1}$, which are equal to 1,

$$\mathbf{a}(s) = (\underbrace{(NA\mathbf{y}(s) + E\mathbf{e})^T, (E\mathbf{e})^T,\dots,(E\mathbf{e})^T}_{K})^T.$$

Let us introduce the matrix

$$V = -NI_K\otimes A - C\otimes E + I_K\otimes H + NE^-\otimes A + (C-I_K)E^-\otimes E.$$

So, system (4) can be rewriten in the following form:

$$(V - sI)\mathbf{w}(s) + \mathbf{a}(s) = \mathbf{0}^T. \tag{5}$$

It can be verified that the diagonal entries of the matrix $V - sI$ dominate in all rows of this matrix. So the inverse matrix exists. Thus we have proved the following assertion.

Theorem 2. *The vector* $\mathbf{w}(s)$ *consisting of the conditional LST* $w(s, l, r)$, $l = \overline{1, K}$, $r = \overline{1, R}$, *is calculated as*

$$\mathbf{w}(s) = -(V - sI)^{-1}\mathbf{a}(s). \tag{6}$$

Formula (6) gives the explicit form of the vector $\mathbf{w}(s)$, but in practice the matrix $V - sI$ usually has a big dimension. Using the fact that this matrix has block form the subvectors $\mathbf{w}(s,l)$, $l = \overline{1,K}$, of the vector $\mathbf{w}(s)$ can be easily calculated by recurrent formulas:

$$\mathbf{w}(s,1) = (NA + E - H + sI)^{-1}(NA\mathbf{y}(s) + E\mathbf{e})^T,$$

$$\mathbf{w}(s,l+1) = (NA + (l+1)E - H + sI)^{-1}[E\mathbf{e} - (NA + lE)\mathbf{w}(s,l)]^T, l = \overline{1,K-1}.$$

Theorem 3. *The LST $v(s)$ of distribution of an arbitrary customer's sojourn time in the system is computed by*

$$v(s) = P^{ent-loss} + P^{esc-loss} + \lambda^{-1}[\sum_{i=0}^{N-1}\sum_{r=1}^{R}\boldsymbol{\pi}(i,r)D_1^{(r)}\mathbf{e}y(s,r) +$$

$$+ \sum_{i=N}^{N+K-1}\sum_{r=1}^{R}(1 - q_{i-N}^{(r)})\boldsymbol{\pi}(i,r)D_1^{(r)}\mathbf{e}w(s,i-N+1,r)].$$

Corollary 1. *The average sojourn time V_{soj} of an arbitrary customer is calculated by*

$$V_{soj} = -v'(s)|_{s=0} = -\lambda^{-1}[\sum_{i=0}^{N-1}\sum_{r=1}^{R}\boldsymbol{\pi}(i,r)D_1^{(r)}\mathbf{e}\frac{\partial y(s,r)}{\partial s}|_{s=0} +$$

$$+ \sum_{i=N}^{N+K-1}\sum_{r=1}^{R}(1 - q_{i-N}^{(r)})\boldsymbol{\pi}(i,r)D_1^{(r)}\mathbf{e}\frac{\partial w(s,i-N+1,r)}{\partial s}|_{s=0}].$$

Here the values $\frac{\partial w(s,l,r)}{\partial s}|_{s=0}$, $l = \overline{1,K}, r = \overline{1,R}$, are calculated as the entries of the vector $\frac{d\mathbf{w}(s)}{ds}|_{s=0} = -V^{-1}[\mathbf{a}'(0) - \mathbf{e}]$, and the values $\frac{\partial y(s,r)}{\partial s}|_{s=0}$ are calculated as the entries of the vector $\frac{d\mathbf{y}(s)}{ds}|_{s=0} = -(A - H)^{-1}\mathbf{e}$.

Corollary 2. *The average waiting time V_{wait} of an arbitrary customer is calculated by formula*

$$V_{wait} = -\lambda^{-1}\sum_{i=N}^{N+K-1}\sum_{r=1}^{R}(1 - q_{i-N}^{(r)})\boldsymbol{\pi}(i,r)D_1^{(r)}\mathbf{e}\frac{\partial z(s,i-N+1,r)}{\partial s}|_{s=0}$$

where the values $\frac{\partial z(s,l,r)}{\partial s}|_{s=0}$, $l = \overline{1,K}, r = \overline{1,R}$, are calculated as the entries of the vector $\frac{d\mathbf{z}(s)}{ds}|_{s=0} = -V^{-2}\mathbf{a}(0)$.

6 Conclusion

In this paper, the multi-server queueing system with a MAP arrival process, a finite buffer and impatient customers operating in random environment is investigated. The process of system states is considered. The numerically stable

algorithm for calculating the steady state probabilities is presented. Expressions for the main performance characteristics of the system and the Laplace-Stieltjes transforms of the sojourn and waiting time distributions are obtained. The presented results can be used for modeling, performance evaluations and optimization of real call centers.

Acknowledgments. This research was supported by Belarusian Republican Foundation for Fundamental Research (grant No. F11M-003).

References

1. Aksin, O.Z., Armony, M., Mehrotra, V.: The modern call centers: a multi-disciplinary perspective on operations management research. Production and Operation Management 16, 655–688 (2007)
2. Kim, J.W., Park, S.C.: Outsourcing strategy in two-stage call centers. Computers & Operations Research 37, 790–805 (2010)
3. Kim, C., Dudin, S., Taramin, O., Baek, J.: Queueing system MAP/PH/N/N + R with impatient heterogeneous customers as a model of call center. Applied Mathematical Modelling (2012) dx.doi.org/10.1016/j.apm, 03.021
4. Dudin, S., Dudina, O.: Call center operation model as a $MAP/PH/N/R-N$ system with impatient customers. Problems of Information Transmission 47, 364–377 (2011)
5. Kim, C., Dudin, A., Klimenok, V., Khramova, V.: Erlang loss queueing system with batch arrivals operating in a random environment. Computers & Operations Research 36(3), 674–697 (2009)
6. Kim, C., Klimenok, V., Mushko, V., Dudin, A.: The BMAP/PH/N retrial queueing system operating in Markovian random environment. Computers & Operations Research 37(7), 1228–1237 (2010)
7. He, Q.M.: Queues with marked customers. Advances in Applied Probability 28, 567–587 (1996)
8. Kim, C.S., Dudin, S.A.: Priority tandem queueing model with admission control. Computers & Industrial Engineering 61, 131–140 (2011)
9. Cleveland, B.: Call center management on fast forward: succeeding in today's dynamic customer contact environment. In: ICMI (2006)
10. Graham, A.: Kronecker products and matrix calculus with applications. Ellis Horwood, Cichester (1981)
11. Kesten, H., Runnenburg, J.T.: Priority in waiting line problems. Mathematisch Centrum, Amsterdam (1956)
12. van Danzig, D.: Chaines de Markof dans les ensembles abstraits et applications aux processus avec regions absorbantes et au probleme des boucles. Ann. de l'Inst. H. Pioncare 14, 145–199 (1995)

Optimal Choice of the Capacities of Telecommunication Networks to Provide QoS-Routing

E. Girlich[1], M.M. Kovalev[2], and N.I. Listopad[3,⋆]

[1] Otto-von-Guericke-Universität, Fakultät für Mathematik, Universitätsplatz 2, 39106 Magdeburg
[2] Belarusian State University, Department of Economics, Karl Marx Street 31, 220050, Minsk, Belarus
[3] Belarusian State University of Informatics and Radioelectronics, Department of Radioelectronics, Brovky Street 6, 220050, Minsk, Belarus
eberhard.girlich@mathematik.uni-magdeburg.de, kovalev@bsu.by, listopad@unibel.by

Abstract. The problems of telecommunication networks designing could be presented into three aspects: 1) choice of the capacity for each telecommunication link with total minimum network cost; 2) QoS-routing of multicommodity flows in the synthesized network for all forecasting demands and 3) providing a necessary level of survivability. We consider QoS-routing, taking into account various performance requirements: delay, variation of the delay (jitter), bandwidth, packet loss probability. In this article we consider QoS-routing adding to consideration new constraints which provide the delay requirements as the important part of QoS.

Keywords: telecommunication networks design, optimization in telecommunication, multicommodity flows, QoS-routing, survivability.

1 Introduction

The development of the WWW-service and Virtual Private Networks (VPNs) has greatly changed the nature of telecommunication network design [1]. New applications, such as video conferencing, Internet telephony, various forms of e-commerce, e-government and e-learning represent specific performance requirements. Many of these applications are typically delay-sensitive with performance guarantees, sufficient resources (for example, bandwidth, processing time of the routers) are made available to various classes of traffic so that certain specified performance requirements (delay, variation delay) will be explicitly met. For such applications the best efforts service is no longer acceptable. Till today, the Internet was dominated by applications such as file transfer and e-mail. Since these applications could tolerate considerable delays, so-called the best efforts

⋆ Corresponding author.

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 93–104, 2013.

service, which not provide any performance guarantees, were acceptable. Below we should see that these problems are interconnected and it is possible to use the common mathematical model for their simultaneous decision. The term Quality of Service (QoS) describes network features that are used to provide the better than the best efforts performance that is required by any applications [1], [2]. In particular, we introduce QoS-routing by involving the multiple constrains: delay, variation of the delay (jitter), bandwidth, packet loss probability.

Our approach of telecommunication network designing could be presented into three following models.

Model 1 (Choice of Capacites). Sets of possible technologies for future telecommunication networks are given. It is necessary to determine the type, technologies and capacity for each telecommunication link.

Model 2 (QoS-Routing). It necessary to route of multicommodity flows in the synthesized network for all forecasting demands and QoS requirements (delay, jitter, packets loss e.g.).

Model 3 (Survivabiliy). Providing the necessary level of survivability is desirable to design networks that are robust with respect to link one node failures. The survivability is an important part of QoS requirements.

The article continues our investigations presented in [3].

2 Choice of the Capacity of the Telecommunication Networks

We shall represent the topology of the telecommunication networks as non-directed (or direct) graph $G = (V, E)$, where V is the set of nodes and E is the set of potential edges (arcs) connecting the nodes. The nodes of the graph G represent user equipment, routers, switchers, cross-connecters etc. The edges $e \in E$ of the graph G represent the telecommunication links which can be potentially used (exist by the current moment or can be established); for example, an optical fiber, copper links, radio-wave links, satellite channels, etc. If between two net nodes there are some various communication links they are represented by the parallel edges responding different technologies: Ethernet, Frame Relay, ISDN, ATM, etc. In telecommunication networks each link at transfer of the information is directed: one node passes the information, and the other accepts. The establishment of the link between i and j allows to pass an amount of the information in unit of time between i and j and the same amount of the information between j and i, if connectivity is synchronous, and total (from i to j plus from j to i) amount of the information, if connectivity is asynchronous.

A traffic demand (or just demand) is a requirement on the network design to provide for predetermined source s and sink t the future volume information $d(s, t)$. Let D be the set of all demands. For all demands $(s, t) \in D$ the positive numbers $d\,(s, t)$ are called *function of the traffic*. The function of the traffic is determined statistically on the basis of the information flows growth forecast.

As capacity $y_e(e)$ of the link $e \in E$, possible physical capacities of telecommunication links (at use the lease links and dial-up links for realization of

connectivity of type a point - point between net nodes), and speed on which connectivity can be carried out to local or global networks are understood. It is measured in bits/s or some resource unit (64 Kbit/s, channels, E1/T1s, fractional T1s, wavelengths, OC-ns). At use of a synchronous link in 1024 Kb/s it is possible to transfer in the same second 1024 kilobits/second (Kb/s) from i to j and 1024 Kb/s from j to i. There are two links of identical capacity in each such link that is essential. At asynchronous connectivity in 1024 Kb/s in the same second the information with total speed in 1024 Kb/s is transferred from i to j and from j to i, that it is necessary to take into account at designing networks.

There are two cases of the formulated problem: designing of topology and technology of a telecommunication network and upgrade of topology and a choice of technology of an existing network. As we shall see below, there are mathematical models which are common for both cases. For this purpose it is possible to consider that initial capacity of the each edge$e$ of the graph G is $C_0(e)$. If today there are not such channels, we are on opinion that $C_0(e) = 0$. That on establishing the links with initial capacity $C_0(e)$ it is not required capital expenses. In modern technologies continuous capacities seldom meet in practice. Much more often communication links have discrete capacities.

For each $e \in E$ set of possible capacities are determined by the following parameters:

$t(e) = |T(e)|$ is the number of possible additional capacities for an edge e;

$C_t(e) \in Z_+(1 \le t \le t(e))$ are the potential technologies for an edge e (it is supposed that $C_0(e) \le C_1(e) \le \cdots \le C_{t(e)}(e)$);

$K_t(e) \in Q_+(1 \le t \le t(e))$ is the cost of establishing the communication link with capacity $C_t(e)$.

For each edge $e \in E$, we introduce the variables

$$x_0(e) \ge x_1(e) \ge \cdots \ge x_t(e), \tag{1}$$

$$x_t\,(e) \in \{0, 1\}\,, \text{ for all } e \in E,\ t = \overline{1, t(e)}. \tag{2}$$

A choice of the capacity $C_\tau(e)(0 \le \tau \le t(e))$ for an edge means that $x_0(e) = x_1(e) ... = x_\tau(e) = 1, x_{\tau+1}(e) = ... = x_{t(e)} = 0$.

Then variables

$$y(e) = \sum_{t=0}^{t(e)} c_t(e)x_t(e), \text{for all } e \in E, \tag{3}$$

representing the capacities are installed on the edges e. Here

$$c_t(e) = C_t(e) - C_{t-1}(e)(1 \le t \le t(e)), k_t(e) = K_t(e) - K_{t-1}(e)(1 \le t \le t(e)).$$

For convenience of denotations we shall put $c_0(e) = C_0(e)$ and $k_0(e) = K_0(e)$.

If we should designate the cost $K(e)$ of establishing the telecommunication link $e \in E$, the common problem of design of topology and a choice of capacities $y(e)$ of networks could be formulated as follows.

Problem of the Capacity Choice: it is necessary to find subgraph $G = (V, E)$ of the complete graph on set of vertices V with the minimal total cost of edges

$K(G) = \sum_{e \in E} K(e)$ *such that the capacities of edges of the graph* G' *provide QoS-routing (probably, in view of constraints on lengths, delay, ..., of telecommunication links) the required amount of the information in accident-free and in failures.*

In other words, it means the following. The general problem of the telecommunication network design, including routing of information flows, will consist in definition of capacities of all telecommunication at which there is minimum cost network providing the transfer of flows under all demands (s,t) traffic $d(s,t)$.

In the case of considering networks with multiple edges, models of the choice of technologies could be simplified. Namely, if we should suggest that each line (edge) is responded with unique technology $C_\tau(e)$ and with cost $K_\tau(e)$ the problem of a choice of technology becomes simpler and can be formulated in the following ways:

to determine subgraph $G = (V, E)$ with the minimum cost of edges $\sum_{e \in E} K(e)$ and capacities y $c(e) = C(e)$ for $e \in E$ and $y(e) = 0$ for $e \notin E$, providing routing the traffic for all demands.

The above mentioned problem could be a little bit complicated: it is necessary to minimize a total cost of establishing of additional telecommunication links, and also to determine the routing paths on which the data will be transferred for satisfaction of all demands (a problem of routing). Understandably, that not at any capacities of links and topology of a network the solution of a problem of routing is possible.

In the optimal solution capacities $y(e)$ could be equal to 0, therefore the problem of a choice of technologies automatically also includes the problem of a choice of topology of a network.

Designing of a network without taking into account capacities can be applied at building of a new network. Sometimes at the initial stages of development more attention is given to topology, and capacities of separate communication links are determined at later stages (only topological models are used). For optimization of already existing network models, which will take into account at capacities of existing topology, are more preferable.

In view of the introduced denotations and assumptions the problem of a choice of the capacities can be formulated as the follows:

$$\min \sum_{e \in E} \sum_{t=1}^{t(e)} k_t(e) x_t(e), \tag{4}$$

subjects to (1)-(3) and supplement of constraints to provide QoS-routing in a network (V, E, y) the traffic $d(s,t)$ for all demands $(s,t) \in D$.

3 QoS-Routing of the Multicommodity Flows in the Form of "Flows-Arcs"

The problem of routing is an identification of one or several paths along which there will be traffic $d(s,t)$ from a source s to a sink t. A flow between s and t

nodes we shall call a flow of the type (s,t). Thus, in a problem of routing it is necessary to segment the traffic and, for each of the segments, to find the path of data transfer. Thus on one link there could be different flows, but in the sum they should not exceed its real opportunities on data transfer.

The initial base model used for the analysis paths are multiflows in the networks [4], [5], [6], [7], [8], [9]. There are possible two formulations of the model in terms of "flows - arcs" and in terms of "flows - paths". We shall also use both of them below and consider the models for two technologies: synchronous and asynchronous.

If an edge $e = (v, w) \in E$ is directed from v to w we say, that e leaves from v and enters in w. Set of the arcs, which are incoming in v, we shall designate through $E_{int}(v)$ and the arcs leaving v is denotes as $E_{out}(v)$.

The nonnegative numbers

$$f(s,t;e) > 0, e \in E, \text{ for all } (s,t) \in D \tag{5}$$

are called *as a multiflow*, if they satisfy the following linear balance equations:

$$\sum_{e \in E_{int}(v)} f(s,t,e) - \sum_{e \in E_{out}(v)} f(s,t,e) = \begin{cases} -d(s,t), & v = s, \\ 0, & v \neq s,t, \\ d(s,t), & v = t, \end{cases} \quad \text{for all } v \in V, (s,t) \in D. \tag{6}$$

and also inequations on capacities of edges:

for synchronous technologies:

$$\sum_{(s,t) \in D} f(s,t,e) \leq y(e), \text{ for all } e \in E; \tag{7}$$

or for asynchronous technologies:

$$0 \leq \sum_{(s,t) \in D} \left(f^{+}(s,t,e) + f^{-}(s,t,e) \right) \leq y(e), \text{ for all } e \in E \tag{7'}$$

which express that fact that the total amount of the flows of all types in both directions on any edge cannot exceed capacity of this edge.

The problem of designing of an optimal telecommunication network with discrete capacities $y(e)$ is formulated as the multi-commodity flow models, described by (1)-(7) for synchronous network or (1)-(6), (7') for asynchronous networks.

QoS-routing should be provided by including constraint for the average packet delay. The queuing plus transmission delay have frequently been approximated using $M/M/1$ model. As the results by the Kleinrock-formula [1] for average packet delay in the network the following constraints should be noted:

$$\frac{1}{\gamma} \sum_{e \in E} f_e(s,t,e) \left[\frac{1}{y_e(s,t,e) - f_e(s,t,e)} + \mu(P_e + K_e) \right] \leq T_{\max}(s,t), \text{ for all} (s,t) \in D \tag{8}$$

where: $T_{\max}(s,t)$ is maximum possible delay; $1/\mu$ is the average packet length (bits/packet); λ_e is the average packet arrival rate to link e (packets/second);

P_e is propagation delay on link e; K_e is node processing delay entering link e; γ is total traffic in the network (packets/second).

There are some ways to determine maximum possible delay. First of all, you should allocate $T_{\max}(s,t)$ empirically, for example, from performance required by any application.

Klincewicz [1] proposed the algorithms to allocate maximum delay for each route any links and the network at the whole. An objective function could be more complicated and include other requirements of QoS, not only delay, but, for example, cost of delay for each link e [1]:

$$T(s,t,e) = \beta \frac{f_e(s,t,e)}{y_e(s,t,e) - f_e(s,t,e)} \tag{9}$$

where $T(s,t,e)$ is cost delay for link of the demand (s,t); β is cost factor.

Function (9) is derived from $M/M/1$ expression for queuing and insertion delay.

4 QoS-Routing in the Form of "Flows-Paths"

Let's designate through $P(s,t)$ the set of all paths from s in t in graph $G = (V,E)$. Concrete path P from $P(s,t)$, containing an edge (or vertex u), we shall designate $P \in P(s,t) : e \in P, u \in P$. Let $f(s,t;P)$ be the flow of type (s,t) along path $P \in P(s,t)$.

It is known that always there is a decomposition of the flow $f(s,t;e)$ as flows on paths (the theorem of decomposition [5]) so, that

$$f(s,t;e) = \sum_{P \in P(0;s,t): e \in P} f(s,t;P).$$

Cost of transfer of unit of the flow on path $P \in P(s,t)$ is determined by the following.

$$K(s,t;P) = \sum_{e \in P} K(s,t,e),$$

where $K(s,t;e)$ is cost of transfer on an arc e the unit of the information on demand (s,t). Cost of transfer can not depend on type of the demand (s,t). In general $K(s,t;e)$ can differ from cost $k_\tau(e)$, escalating capacity of a line e on technology τ.

The problem of designing of an optimal telecommunication network for discrete capacities $y(e)$ in the form of "flows-paths" can be represented as the following model:

$$\sum_{(s,t) \in D} \sum_{P \in P(0,s,t): e \in P} K(s,t;e) f(s,t;P) \to \min \tag{10}$$

subject to (1)-(2) and constraints on capacities of the arcs for synchronous links

$$\sum_{(s,t) \in D} \sum_{P \in P(0;s,t): e \in P} f(s,t;P) \le y(e);\ y(e) = \sum_{t=0}^{t(e)} c_t(e) x_t(e), \text{ for all } e \in E; \tag{11}$$

on constraints on volume of the demands

$$\sum_{P \in P(0;s,t;)} f(s,t;P) = d(s,t) \text{ for all} (s,t) \in D \tag{12}$$

$$f(s,t;P) \geq 0 \text{ for all } (s,t) \in D \text{ and } P \in P(s,t). \tag{13}$$

For asynchronous links constraints (11) are replaced by the following:

$$\sum_{(s,t)\in D} \sum_{P \in P(0;s,t):e\in P} f(0;s,t;P) \leq y(e), \text{ for all } e \in E^{+}(\text{for direct arcs}); \tag{11'}$$

$$\sum_{(s,t)\in D} \sum_{P \in P(0;s,t):e\in P} f(0;s,t;P) \leq y(e), \text{ for all } e \in E^{-}(\text{for return arcs}).$$

It is not difficult to notice, that the problem of routing formulated through flow variables on the ways, has simple enough structure of constraint s. For each edge $e \in E$ there is a unique constraint on total amount of the flows on an edge. It is limited by its bandwidth. For each demand (s,t), there is one constraint which guarantees that the amount of the flow from a source s to a sink t will be equal $d(s,t)$.

The main feature of the given model is polynomial number of the constraints and exponential number of unknown variables. For the solution of relaxation problems LP for the given model the simplex - method with procedure of generation columns is effective [10], [11].

In case of loading of telecommunication links some paths, carrying the information between pairs of s and t nodes from set of demand D, could appear very long. By the various reasons (for example, to reduce the delays to establish the connectivity or to reduce loading computers) it happens desirable to limit length of communication ways [1]. We described above the models (1)-(7) or (1)-(6), (7') of designing of optimal networks with unlimited length of paths and models without unlimited length of ways. More complicated models include multi-constraints (see, for example, [1], [2], [12]).

Let each link $(u,v) \in E$ be specified by m additive QoS weights $w_i(u,v) \geq 0, i = 1,\ldots,m$. The weights correspond to the QoS metrics: delay, variation of delay and so on. Delay is the amount of time between the moments when a packet enters the network and leaves the network. It is the most common factor considered in QoS metrics. So, in this paper we pay more attention for consideration of delay factor. The various possible components of delay include: insertion delay, queuing delay, node processing delay and propagation delay. Insertion (transmission) delay refers to the time required to insert a packet of given size (in kilobits) on transmission facility that severs packets at a given rate (kilobits per second). Queuing delay refers to the time that the packet has to wait at the output buffer to be served by the transmission facilities. Node processing delay includes time required for the router to examine and route the packet and to perform other operations, such as encryption/decryption of data compression. The propagation delay refers to the time required to traverse the transmission facility (related to the length of the link and the speed of light).

A path from source s to sink t such that

$$\sum_{(u,v)\in P} w_{i_i}(u,v) \le L_i \text{ for all } i = 1,\dots,m, \tag{14}$$

is called *QoS-feasible* path. Let $P_{fes}(s,t)$ be the set of QoS-feasible paths from s to t. Then the problem of QoS-routing should be formulated as the models described by (10-14) where the set of all paths $P(s,t)$ is changed on the set of QoS-feasible paths $P_{fes}(s,t)$.

5 Solution Strategy

The optimization models described in previous section lead to mixed-integer linear programs with Boolean variables x_i which are presented by discrete capacities $y(e)$ and continues variables $f(s,t,e)$ or $f(s,t,P)$ [3], [13], [14].

Standard approach allows to combine enumeration algorithm for definition discrete capacities $y(e)$ and special linear programm for multicommodity flow with additional constraints (constraints paths, QoS-feasible path, etc.).

Without loss of a generality, it is possible to assume that all costs $K(e)$ of edges of the graph G are integer. It implies that if K_{opt} is a cost of an optimal network then $K_{opt} \in \left[0, K_{\max} = \sum_{e\in E} K_t(e)\right]$. Thus, the problem of designing the network can be solved with use of the dichotomy on a cost interval $[0, K_{\max}]$. The chosen approach could be reduced to the solution $O(\log K)$ of the following ND(K) problem:

ND(K) problem: for each integer K, to find subgraph $G = (V,E)$ of the graph, taking in account that $\sum\limits_{e\in E} K(e) \le K$ and capacities $y(e)$ of edges of the graph G provide routing, probably taking into account the constraints on lengths of paths (strategy of short ways or QoS-feasible paths) in normal and in all failure states.

Let consider the solution of the problem ND(K). Take for example the problem of providing the survivability of the network by strategy of reservation in case of asynchronous networks. We shall introduce the following objective function:

$$z_u(f_u^+, f_u^-, y_u) =$$

$$= \sum_{(s,t)\in D} \sum_{e\in E} \alpha_{u,st}(e)(f^+(u;s,t;e) + f^-(u;s,t;e)) + \sum_{(s,t)\in D} \beta_{u,st} y_{u,st} + \mu y \to \max$$

where: $\alpha_{u,st}(e) (0 \le \alpha_{u,st}(e) \le d(u,s,t))$ is parameter which determines the amount of the minimal demand.

Let's determine the coefficients of objective function as follows:

$$\alpha_{u,st}(e) = -1 \text{ for all } e \in E,\ (s,t) \in D; \beta_{u,st} = \begin{cases} K + \frac{1}{2}, & u = 0; \\ |V|, & u \ne 0. \end{cases}$$

Parameter m can get one of two values: 0 or M where, as it is traditional in linear programming formulation, M is big enough number.

If for providing survivability of a network strategy of reservation is used the ND(K) problem could be written down as the following ND1(K) linear problem: to maximize objective function $z_0\left(f_0^+, f_0^-, y_0\right)$ subject to the following constraints on nonnegative flow variables:

on equality to zero of the flows on the failure nodes:

$$f^+(u;s,t;u)=0,\ f^-(u;s,t;u)=0,\ \text{for all } (s,t)\in D,\ u\in E;$$

$$f^+(u;s,t;e)=0,\ f^-(u;s,t;e)=0,\ \text{for all } e\in E_{int}(u)\cup E_{out}(u),(s,t)\in D,u\in V;$$

on the numbers of the demands which have been written down for all $s\in\{0\}\cup\bar{E}\cup\bar{V}$ as:

$$d(u;s,t)\le y(u;s,t)\le d(s,t),\ (s,t)\in D,$$

and also under following conditions imposed by dichotomy process:

$$\sum_{e\in E}K(e)x(e)\le A;\ x(e)=0,1;\text{for all } e\in E.$$

Optimal routing of the flows for a state u in a network determined by the network $G_i(V_i,E_i,y)$ is the solution of the following problem of linear programming:

Problem LP(u): *to maximize objective function* $z_u\left(f_u^+,\ f_u^-,\ y_u\right)$ *with the purpose of definition of the new paths for transfer of the information at the following constraints*:

on nonnegative flow variables:

$$f^+\left(u;s,t;e\right)\ge 0, f^-\left(u;s,t;e\right)\ge 0,\ \text{for all } e\in E,\ (s,t)\in D;$$

flows on the failed nodes of the network are equal to zero:

$$f^+(u;s,t;u)=0, f^-(u;s,t;u)=0,\ (s,t)\in D, u\in E;$$

$$f^+(u;s,t;e)=0, f^-(u;s,t;e)=0,\ \text{ for all } e\in E_{int}(u)\cup E_{out}(u),\ (s,t)\in D,\ u\in V;$$

on capacities: $\sum_{(s,t)\in D}\left(f^+(u;s,t;e)+f^-(u;s,t;e)\right)\le y(e),\ \text{ for all } e\in E$
balance constraints:

$$\left(\sum_{e\in E_{int}(v)}f^+(u;s,t;e)+\sum_{e\in E_{out}(v)}f^-(u;s,t;e)\right)-$$

$$-\left(\sum_{e\in E_{int}(v)}f^+(u;s,t;e)-\sum_{e\in E_{out}(v)}f^-(u;s,t;e)\right)=$$

$$=\begin{cases}-y(u;s,t), & v=s;\\ 0, & v\in V\backslash\{s,t\};\\ y(u;s,t), & v=t;\end{cases}\ \text{for all } (s,t)\in D$$

on number of demands:

$$d(u;s,t)\le y(u;s,t)\le d(s,t),\ \text{for all } (s,t)\in D.$$

It is easy to understand that $LP(u)$ problem has the solution for a state u when a capacity of network G is sufficient for rerouting a given percent of the information for each of the demand.

Let's notice, that in problem LP(u), $u \in \{0\} \cup E \cup V$ variables $\{x(e)\}_{e \in E}$ are fixed and are not unknown.

Let's stop on the analysis of the optimal solution of LP(u) problem for the state u, and actually we shall determine the new paths for transfer of the information. Let

$$\bar{f}^+(u; s, t; e),\ \bar{f}(u; s, t; e),\ \bar{y}(u; s, t),\ e \in E,\ (s, t) \in D$$

be the components of the optimal basic solution of the problem LP(u). As the columns of the matrix of the constraints of a problem, which are corresponded to variables $\bar{f}^+(u; s, t; e)$, $\bar{f}(u; s, t; e)$, are linearly dependent, then one of values $\bar{f}^+(u; s, t; e)$, $\bar{f}^-(u; s, t; e)$ should be equal to zero. For $(s, t) \in D$ on set of the nodes V we shall determine the subgraph $G_{s,t}$ with set of arcs $A_{s,t}$ and the valid function $g_{s,t}$ on $A_{s,t}$ as follows: $(s, t) \in A_{s,t}, g_{s,t}(s, t) = y_{(0,s,t)}$; for $e = (v, w) \in E$ (orientation from $v \quad w$); if $\bar{f}^+(0; s, t; e) > 0$, $(v, w) \in A_{s,t}$ then $g_{s,t}(v, w) = \bar{f}^+(0; s, t; e)$, and if $\bar{f}^-(0; s, t; e) > 0$, $(v, w) \in A_{s,t}$ then $g_{s,t}(v, w) = \bar{f}^-(0; s, t; e)$. We shall note that $g_{s,t}$ is circulation. Therefore, in time $O(|V||E|)$ it is possible to find family $G^1_{s,t}, ..., G^{k(s,t)}_{s,t}$ of simple subcycles in $G_{s,t}$ and a set of positive real numbers $\varepsilon^1_{s,t}, ..., \varepsilon^{k(s,t)}_{s,t}$ such that

$$g_s(e) = \sum_{1 \le j \le k(s,t),\ e \in E(G^j_{s,t})} \varepsilon^j_{s,t}, \qquad \forall e \in A_{s,t}.$$

Here $E(G)$ is a set of arcs of the cycle G. We shall determine the weights of arcs of the subgraph $G_{s,t}$ as follows: $w_{s,t}(s, t) = \beta_{u,s,t}$; $\quad w_{s,t}(e) = a_{u,s,t}(e)$, if $e \in A_{s,t} \backslash \{s, t\}$. Then by virtue of definition of circulationg_{st} we have:

$$\sum_{j=1}^{k(s,t)} \varepsilon^j_{s,t} \sum_{e \in E(G^j_{s,t})} w_{s,t}(e) = \sum_{e \in A_j} w_{s,t}(e) g_{s,t}(e) = z_u(\bar{f}^+_u,\ \bar{f}^-_u,\ y_u).$$

We should see that weight $w_{s,t}\left(G^j_{s,t}\right) = \sum\limits_{e \in E\left(G^j_{s,t}\right)} w_{s,t}\,(e)$ of each of cycles $G^j_{s,t}$ is nonnegative. Having removed an arc (s, t) from cycles $G^1_{s,t}, ..., G^{k(s,t)}_{s,t}$ *we shall receive the family of simple* (s, t) *ways* $P^1_{s,t}, ..., P^{k(s,t)}_{s,t}$ *on which* $\varepsilon^1_{s,t}, ..., \varepsilon^{k(s,t)}_{s,t}$ *units of the information could be transferred accordingly between* s *and* t. As $d(u; s, t) \le \bar{y}_{u,s,t} = \sum\limits_{j=1}^{k(s,t)} \varepsilon^j_{s,t}$, then on paths $P^1_{s,t}, ..., P^{k(s,t)}_{s,t}$ it is possible to transfer required amount of the information.

The survey of another methods of the optimal network design one can find, for example, in [1].

6 Conclusions and Extensions

The main features of the presented models of the optimal routing are taking in account the QoS requirements, in particular, delay metric as very important requirement for a lot of delay-sensitive applications.

To develop the models in the form of "flows - paths", objection function can be a little bit another, for example, where the function of the cost includes cost of transferring the unit of information. Objective function can be even more complex, for example, as expression (4) or (10) where along with cost of transferring of unit of the information expenses for building of a network (a building of the additional telecommunication links) and cost of delay (9) are taken into account.

Some researchers develop the models, where the decisions suggested by these models would not necessary be implemented immediately. But future advance in technologies (e.g., integration of the IP layer and optical layer) will likely make it more possible for networks to respond in real time to short-term changes in traffic demands. So, optimization models that address various types of real-time decisions will be needed [1].

In majority of the works the traffic on the link is presented be $M/M/1$ model. But more realistic model, that could be described the traffic at the networks is model, based on BMAP-flows. This allows for network design procedures to utilize more realistic models and characterizations of traffic behavior both in the calculation of network delay and in the sizing of network links.

In our consideration we analyze the different types of delay: insertion, queuing, node processing delay and the propagation delay. Incorporating new QoS metrics, such as application delay, into network design models will allow users to make more direct connection between system requirements and model inputs.

Network design problems with QoS consideration are typically difficult solved combinatorial problems. Success in these directions of research will enable network designers in any practical problems by optimal routing information flows with QoS requirements. For example, using the model (5)-(9) for upgrade educational network Unibel of the Ministry of Education of Belarus it's needed more than ten hours of PC work.

References

1. Resende, M., Pardalos, P.: Handbook of Optimization in Telecommunications. Springer Science, Business Media (2006)
2. Van Mieghem, P., Kuipers, F.A., Korkmaz, T., Krunz, M., Curado, M., Monteiro, E., Masip-Bruin, X., Solé-Pareta, J., Sánchez-López, S.: Quality of Service Routing. In: Smirnov, M. (ed.) COST 263 Final Report. LNCS, vol. 2856, pp. 80–117. Springer, Heidelberg (2003)
3. Girlich, E., Kovalev, M.M., Listopad, N.I.: Optimization of the Topology and the Capacities of Telecommunications Networks, http://www.math.uni-magdeburg.de/~girlish/preprints/final_eng.pdf

4. Chifflet, J., Hadjiat, M., Maurras, J.-F., Vaxes, Y.: A Primal Partitioning Approach for Single and Non-Simultaneous Multicommodity Flow Problems. Discrete Mathematics 1 (1998)
5. Grötschel, M., Monma, C.L., Stoer, M.: Computational results with a cutting plan algorithm for designing communication networks with low-connectivity constraints. Operations Research 2(3), 474–504 (1992)
6. Gröstchel, M., Monma, C.I., Stoer, M.: Polyhedral anad Computational Investigations for designing Communication networks with High Survivability Requirements. Konrad-Zuse-Zentrum für Informationstechnik Berlin. Preprint SC 92-94, (1992)
7. Hu, T.: Integer programming and flows in networks. Addison-Wesley, CityLondon (1970)
8. Maurras, J.F., Vaxes, Y.: Multicommodity network flow with jump constraints. Discrete Mathematics 165–166 (1997)
9. Minoux, M.: Optimum synthes of a network with non-simultaneous multicommodity flow requirements in Stadies on Graphs and Diskrete Programming, pp. 269–277. North Holland (1981)
10. Grötschel, M., Monma, C.L., Stoer, M.: Design of survivable networks. In: Handbook in Operations Research and Management Science. Network Models, pp. 617–672. North-Holland (1995)
11. Listopad, N.: Modelling and optimization of global networks. Byelorussian State University, Minsk (2000)
12. Kuipers, F.A.: Quality of Service Routing in the Internet. Theory, Complexity and Algorithms. Delft University Press (2004)
13. Holnberg, K., Yuan, D.: A lagrangian hueristic based branch-and-bound approch for the capacitated network design problem. Operation Research 48(3), 461–481 (2000)
14. Kovalev, M., Pisaruk, M.: Modern linear programming. Byelarussian State University, Minsk (1988)

A Retrial Tandem Queue with Two Types of Customers and Reservation of Channels

Valentina Klimenok and Roman Savko

Department of Applied Mathematics and Computer Science
Belarusian State University
Minsk 220030, Belarus
klimenok@bsu.by

Abstract. We consider a retrial tandem queue with two multi-server stations which can be considered as a mathematical model of a call center with two types of customers classified by their ability to wait for the connection to the agent. Customers arrive at Station 1 according the stationary Poisson flow. If an arriving customer meets all servers busy he/she goes to the infinite size orbit and retries after a random time. The type of a customer is randomly determined upon completion of the service at Station 1. If all servers of Station 2 are busy type 1 (priority) customer leaves the system forever while type 2 (non-priority) customer is queued in the buffer of limited size. If the buffer is full this customer leaves the system. The customers staying in the queue are impatient. This means that they might decide to leave the system before their service at Station 2 begins. It is assumed that a number of servers of Station 2 can be reserved to serve priority customers only. We calculate the stationary distribution and the main performance measures of the system. The cost function evaluating quality of service under different number of reserved server is constructed. Illustrative numerical example is presented.

Keywords: tandem queue, retrials, impatient customers, reservation of servers, stationary performance measures.

1 Introduction

Call centers provide customer support, help-desk services, reservation and sales support, order-taking functions for catalog and Web-based merchants. To offer high quality services, call center managers and designers should consider the complex factors associated with random arrivals of customers and a variety of customer requirements for quality of service. Queueing models can be effectively used for call center design and support of their management. The survey of research works devoted to mathematical modeling of call center can be seen in survey [1], papers [5,7,6] and references therein.

In this paper we consider two-stations tandem queue with retrial phenomena at Station 1. The retrial phenomena is that a customer who finds all lines busy

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 105–114, 2013.

upon arrival joins the virtual group of blocked customers called orbit and, independently of other orbital customers, retries for the service after a random amount of time. Such a behavior is typical for outside calls calling to the center and receiving a signal that all trunk lines are busy. Once the call is connected, it are handled in automatic call distributor that is designed to root calls connected via Station 1. In our consideration all calls are directed to service at Station 2 which are modeled by a multi-server queue with a limited buffer. We assume that all agents (servers of Station 2) are flexible enough to answer all requirements of service but the customers to be served at Station 2 is divided into two different classes according to their ability to wait for the connection to the agent. We assume that the company that owns the call center provides strict preferences to high-valued clients who show absolute impatience. If such a client is not connected to an agent immediately after dialing, he/she leaves the system without service. To prevent the loss of most of these clients, management of call center may decide that a group of agents will serve high-valued (priority) clients only. The non-priority customers can also abandon service. They wait for only a limited time, and hang up after this time expires. As it was shown (see, e.g., [3]), the performance evaluation of the models with abandonment is essentially differs from the models without abandonment.

Our purpose is to calculate the stationary performance measures of the tandem queue modeling the above call center and discuss the problem of optimal reservation of servers (agents) at Station 2.

2 Model Description

We consider a tandem queueing system consisting of two stations in series. Station 1 is represented by the N-server queue without a buffer. Customers arrive at Station 1 according the stationary Poisson flow with parameter λ. If an arriving customer meets all servers busy he/she goes to the infinite size orbit and tries his/her luck later on after a random amount of time. We assume that the total flow of retrials is such as the probability of generating a retrial attempt in the interval $(t, t + \Delta t)$ is equal to $\alpha_i \Delta t + o(\Delta t)$ when the number of customers in the orbit is equal to i, $i > 0$, $\alpha_0 = 0$. We do not fix the explicit dependence of the intensity α_i on i assuming only that $\lim\limits_{i\to\infty} \alpha_i = \infty$. Note that such a dependence describes the classic retrial strategy ($\alpha_i = i\alpha$) and the linear strategy ($\alpha_i = i\alpha + \gamma,\ \alpha > 0$) as special cases.

We assume that customers arriving at Station 1 are not homogeneous. They can be of two types. The type of a customer is determined by a randomized manner upon completion of a service at Station 1: with probability p the customer is classified as type 1 (priority) customer and with probability $q = 1 - p$ he/she is classified as type 2 (non-priority) one. Customer's type is determined based on his/her ability to wait for the service at Station 2. If all servers of Station 2 are busy type 1 (priority) customer leaves the system forever while type 2 (non-priority) customer is queued in the buffer of size M. If the buffer is full type 2 customer leaves the system. We assume that some number R, $R \leq N$, servers of Station 2 are reserved to serve priority customers only.

For each customer placed into the buffer, the waiting time is restricted by the random value having the exponential distribution with parameter γ. If this time (obsolescence time) expires before the customer is picked-up from the buffer to the server, it is assumed that this customer immediately leaves the buffer and is lost. The obsolescence times of different customers are independent of each other and identically distributed.

All servers of the tandem are independent of each other. The service time of a customer at Station k is exponentially distributed with parameter μ_k, $k = 1, 2$.

The structure of the system is presented in Figure 1.

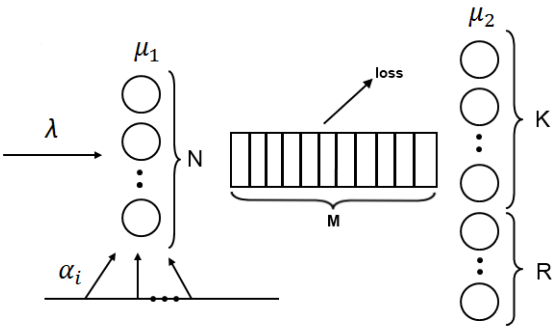


Fig. 1. The structure of the system

For further use in the sequel, we introduce the following notation:

- I ($\mathbf{e}$) is an identity matrix (a row vector of units) of appropriate dimension. When needed we will identify the dimension of the matrix (the vector) with a suffix;
- 0_l is a square matrix of size l consisting of zeroes;
- $\otimes$ is a symbol of Kronecker's product of matrices, see [4];
- $\tilde{I}$ and $\hat{I}$ are square matrices of size $M + 1$ that are defined as

$$\tilde{I} = \begin{pmatrix} 0\,1\,0\,\ldots\,0\,0 \\ 0\,0\,1\,\ldots\,0\,0 \\ \vdots\,\vdots\,\vdots\,\ddots\,\vdots\,\vdots \\ 0\,0\,0\,\ldots\,0\,1 \\ 0\,0\,0\,\ldots\,0\,0 \end{pmatrix}, \quad \hat{I}_H = \begin{pmatrix} 0\,0\,0\,\ldots\,0\,0 \\ 1\,0\,0\,\ldots\,0\,0 \\ 0\,1\,0\,\ldots\,0\,0 \\ \vdots\,\vdots\,\vdots\,\ddots\,\vdots\,\vdots \\ 0\,0\,0\,\ldots\,1\,0 \end{pmatrix}$$

- $\overline{I}$ ($\overline{\overline{I}}$) is a square matrix of size $M + 1$ whose first (last) diagonal entry is equal to 1 and others entries are equal to zero;
- $diag\{a_l,\ l = \overline{1, L}\}$ is a diagonal matrix with diagonal entries or blocks a_l;

3 Process of the System States

Let

- i_t be the number of calls in the orbit;
- n_t be the number of busy servers at Station 1;
- r_t be the number of busy servers at Station 2;
- m_t be the number of customers staying in the buffer at time t, $t \geq 0$.

The process of the system states is described in terms of the irreducible four-dimensional continuous-time Markov chain $\xi_t = \{i_t, n_t, r_t, m_t\}, t \geq 0$ with state space $X = \{(i, n, r, m),\ i \geq 0,\ n = \overline{0, N},\ r = \overline{0, K+R},\ m = \overline{0, M}\}$.

Enumerate the states of this chain in lexicographic order, and denote by $Q_{i,j}$, $i, j \geq 0$, the square matrix of order $(N+1)(K+R+1)(M+1)$ governing the transition rates of the chain from the set of states $\{i, \cdot, \cdot\}$ to the set $\{j, \cdot, \cdot\}$.

Lemma 1. *Infinitesimal generator of the Markov chain ξ_t, $t \geq 0$, has the following block structure:*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & 0 & 0 & 0 & \dots \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & 0 & 0 & \dots \\ 0 & Q_{2,1} & Q_{2,2} & Q_{2,3} & 0 & \dots \\ 0 & 0 & Q_{3,1} & Q_{3,2} & Q_{3,3} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where sub-diagonal and over-diagonal blocks are calculated as follows:

$$Q_{i,i-1} = \alpha_i \tilde{I}_{N+1} \otimes I_{(K+R+1)(M+1)},\ i \geq 1,$$

$$Q_{i,i+1} = \lambda \overline{I}_{N+1} \otimes I_{(K+R+1)(M+1)},\ i \geq 0,$$

diagonal blocks are represented as block matrices $Q_{i,i} = ((Q_{i,i})_{n,n'})_{n,n'=\overline{0,N}}$ with non-zero blocks of the following form:

$$(Q_{i,i})_{n,n-1} =$$

$$= n\mu_1 \begin{pmatrix} 0 & I_{M+1} & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & I_{M+1} & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & I_{M+1} & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & q(\tilde{I}+\overline{I}) & pI & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & q(\tilde{I}+\overline{I}) & pI & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & q(\tilde{I}+\overline{I}) & pI \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & q(\tilde{I}+\overline{I})+pI \end{pmatrix},$$

$$i \geq 0,\ n = \overline{1, N};$$

$$(Q_{i,i})_{n,n} =$$

$$= \mu_2 \begin{pmatrix} 0 & 0 \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ I_{M+1} & 0 \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 \dots & (K-1)I_{M+1} & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 \dots & 0 & K\overline{\overline{I}}_{M+1} & K\hat{I} & \dots & 0 & 0 \\ 0 & 0 \dots & 0 & 0 & (K+1)\overline{\overline{I}} & \dots & 0 & 0 \\ \vdots & \vdots \ddots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 \dots & 0 & 0 & 0 & \dots & (K+R)\overline{\overline{I}} & (K+R)\hat{I} \end{pmatrix}$$

$$+\gamma diag\{0_{K(M+1)}, I_{R+1} \otimes diag\{0, 1, \dots, M\}(\hat{I} - I)\}$$

$$-diag\{n\mu_1 + (1 - \delta_{n,N})\alpha_i + \lambda + k\mu_2, k = \overline{0, K+R}\} \otimes I_{M+1},\ i \geq 0,\ n = \overline{0, N};$$

$$(Q_{i,i})_{n,n+1} = \lambda I_{(K+R+1)(M+1)},\ i \geq 0,\ n = \overline{0, N-1}.$$

In the further investigation of the Markov chain $\xi_t, t \geq 0$, we will use the results for continuous time asymptotically quasi-toeplitz Markov chains (AQTMC) presented in [8].

Corollary 1. *The Markov chain $\xi_t, t \geq 0$, belongs to the class of continuous time asymptotically quasi-toeplitz Markov chains.*

Proof. According to the definition given in [8], the chain $\xi_t, t \geq 0$, belongs to the class of continuous time AQTMC if there exist the limits

$$Y_0 = \lim_{i \to \infty} C_i^{-1} Q_{i,i-1},\ Y_1 = \lim_{i \to \infty} C_i^{-1} Q_{i,i} + I,\ Y_2 = \lim_{i \to \infty} C_i^{-1} Q_{i,i+1}, \quad (1)$$

and the matrix $\sum_{k=0}^{\infty} Y_k$ is a stochastic one.

Here C_i is a diagonal matrix defined by modules of diagonal entries of the matrix $Q_{i,i}, i \geq 0$.

It is easy to see that the diagonal entries of the matrix C_i corresponding to the first N block rows of the matrix $Q_{i,i}$ include the term α_i while the rest of diagonal does not depend on i. Then this matrix can be represented as

$$C_i = \begin{pmatrix} C_1(i) & 0 \\ 0 & C_2 \end{pmatrix}.$$

Taking into account dependence (or not dependence) of the blocks of the matrices $Q_{i,i+k}$, $k = -1, 0, 1$, of α_i we calculate the limits (1) as follows:

$$Y_0 = \tilde{I}_{N+1} \otimes I_{(K+R+1)(M+1)},\ Y_2 = \begin{pmatrix} 0_{N(K+R+1)(M+1)} & 0 \\ 0 & C_2^{-1}(Q_{i,i+1})_{N,N} \end{pmatrix},$$

$$Y_1 = \begin{pmatrix} 0_{N(K+R+1)(M+1)} & 0 \\ C_2^{-1}(Q_{i,i})_{N,N-1} & C_2^{-1}(Q_{i,i})_{N,N} + I \end{pmatrix}.$$

It is easy to see that the sum of these matrices is a stochastic matrix. Thus, the chain $\xi_t, t \geq 0$, is asymptotically quasi-toeplitz Markov chain.

4 Stationary Distribution

It is clear that the condition for existing the stationary distribution of the tandem under consideration coincides with such a condition for the retrial queue $M/M/N$ representing the first station of the tandem. Using the results of the paper [2] where the more general retrial queue $BMAP/PH/N$ has been investigated we immediately get the following statement.

Theorem 1. *(i) The stationary distribution of the Markov chain ξ_t, $t \geq 0$, exists if the following inequality*

$$\rho = \lambda/(N\mu_1) < 1. \tag{2}$$

holds.

(ii) The stationary distribution of the chain ξ_t, $t \geq 0$, does not exist if inequality (2) has an opposite sign.

In what follows we assume inequality (2) be fulfilled.

Let $p(i,n,r,m)$, $i \geq 0, n = \overline{0,N}, r = \overline{0,K+R}, m = \overline{0,M}$, be the steady state probabilities of the chain $\xi_t, t \geq 0$. Enumerate the states of the $\xi_t, t \geq 0$, in the lexicographic order and form the row vector $\mathbf{p}_i$ of steady state probabilities corresponding the value i of the first component, $i \geq 0$. To calculate the vectors $\mathbf{p}_i$, $i \geq 0$, we use the numerically stable algorithm (see [8]) which has been elaborated for calculating the stationary distribution of the multi-dimensional continuous time asymptotically quasi-toeplitz Markov chain. Taking into account the specifics of the chain under consideration, this algorithm takes the following form.

1. Compute Neuts' matrix G (see [9]) as the minimal nonnegative solution of the matrix equation $G = Y_0 + Y_1 G + Y_2 G^2$.
2. For preassigned sufficiently large integer i_0 compute the matrices G_{i_0-1}, $G_{i_0-2}, \ldots, G_0$ using the equation of the backward recursion

$$G_i = \left(-Q_{i+1,i+1} - Q_{i+1,i+2}G_{i+1}\right)^{-1} Q_{i+1,i},\ i = i_0 - 1, i_0 - 2, \ldots, 0$$

   with the boundary condition $G_i = G$, $i \geq i_0$.
3. Compute the matrices $\bar{Q}_{i,j}$, $j \geq i$, by the formulas

$$\bar{Q}_{i,i} = Q_{i,i} + Q_{i,i+1}G_i,\ \ \bar{Q}_{i,i+1} = Q_{i,i+1}.$$

4. Compute the matrices F_i from the recursion

$$F_0 = I, F_i = F_{i-1}\bar{Q}_{i-1,i}\left(-\bar{Q}_{i,i}\right)^{-1},\ i \geq 1.$$

5. Compute the vector $\boldsymbol{p}_0$ as the unique solution to the system

$$\boldsymbol{p}_0(-\bar{Q}_{0,0}) = \mathbf{0},\ \ \boldsymbol{p}_0 \sum_{i=0}^{\infty} F_i \mathbf{e} = 1.$$

6. Compute the vectors $\boldsymbol{p}_i$ by the formulas $\boldsymbol{p}_i = \boldsymbol{p}_0 F_i$, $i \geq 0$.

5 Performance Measures

- Mean number of calls in the orbit $L_{orb} = \sum_{i=1}^{\infty} i\boldsymbol{p}_i\mathbf{e}$.
- Joint stationary distribution of the number of busy servers at Station 1 and the number of customers at Station 2 $\boldsymbol{P}^{(1,2)} = \sum_{l=0}^{\infty} \boldsymbol{p}_i$.
- Stationary distribution of the number of busy servers at Station 1

$$\boldsymbol{P}^{(1)} = \boldsymbol{P}^{(1,2)}(I_{N+1} \otimes \mathbf{e}_{(K+R+1)(M+1)}).$$

- Mean number of busy servers at Station 1 $\overline{N}^{(1)} = \boldsymbol{P}^{(1)} diag\{0, 1, \ldots, N\}\mathbf{e}$.
- Stationary distribution of the number of busy servers at Station 2

$$\boldsymbol{P}^{(2)} = \boldsymbol{P}^{(1,2)}(\mathbf{e}_{N+1} \otimes I_{K+R+1} \otimes \mathbf{e}_{M+1}).$$

- Mean number of busy servers at Station 1

$$\overline{N}^{(2)} = \boldsymbol{P}^{(2)} diag\{0, 1, \ldots, K + R + 1\}\mathbf{e}.$$

- Stationary distribution of the number of non-priority customers in the buffer

$$\boldsymbol{P}^{(buff)} = \boldsymbol{P}^{(1,2)}(\mathbf{e}_{N+1} \otimes \mathbf{e}_{K+R+1} \otimes I_{M+1}).$$

- Probability that an arbitrary arriving customer will be lost

$$P_{loss} = 1 - \lambda^{-1}\boldsymbol{P}^{(2)}\,\mu_2 diag\{0, 1, \ldots, K + R + 1\}\mathbf{e}.$$

- Intensity of output flow from Station 1

$$\bar{\mu}^{(1)} = \mu_1 \boldsymbol{P}^{(1,2)}(diag\{0, 1, \ldots, N\} \otimes I_{(K+R+1)(M+1)})\mathbf{e}.$$

- Probability that an arbitrary priority customer will be lost due to lack of idle servers at Station 2

$$P_{loss}^{(prior)} = \mu_1 \boldsymbol{P}^{(1,2)}(diag\{0, 1, \ldots, N\} \otimes \bar{I}_{K+R+1} \otimes I_{M+1})\mathbf{e}/\bar{\mu}^{(1)}.$$

- Probability that an arbitrary non-priority customer will be lost due to lack of free space at Station 2

$$P_{loss}^{(non-prior,buff)} = \mu_1 \boldsymbol{P}^{(1,2)}(diag\{0, 1, \ldots, N\} \otimes I_{K+R+1} \otimes \bar{I}_{M+1})\mathbf{e}/\bar{\mu}^{(1)}.$$

- Probability that an arbitrary non-priority customer will be lost due to impatience

$$P_{loss}^{(non-prior,ipm)} = P_{loss} - qP_{loss}^{(prior)} - pP_{loss}^{(non-prior,buff)}.$$

6 Numerical Example

In the numerical example we provide graphics for the loss probabilities associated with tandem and consider the problem of optimal choice of the number R of reserved servers at Station 2. We introduce the following cost criterion (an average penalty per unit time under the steady-state operation of the system):

$$I = aR + q\bar{\mu}^{(1)}(c_1 P_{loss}^{(non-prior,buff)} + c_2 P_{loss}^{(non-prior,ipm)}) + c_3 p\bar{\mu}^{(1)} P_{loss}^{prior}$$

where a is a cost of maintenance of a reserved server per unit time, $c_1(c_2)$ is a penalty for the loss of non-priority customer due to absence of free space in the buffer (due to impatience), c_3 is a penalty for the loss of priority customer.

The parameters of the queue under consideration are assumed to be as follows. The number of servers at Stations 1 and Station 2 are $N = 6$ and $K + R = 14$ respectively. The size of the buffer is M = 8. Intensity of arrival is $\lambda = 5$. The service rate at stations are defined by $\mu_1 = 1; \mu_2 = 0.4$. We consider the classic retrial strategy $\alpha_i = i\alpha$ where $\alpha = 1$. The impatience rate is $\gamma = 1$.

Figures 2-4 depict the probabilities $P_{loss}^{(prior)}$, $P_{loss}^{(non-prior,buff)}$, $P_{loss}^{(non-prior,imp)}$ as functions of the number R of reserved server of Station 2 for different values of probability p.

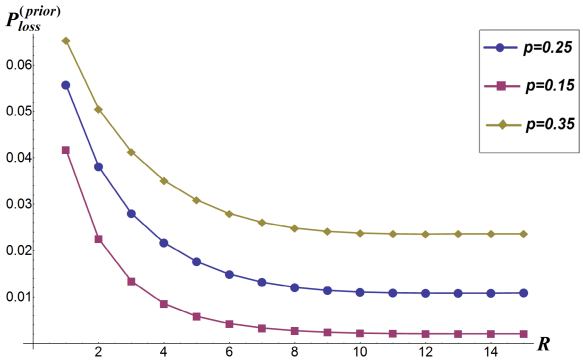


Fig. 2. The probability $P_{loss}^{(prior)}$ vs the number R of reserved server

Let now consider an example of numerical solution of optimization problem. The cost coefficients a, c_1, c_2, c_3 are assumed to be as follows: $a = 0, c_1 = 10, c_2 = 5, c_3 = 60$. Our aim is to find numerically the optimal number R of reserved servers at Station 2 that provides the minimal value to the cost criterion for different share p of priority customers. The value of criterion I as a function of R and p is presented in Figure 5. The minimum value of I is achieved at the point $R = 2$, $p = 0.1$ and is equal to 2.614.

For better understanding the behavior of cost criterion we present in Figure 6 the values of the criterion as function of R under three different values of p. It

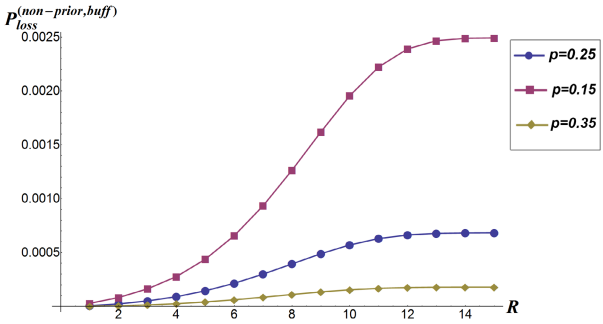


Fig. 3. The probability $P_{loss}^{(non-prior,buff)}$ vs the number R of reserved servers

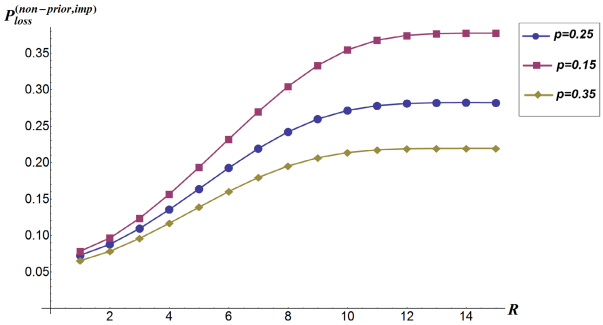


Fig. 4. The probability $P_{loss}^{(non-prior,imp)}$ vs the number R of reserved servers

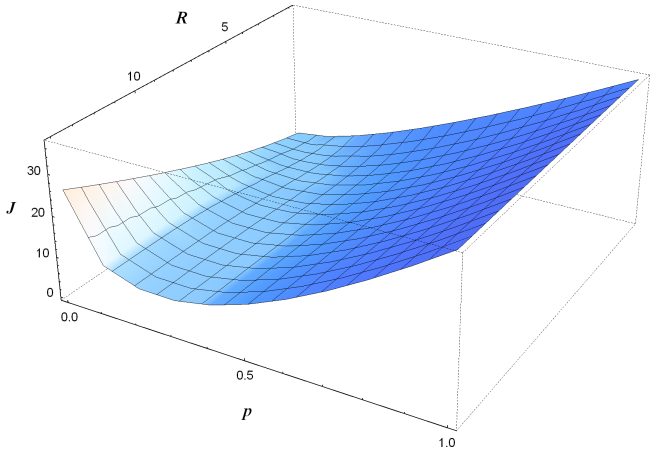


Fig. 5. The cost criterion as a function of the number R of reserved servers at Station 2 and the share p of priority customers

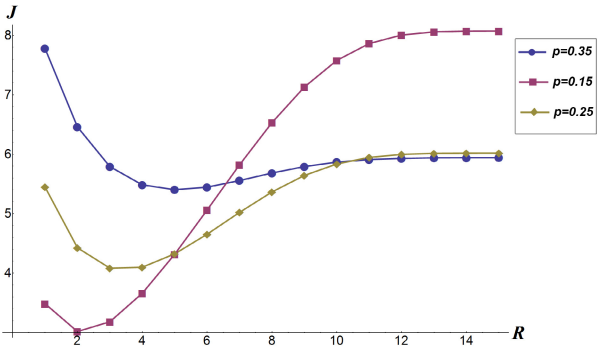


Fig. 6. The cost criterion as a function of the number R of reserved servers at Station 2 for different shares p of priority customers

is seen from the figure that the optimal number of reserved servers at Station 2 increases from 2 to 5 when the share p of priority customers increases from 0.15 to 0.35.

References

1. Aksin, Z., Armony, M., Mehrotra, V.: The modern Call centers: a multidisciplinary perspective on operation management research. Production and Operation Management 16, 655–688 (2007)
2. Breuer, L., Dudin, A., Klimenok, V.: A retrial $BMAP/PH/N$ system. Queueing Systems 40, 433–457 (2002)
3. Garnett, O., Mandelbaum, A., Reiman, M.: Designing a call center with impatient customers. Manufacturing and Service Operation Management 4, 208–227 (2002)
4. Graham, A.: Kronecker Products and Matrix Calculus with Applications. Ellis Horwood, Cichester (1981)
5. Jouini, O., Dallery, Y., Aksin, Z.: Queueing models for full-flexible multi-class Call centers with real-time anticipated delays. Int. J. Production Economics 120, 389–399 (2009)
6. Jouini, O., Pot, A., Dallery, Y.: Online scheduling policies for multicalass call canters with inpatient customers. European Journal of Operational Research 207, 258–268 (2010)
7. Khudyakov, P., Feigin, P., Mandelbaum, A.: Designing a call center with an IVR (Interactive Voice Response). Queueing Systems 66, 215–237 (2010)
8. Klimenok, V.I., Dudin, A.N.: Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. Queueing Systems 54, 245–259 (2006)
9. Neuts, M.: Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach. Johns Hopkins University Press (1981)

Some Aspects of Waiting Time in Cyclic-Waiting Systems

Laszlo Lakatos[1] and Dmitry Efroshinin[2]

[1] Eötvös Loránd University, Budapest, Hungary
[2] Johannes Kepler Universität, Linz, Austria

Abstract. We consider a queueing system with Poisson arrivals and exponentially distributed service time and FCFS service discipline. The service of a customer is started at the moment of arrival (in case of free system) or at moments differing from it by the multiples of a given cycle time T (in case of occupied server or waiting queue). The waiting time is always the multiple of cycle time T, one finds its generating function and mean value. The characteristics of service are illustrated by numerical examples. If we measure the waiting time by means of number of cycles, we can optimize the cycle time T.

1 Introduction

According to the Kendall notation a queueing system is characterized by the interarrival and service times, the number of servers and the waiting room. This notation does not include the service discipline which plays key role, too. It determines the order of service, these rules may be rather simple (first-come-first-served, last-come-first-served, random, etc.) or more complex depending on the waiting time, number of present customers or priorities and so on. The analysis of queueing system with simple probabilistic characteristics may be rather complicated because of the service discipline.

We propose to consider a single-server queueing system, where an entering customer may be accepted for service either at the moment of arrival or at moments differing from it by the multiples of a given so-called cycle time. In order to illustrate the problem we give two practical examples.

1. Airplanes arrive at the airport in optimal position for landing. If there is no queue and the previous one is far enough, they start the landing process. If the distance is too small or there are some waiting ones, they start cycling. The next request for service may be put when the the airplane arrives at the starting geometrical point and this procedure is repeated.

2. Optical signals enter a node and they should be transmitted according to the FCFS rule. This information cannot be stored, if it cannot be serviced at once is sent to a delay line and returns to the node after having passed it. Clearly, the signal can be transmitted from the node at the moment of its arrival or at the time that differs from it by a multiple of time necessary to pass the delay line.

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 115–121, 2013.

The original problem was raised in connection with the landing process of airplanes [2], later it appeared to be an exact model for the transmission of optical signals where because of lack of optical RAM the fiber delay lines are used. First the system was studied characterizing it by the number of present customers [2], Koba [1] found sufficient condition for the stability of GI/G/1 system and gave the system of equations determining the waiting time's ergodic distribution. Koba and Mykhalevich [3] compared the classical retrial M/G/1 system with the cyclic-waiting one. [4], [5] describe the application of model for the transmission of optical signals.

The queueing systems may be considered from the viewpoints of the system and the individual customers. From the viewpoint of the system the number of present customers is important, from the viewpoint of individual customers the waiting time plays essential role. In this paper we concentrate our attention on the waiting time for such systems.

2 Theoretical Results

We consider a queueing system with Poisson arrivals and exponentially distributed service time, the corresponding parameters are λ and μ, respectively. The service is realized according to the order of arrivals and fix some cycle time T. If the system is free, the entering customer is immediately taken for service. If the server is occupied or there is a waiting queue, the customer starts cycling with cycle length T, it can put the next request for service arriving at the starting geometrical point. If it is at the head of queue and the server is free its service begins, in the opposite case this process is repeated.

We will use Koba's results [1] to find the waiting time distribution. We shortly repeat them.

Let t_n denote the time of arrival of the n-th customer; its service will begin at the moment $t_n + T \cdot X_n$, where X_n is a nonnegative integer. Let $\xi_n = t_{n+1} - t_n$, and η_n be the service time of n-th customer. Furthermore, let $X_n = i$, if

$$(k-1)T < iT + Y_n - Z_n \le kT \qquad (k \ge 1),$$

then $X_{n+1} = k$, and if $iT + \eta_n - \xi_n \le 0$, then $X_{n+1} = 0$. Hence, X_n is a homogeneous Markov chain with transition probabilities p_{ik}, where

$$p_{ik} = P\{(k-i-1)T < \eta_n - \xi_n \le (k-i)T\}$$

if $k \ge 1$, and $p_{i0} = P\{\eta_n - \xi_n \le -iT\}$. Introduce the notations

$$f_j = P\{(j-1)T < \eta_n - \xi_n \le jT\}, \tag{1}$$

$$p_{ik} = f_{k-i} \quad \text{if} \quad k \ge 1, \quad p_{i0} = \sum_{j=-\infty}^{-i} f_j = \hat{f}_i. \tag{2}$$

The ergodic distribution of this chain satisfies the system of equations

$$p_j = \sum_{i=0}^{\infty} p_i p_{ij} \quad (j \ge 0), \quad \sum_{j=0}^{\infty} p_j = 1.$$

Theorem 1. *Let us consider the above described system and introduce a Markov chain whose states correspond to the waiting time (in the sense that the waiting time is the number of actual state multiplied by T) at the arrival time of customers. The matrix of transition probabilities for this chain is*

$$\begin{bmatrix} \sum_{j=-\infty}^{0} f_j & f_1 & f_2 & f_3 & f_4 & \dots \\ \sum_{j=-\infty}^{-1} f_j & f_0 & f_1 & f_2 & f_3 & \dots \\ \sum_{j=-\infty}^{-2} f_j & f_{-1} & f_0 & f_1 & f_2 & \dots \\ \sum_{j=-\infty}^{-3} f_j & f_{-2} & f_{-1} & f_0 & f_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

its elements are defined by (1) and (2). The generating function of the ergodic distribution is

$$P(z) = \left[1 - \frac{\lambda}{\mu}\frac{1-e^{-\lambda T}}{e^{-\lambda T}(1-e^{-\mu T})}\right] \times \tag{3}$$
$$\times \frac{\frac{\mu}{\lambda+\mu} - \frac{\mu(1-e^{-\lambda T})}{\lambda+\mu}\frac{z}{z-e^{-\lambda T}}}{1 - \frac{\lambda(1-e^{-\mu T})}{\lambda+\mu}\frac{z}{1-ze^{-\mu T}} - \frac{\mu(1-e^{-\lambda T})}{\lambda+\mu}\frac{z}{z-e^{-\lambda T}}},$$

the condition of existence of ergodic distribution is

$$\frac{\lambda}{\mu} < \frac{e^{-\lambda T}(1-e^{-\mu T})}{1-e^{-\lambda T}}. \tag{4}$$

Proof. For the system we have

$$P\{Z < x\} = 1 - e^{-\lambda x}, \qquad P\{Y < x\} = 1 - e^{-\mu x}.$$

$\eta - \xi$ has the distribution

$$F(x) = \begin{cases} \frac{\mu}{\lambda+\mu}e^{\lambda x} & \text{if } x \le 0, \\ 1 - \frac{\lambda}{\lambda+\mu}e^{-\mu x} & \text{if } x > 0. \end{cases}$$

The transition probabilities of the Markov chain are, if $j > 0$

$$f_j = 1 - \frac{\lambda}{\lambda+\mu}e^{-\mu(j-1)T} - 1 + \frac{\lambda}{\lambda+\mu}e^{-\mu jT} = \frac{\lambda}{\lambda+\mu}(1-e^{-\mu T})e^{-\mu(j-1)T},$$

for the negative values ($j \ge 0$)

$$f_{-j} = \frac{\mu}{\lambda+\mu}e^{-\lambda jT} - \frac{\mu}{\lambda+\mu}e^{-\lambda(j+1)T} = \frac{\mu}{\lambda+\mu}(1-e^{-\lambda T})e^{-\lambda jT},$$

$$p_{i0} = \hat{f}_i = \sum_{j=-\infty}^{-i} f_j = \sum_{j=i}^{\infty} \frac{\mu}{\lambda+\mu}(1-e^{-\lambda T})e^{-\lambda j T} = \frac{\mu}{\lambda+\mu}e^{-\lambda i T}.$$

Using the matrix of transition probabilities, we obtain the system of equations

$$\begin{aligned} p_0 &= p_0\hat{f}_0 + p_1\hat{f}_1 + p_2\hat{f}_2 + p_3\hat{f}_3 + \dots \\ p_1 &= p_0 f_1 + p_1 f_0 + p_2 f_{-1} + p_3 f_{-2} + \dots \\ p_2 &= p_0 f_2 + p_1 f_1 + p_2 f_0 + p_3 f_{-1} + \dots \\ &\vdots \end{aligned}$$

Multiplying the j-th equation by z^j, summing up from zero to infinity, for the generating function $P(z) = \sum\limits_{j=0}^{\infty} p_j z^j$ we obtain

$$P(z) = P(z)F_+(z) + \sum_{j=1}^{\infty} p_j z^j \sum_{i=0}^{j-1} f_{-i} z^{-i} + \sum_{j=0}^{\infty} p_j \hat{f}_j. \tag{5}$$

For our system

$$\begin{aligned} F_+(z) &= \sum_{i=1}^{\infty} f_i z^i = \frac{\lambda z}{\lambda+\mu}(1-e^{-\mu T}) \sum_{i=1}^{\infty} e^{-\mu(i-1)T} z^{i-1} = \\ &= \frac{\lambda(1-e^{-\mu T})}{\lambda+\mu} \frac{z}{1-ze^{-\mu T}}, \\ \sum_{i=0}^{j-1} f_{-i} z^{-i} &= \frac{\mu(1-e^{-\lambda T})}{\lambda+\mu} \sum_{i=0}^{j-1} e^{-\lambda i T} z^{-i} = \frac{\mu(1-e^{-\lambda T})}{\lambda+\mu} \frac{1-\left(\frac{e^{-\lambda T}}{z}\right)^j}{1-\frac{e^{-\lambda T}}{z}}, \\ \sum_{i=0}^{\infty} p_i \hat{f}_i &= \sum_{i=0}^{\infty} p_i \frac{\mu}{\lambda+\mu} e^{-\lambda i T} = \frac{\mu}{\lambda+\mu} P\left(e^{-\lambda T}\right). \end{aligned}$$

Substituting these expressions (5) yields

$$P(z) = P(z)F_+(z) + \sum_{j=1}^{\infty} p_j z^j \frac{\mu(1-e^{-\lambda T})}{\lambda+\mu} \frac{1-\left(\frac{e^{-\lambda T}}{z}\right)^j}{1-\frac{e^{-\lambda T}}{z}} + \frac{\mu}{\lambda+\mu} P\left(e^{-\lambda T}\right) =$$

$$= P(z)F_+(z) + \frac{\mu(1-e^{-\lambda T})}{\lambda+\mu} \frac{z}{z-e^{-\lambda T}} \left[P(z) - P\left(e^{-\lambda T}\right)\right] + \frac{\mu}{\lambda+\mu} P\left(e^{-\lambda T}\right),$$

or

$$P(z)\left[1 - F_+(z) - \frac{\mu(1-e^{-\lambda T})}{\lambda+\mu} \frac{z}{z-e^{-\lambda T}}\right] =$$

$$= P\left(e^{-\lambda T}\right)\left[\frac{\mu}{\lambda+\mu} - \frac{\mu(1-e^{-\lambda T})}{\lambda+\mu} \frac{z}{z-e^{-\lambda T}}\right].$$

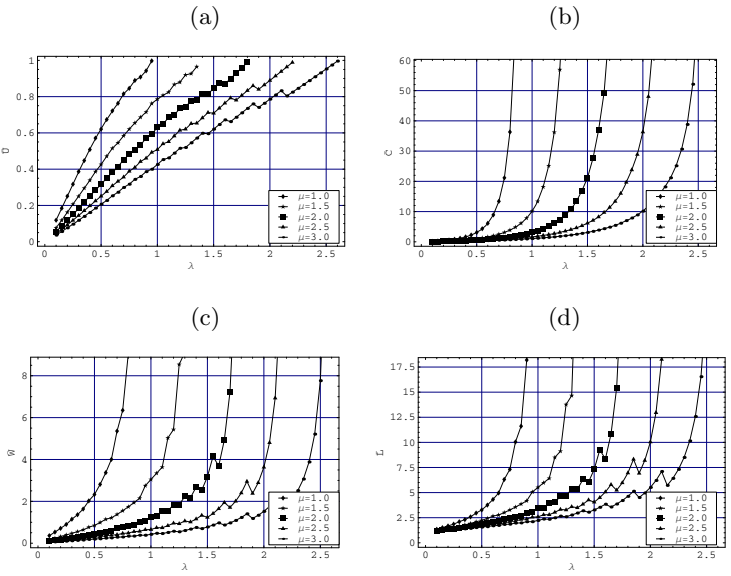


Fig. 1. $\bar{U}$ (a), $\bar{C}$ (b), $\bar{W}$ (c) and $\bar{L}$ (d) versus λ

The value of $P(e^{-\lambda T})$ can be found from the fact $P(1) = 1$,

$$P\left(e^{-\lambda T}\right) = 1 - \frac{\lambda}{\mu} \frac{1 - e^{-\lambda T}}{e^{-\lambda T}(1 - e^{-\mu T})}.$$

For the generating function of waiting time we obtain the above expression, whence the probability of zero waiting time is

$$p_0 = \left[1 - \frac{\lambda}{\mu} \frac{1 - e^{-\lambda T}}{e^{-\lambda T}(1 - e^{-\mu T})}\right] \frac{\mu}{\lambda + \mu}.$$

Because of ergodicity $p_0 > 0$ must hold, so the inequality

$$\frac{\lambda}{\mu} \frac{1 - e^{-\lambda T}}{e^{-\lambda T}(1 - e^{-\mu T})} < 1$$

must be fulfilled. It leads to the condition (4), and coincides with the stability condition for the number of customers [2].

3 Mean Performance Measures

As soon as the probabilities p_i, $i \geq 0$ are known, different performance characteristics of the system can be evaluated. Some of them are enumerated below.

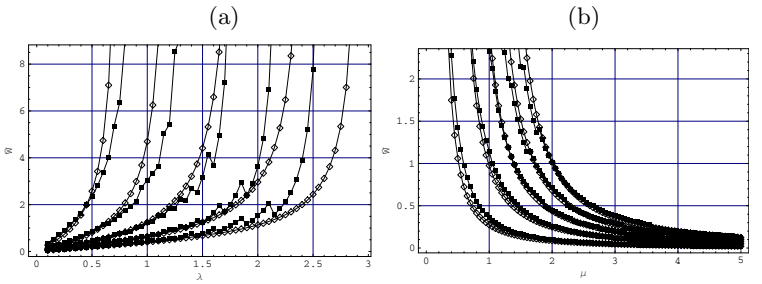


Fig. 2. $\bar{W}$ versus λ for $\mu = \{1.0, 1.5, 2.0, 2.5, 3.0\}$ (a) and versus μ for $\lambda = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ (b) in case of exact and approximated values of T^*

Utilization of the system $\bar{U} = 1 - p_0 = \dfrac{\lambda}{\lambda + \mu} \dfrac{1 - e^{-(\lambda+\mu)T}}{e^{-\lambda T}(1 - e^{-\mu T})}$.

Mean number of retrial cycles (from [3])

$$\bar{C} = \frac{\lambda[1 - e^{-(\lambda+\mu)T}]}{(1 - e^{-\mu T})[\mu e^{-\lambda T}(1 - e^{-\mu T}) - \lambda(1 - e^{-\lambda T})]}.$$

Mean waiting and sojourn time

$$\bar{W} = T\bar{C}, \quad \bar{S} = \bar{W} + \frac{1}{\mu}.$$

Mean number of customers in orbit and system $\bar{Q} = \lambda\bar{W}, \quad \bar{N} = \lambda\bar{S}$.

Mean busy period

$$\bar{L} = \frac{1}{\lambda}\left(\frac{1}{p_0} - 1\right) = \frac{1 - e^{-(\lambda+\mu)T}}{(\lambda + \mu)e^{-\lambda T}(1 - e^{-\mu T}) - \lambda(1 - e^{-(\lambda+\mu)T})}.$$

Mean number of customers served in a busy period $\bar{N}_L = \mu\bar{L}$.

4 Optimization of the Retrial Cycle

Our numerical experiments indicate that optimization of the values like $\bar{U}$, $\bar{W}$, $\bar{S}$, $\bar{Q}$, $\bar{N}$, $\bar{L}$ and $\bar{N}_L$ for any fixed parameters λ and μ leads to trivial solution, i.e. when the length of retrial cycle is $T = 0$.

Assuming $T > 0$, it is interesting to study the value $\bar{C}$ of the mean number of cycles. The formal optimization problem can be written as follows

$$\bar{C} = \bar{C}(T) \Rightarrow \min_T .$$

Figure 3 illustrates the concave structure of the curve $\bar{C}$ upon varying the parameter T for different values of λ and μ. Hence the optimal value $T^* > 0$ exists and can be evaluated. This figure shows that T^* takes a lower value while λ increases and/or μ decreases. In this case the server with a higher probability will

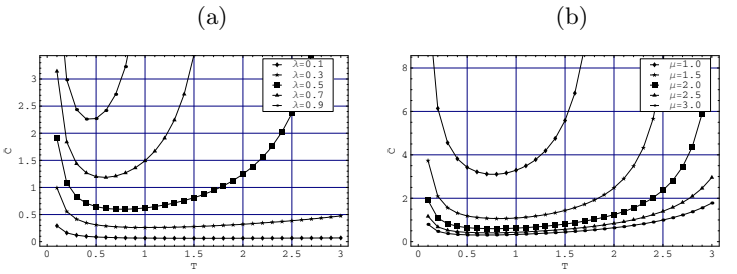


Fig. 3. $\bar{C}$ versus T, λ (a) and μ (b)

be busy and obviously the cycle length should be decreased in order to minimize the mean number of retrial cycles during the waiting time.

It is not possible to derive an explicit formula for the value T^*. The function $\bar{C}(T)$ can be minimized numerically by evaluation of the derivative which must be set to be equal to 0. Another way is a simple enumerative technique if the parameter T is changed in some interval with a small step.

Depending on the value of T the number of cycles first decreases and achieving some optimal value it increases. This fact may be explained on a simple way. In case of small T till the beginning of service of the next customer a large number of cycles is required (the service time is significantly greater than the length of a cycle). Approaching the optimal value T^* the number cycles decreases, leaving it the cycle is longer and longer and the waiting time will mainly be determined not by the service time, but the length of the cycle. It makes the waiting time large, consequently the required number of cycles will grow.

References

1. Koba, E.V.: On a GI/G/1 queueing system with repetition of requests for service and FCFS service discipline. Dopovidi NAN Ukrainy (6), 101–103 (2000) (in Russian)
2. Lakatos, L.: On a simple continuous cyclic-waiting problem. Annales Univ. Sci. Budapest. Sect. Comp. 14, 105–113 (1994)
3. Mykhalevich, K.V.: A comparison of a classical retrial M/G/1 queueing system and a Lakatos-type M/G/1 cyclic-waiting time queueing system. Annales Univ. Sci. Budapest. Sect. Comp. 23, 229–238 (2004)
4. Rogiest, W., Laevens, K., Fiems, D., Bruneel, H.: Analysis of a Lakatos-type queueing system with general service times. In: Proc. of ORBEL 20. Quantitative Methods for Decision Making, Ghent, January 19-20, pp. 95–97 (2006)
5. Rogiest, W., Laevens, K., Walraevens, J., Bruneel, H.: Analyzing a degenerate buffer with general inter-arrival and service times in discrete time. Queueing Systems 56, 203–212 (2007)

Gaussian Approximation of Multi-channel Networks in Heavy Traffic

Eugene Lebedev and Ganna Livinska

Taras Shevchenko National University of Kyiv

Abstract. In the paper the multichannel stochastic networks are considered. From the outside on each node of the network a Poisson input flow of calls arrives. An approximate method of studying of the service process at heavy traffic regime is developed. The limit process is represented as a multidimensional diffusion.

Keywords: multichannel stochastic network, heavy traffic regime, Gaussian approximation.

1 Introduction and Main Result

The basic mathematical model under consideration in the paper is a queuing network consisting of "r " service nodes. From the outside a Poisson input flow of calls $\nu_i(t)$ with the leading function $\Lambda_i(t)$ arrives at the i -th node, $i = 1, 2, ..., r$. Each of "r" nodes operates as a multi-channel stochastic system. If the call arrives at such a system then its service immediately begins. In the i-th node service time is exponential distributed with parameter μ_i , $i = 1, 2, ..., r$. After service in the i -th node the call arrives in the j-th node with probability p_{ij} and leaves the network with probability $p_{i,r+1} = 1 - \sum_{j=1}^{r} p_{ij}$. $P = \|p_{ij}\|_1^r$ is a switching matrix of the network. An additional node numbered "$r + 1$" is interpreted as "output" from the network.

According to the notation system, which is adopted in the theory of stochastic networks, such the model will be marked by the symbol $[M_t|M|\infty]^r$.

Let $Q_i(t)$, $i = 1, 2, ..., r$ be the number of calls in the i-th node of the network at t moment time.To the r - dimensional process $Q'(t) = (Q_1(t), ..., Q_r(t))$ we will refer as to a service process of calls in the network of the $[M_t|M|\infty]^r$- type. Our main goal is to study the process $Q(t)$ in conditions of heavy traffic.

The heavy traffic regime is determined by the following behavior of network parameters.

Condition 1. Input flows depend on n (series number) so that in any finite interval $[0, T]$

$$n^{-1}\Lambda_i^{(n)}(nt) \Rightarrow_{n\to\infty}^{U} \Lambda_i^{(0)}(t) \in C[0, T], \quad i = 1, 2, ..., r \tag{1}$$

where $C[0, T]$ is a set of continuous functions, symbol $\overset{U}{\Rightarrow}$ means convergence in uniform metric.

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 122–130, 2013.

Let us consider two cases that are important for applications when the Condition 1 is held.

We temporarily assume that the Poisson flow $\nu_i(t)$ is regular: $\Lambda_i(t) = \int\limits_0^t \lambda_i(u)du$, where $\lambda_i(u)$ is an instant value of the parameter (see, for example [1], page 100). It is naturally to call this flow as a Poisson flow with variable parameter.

If for the regular flow

$$\lim_{t\to\infty} \lambda_i(t) = \lambda_i > 0,$$

then Condition 1 holds with $\Lambda_i^{(0)}(t) = \lambda_i t$.

This follows from the estimates:

$$\left|\frac{1}{t}\int_0^t \lambda_i(u)du - \lambda_i\right| \le \frac{1}{t}\int_{\varepsilon t}^t |\lambda_i(u) - \lambda_i|du + \frac{1}{t}\int_0^{\varepsilon t} |\lambda_i(u) - \lambda_i|du \le$$

$$\le (1-\varepsilon)\delta(\varepsilon t) + (\lambda_i^* + \lambda_i)\varepsilon,$$

where $\varepsilon \in (0,1)$, $\sup_{u\ge 0} \lambda_i(u) = \lambda_i *$, $\delta(t') \to 0$, when $t' \to \infty$.

Now let $\lambda_i(t)$ be a periodic function with period T_i:

$$\lambda_i(nT_i + t) = \lambda_i(t) \qquad \text{for both} n = 1, 2, \ldots \text{ and } 0 \le t < T_i.$$

Then Condition 1 holds with $\Lambda_i^{(0)}(t) = \left(\int\limits_0^{T_i} \lambda_i(u)du\right) t$. Really,

$$\frac{\Lambda_i(t)}{t} = \frac{\int_0^t \lambda_i(u)du}{t} = \frac{[t]_{T_i}\int_0^{T_i} \lambda_i(u)du + \int_0^{(t)_{T_i}} \lambda_i(u)du}{[t]_{T_i} + \{t\}_{T_i}} \xrightarrow{t\to\infty} \lambda_i,$$

where $[t]_{T_i} = \max\{n \in Z_+ : nT_i \le t\}$, $\{t\}_{T_i} = t - [t]_{T_i}$, Z_+ is the set of nonnegative integer numbers.

Condition 2. A service rate in each node depends on the "n" (series number) so that

$$\lim_{n\to\infty} n\mu_i(n) = \mu_i > 0, \quad i = 1, 2, \ldots, r.$$

Together Conditions 1 and 2 mean that $[M_t|M|\infty]^r$ - network operating in heavy traffic regime.

In the context of Conditions 1, 2 we consider the sequence of stochastic processes:

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

where $q^{(n)'}(nt) = (q_1^{(n)}(nt), \ldots, q_r^{(n)}(nt))$, $q_j^{(n)}(nt) = \sum\limits_{i=1}^r \int\limits_0^{nt} d\Lambda_i^{(n)}(u) p_{ij}^{(n)}(nt - u)$, $j = 1, \ldots, r$, $p_{ij}^{(n)}(\tau)$ are elements of the matrix $P^{(n)}(\tau) = \|p_{ij}^{(n)}(\tau)\|_1^r = \exp\{\Delta(\mu^{(n)})(P - I)\tau\}$, $\mu^{(n)'} = (\mu_1^{(n)}, \ldots, \mu_r^{(n)})$, $\Delta(x) = \|\delta_{ij}x_i\|_1^r$ is a diagonal matrix with the vector $x' = (x_1, \ldots, x_r)$ at the principal diagonal, $I = \|\delta_{ij}\|_1^r$ is the identity matrix.

To describe the limit of the sequence of stochastic processes $\xi^{(n)}(t)$, $n \geq 1$, we introduce two independent Gaussian processes $\xi^{(i)'}(t) = (\xi_1^{(i)'}(t), \ldots, \xi_r^{(i)'}(t))$, $i = 1, 2$.

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

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

and by correlation matrixes:

$$R^{(1)}(t) = E\xi^{(1)}(t)\xi^{(1)'}(t) - E\xi^{(1)}(t)E\xi^{(1)'}(t) = \int_0^t P'(t-\tau)\Delta[d\Lambda^{(0)}(\tau)]P(t-\tau),$$

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

where $\Lambda^{(0)'}(t) = (\lambda_1^{(0)}(t), \ldots, \lambda_r^{(0)}(t))$, $P(\tau) = \exp\{\Delta(\mu)(P-I)\tau\}$.

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

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

$$R^{(2)}(t) = \int_0^t [\Delta[(d\Lambda^{(0)}(\tau))'P(t-\tau)] - P'(t-\tau)\Delta[d\Lambda^{(0)}(\tau)]P(t-\tau)],$$

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

The following theorem is the main result of the work.

Theorem 1. *Let for the $[M_t|M|\infty]^r$- network conditions 1, 2 take place. At the initial moment of time $t = 0$ the network is empty: $Q_i(0) = 0$, $i = 1, 2, \ldots, r$. Then for any finite interval $[0, T]$ the sequence of stochastic processes $\xi^{(n)}(t)$, $n \geq 1$, converges in the uniform topology to $\xi^{(1)}(t) + \xi^{(2)}(t)$.*

2 Proof of Theorem 1

Before proof of Theorem 1 we obtain some auxiliary results.

Lemma 1. *Let $\nu^{(n)}(t)$ be the Poisson process with leading function $\Lambda^{(n)}(t)$ for which Condition 1 is true. Then for any finite interval $[0, T]$ sequence of stochastic processes $W^{(n)}(t) = n^{-1/2}(\nu^{(n)}(nt) - \Lambda^{(n)}(nt))$, $n \geq 1$, converges in the uniform topology to the Wiener process $W^{(0)}(t)$ with $EW^{(0)}(t) = 0$ and $VarW^{(0)}(t) = \Lambda^{(0)}(t)$.*

Proof. Convergence of finite-dimensional distributions of the process $W^{(n)}(t)$ to $W^{(0)}(t)$ follows from the fact that for any natural number N and time moments $0 < t_1 < \ldots < t_N$ the joint characteristic function of $\nu(t_1), \ldots, \nu(t_N)$ is equal:

$$E\exp\left\{i\sum_{k=1}^{N} s(k)\nu(t_k)\right\} = \prod_{k=0}^{N-1} \exp\left\{[\Lambda(t_{k+1}) - \Lambda(t_k)]\left[\exp\left(i\sum_{m=k+1}^{N} s(m)\right) - 1\right]\right\},$$

where $(s(1), \ldots, s(N)) \in R_N$, $t_0 = 0$.

Now in order to prove convergence in uniform topology it's sufficient to check the following condition

$$\lim_{h\to 0}\overline{\lim_{n\to\infty}}\sup_{|t_1-t_2|\le h} P\left\{\left|W^{(n)}(t_2)-W^{(n)}(t_1)\right|>\varepsilon\right\}=0 \tag{2}$$

from [2] (page 493).

Based on the Chebyshev inequality

$$\sup_{|t_1-t_2|\le h} P\left\{\left|W^{(n)}(t_2)-W^{(n)}(t_1)\right|>\varepsilon\right\}\le \varepsilon^{-2}\sup_{t\in[0,T]}[n^{-1}\Lambda(n(t+h))-n^{-1}\Lambda(nt)].$$

Condition 1 implies that for any $0<\delta<T$

$$\sup_{t\in[0,\delta]}\left[n^{-1}\Lambda(n(t+h))-n^{-1}\Lambda(nt)\right]\le n^{-1}\Lambda(n(\delta+h))\le(\lambda+\varepsilon_n)(\delta+h)$$

and

$$\sup_{t\in[\delta,T]}[n^{-1}\Lambda(n(t+h))-n^{-1}\Lambda(nt)]\le\lambda h+(2T+h)\varepsilon_n.$$

Hence we find

$$\lim_{h\to 0}\overline{\lim_{n\to\infty}}\sup_{|t_1-t_2|\le h} P\left\{\left|W^{(n)}(t_2)-W^{(n)}(t_1)\right|>\varepsilon\right\}\le\lambda\varepsilon^{-2}\delta.$$

Since $\delta>0$ is arbitrary, Condition 1 holds. Lemma is proved.

Hereafter we will denote as $W_i^{(0)}(t)$, $i=1,2,\dots,r$, independent Wiener processes with $EW_i^{(0)}(t)=0$ and $VarW_i^{(0)}(t)=\Lambda_i^{(0)}(t)$. If Condition 1 takes place they approximate the input flows $\nu_i^{(n)}(t)$.

For $W^{(0)'}(t)=(W_1^{(0)}(t),\dots,W_r^{(0)}(t))$ we will need the following result.

Lemma 2. *Finite-dimensional distributions of* $\int_0^t dW^{(0)'}(u)P(t-u)$ *coincide with the finite-dimensional distributions of Gaussian process* $\xi^{(1)}(t)$.

This result is a partial case of Lemma 1 from [3].

Service of a call in nodes of the $[M_t|M|\infty]^r$ -network is independent of other calls. In order to structurally define the service process, we consider the Markov chain $x(t)$, $t\ge 0$, in the set of states $\{1,\dots,r,r+1\}$ with infinitesimal characteristics

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

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

If $p_i(0)=1$ then we will mark the corresponding chain as $x^{(i)}(t)$. State "$r+1$" for the chain $x(t)$ is absorbing. Transitional probabilities of $x(t)$

$$p_{ij}(t)=P\{x(t)=j/x(0)=i\}=P\{x^{(i)}(t)=j\},\quad i,j=1,\dots,r$$

are elements of the matrix $P(t)=\exp\{\Delta(\mu)(P-I)t\}$.

The path of the call from the input moment of time in the network through i-th node to output of it can be described by the chain $x^{(i)}(t)$. The absorption in the state "$r+1$" is interpreted as the output of call from the network.

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

$$\chi^{(i)}(t) = \begin{cases} e_j, \ x^{(i)}(t) = j, \ j = 1, \ldots, r; \\ e_0, \ x^{(i)}(t) = r+1; \end{cases}$$

where e_j is r-dimensional vector with the j-th component equal to 1, while others are zero, e_0 is zero r-dimensional vector.

For arbitrary natural N and $z'(i) = (z_1(i), \ldots, z_r(i))$, $i = 1, 2, \ldots, N$, $|z(i)| \leq 1$, we will denote a joint generating function of the vectors $\chi^{(m)}(t_1), \ldots, \chi^{(m)}(t_N)$, $0 < t_1 < \ldots < t_N$, as $\Phi^{(m)} = \Phi^{(m)}(t_1, \ldots, t_N, z(1), \ldots, z(N))$, $\Phi' = (\Phi^{(1)}, \ldots, \Phi^{(r)})$.

Lemma 3. *For any $N = 1, 2, \ldots$ and $0 < t_1 < \ldots < t_N$*

$$\Phi = \bar{1} + \sum_{i=1}^{N} P(\Delta t_1)\Delta[z(1)] \ldots P(\Delta t_{i-1})\Delta[z(i-1)]P(\Delta t_i)(z(i) - \bar{1}), \qquad (3)$$

where $\bar{1}$ is r-dimensional vector composed of units, $\Delta t_i = t_i - t_{i-1}$ $(t_0 = 0), i = 1, \ldots, N$. .

The proof of (3) can be obtained by mathematical induction on the parameter N.

Proof. (of Theorem 1) Let us analyze the behavior of one-dimensional distributions of the process $\xi^{(n)}(t)$, $t \geq 0$, as $n \to \infty$.

Under the fixed path of input flow $\nu(t)$ the distribution of $Q(t)$ coincides with the distribution of

$$\sum_{m=1}^{r} \sum_{k=1}^{\nu_m(t)} \chi^{(m,k)}(t - \tau_k^{(m)}),$$

where $\chi^{(m,1)}(t), \chi^{(m,2)}(t), \ldots$ is a sequence of independent stochastic processes with finite-dimensional distributions coinciding with $\chi^{(m)}(t)$, $\tau_k^{(m)}$ is the arrival moment of time of k-th call to the m-th node.

Taking into account this fact and the formula (2) under $N = 1$ the generation function $\Phi(t, z)$, $z' = (z_1, \ldots, z_r)$, $|z| \leq 1$, of the vector $Q(t)$ can be represented as:

$$\Phi(t, z) = E \prod_{m=1}^{r} \prod_{k=1}^{\nu_m(t)} [1 - p'_m(t - \tau_k^{(m)})(z - \bar{1})], \qquad (4)$$

where $p'_m(\tau) = (p_{m1}(\tau), \ldots, p_{mr}(\tau))$ is the m-th line of matrix $P(\tau)$.

Let $\phi^{(n)}(s)$, $s' = (s_1, \dots, s_r) \in R_r$ be a characteristic function of $\xi^{(n)}(t)$. In view of (4)

$$\phi^{(n)}(s) = Ee^{i\xi^{(n)'}s} = \exp\left\{-in^{-1/2}q^{(n)'}(nt)s\right\} \times$$

$$\times E\exp\left\{\sum_{m=1}^{r}\sum_{k=1}^{\nu_m^{(n)}(nt)} \ln[1 - p_m^{(n)'}(nt - \tau_k^{(m)})(e^{is/\sqrt{n}} - \bar{1})]\right\},$$

where $\left(e^{is/\sqrt{n}}\right)' = \left(e^{is_1/\sqrt{n}}, \dots, e^{is_r/\sqrt{n}}\right)$.

Let us denote by $(s^2)' = (s_1^2, \dots, s_r^2)$. Then

$$\lim_{n\to\infty}\phi^{(n)}(s) = \lim_{n\to\infty}\exp\left\{-in^{-1/2}q^{(n)'}(nt)s\right\} \times$$

$$\times E\exp\left\{\sum_{m=1}^{r}\sum_{k=1}^{\nu_m^{(n)}(nt)}\left[\frac{i}{\sqrt{n}}p'_m(t - \frac{\tau_k^{(m)}}{n})s - \frac{1}{2}\frac{1}{n}p'_m(t - \frac{\tau_k^{(m)}}{n})s^2 + \right.\right.$$

$$\left.\left. + \frac{1}{2}\frac{1}{n}s'p_m(t - \frac{\tau_k^{(m)}}{n})p'_m(t - \frac{\tau_k^{(m)}}{n})s\right]\right\} =$$

$$= \lim_{n\to\infty}\exp\left\{-in^{-1/2}q^{(n)'}(nt)s\right\} E\exp\left\{in^{-1/2}\int_0^t d\nu^{(n)'}(n\tau)P(t-\tau)s - \right.$$

$$\left. - \frac{1}{2}\frac{1}{n}\int_0^t d\nu^{(n)'}(n\tau)P(t-\tau)s^2 + \frac{1}{2}\frac{1}{n}\sum_{m=1}^{r}\int_0^t d\nu_m^{(n)'}(n\tau)s'p_m(t-\tau)p'_m(t-\tau)s\right\} =$$

$$= \lim_{n\to\infty}\exp\left\{-\frac{1}{2}\frac{1}{n}\int_0^t d\Lambda^{(n)'}(n\tau)P(t-\tau)s^2 + \right.$$

$$\left. + \frac{1}{2}\frac{1}{n}\sum_{m=1}^{r}\int_0^t d\Lambda_m^{(n)}(n\tau)s'p_m(t-\tau)p'_m(t-\tau)s\right\} E\exp\left\{i\int_0^t dW^{(n)'}(\tau)P(t-\tau)s\right\},$$

where $W^{(n)'}(\tau) = (W_1^{(n)}(\tau), \dots, W_r^{(n)}(\tau))$, $W_k^{(n)}(\tau) = n^{-1/2}(\nu_k^{(n)}(n\tau) - \Lambda_k^{(n)}(n\tau))$, $k = 1, \dots, r$.

Using Condition 1 and Lemmas 1, 2 we find

$$\lim_{n\to\infty}\phi^{(n)}(s) = \exp\left\{-\frac{1}{2}\int_0^t d\Lambda^{(0)'}(\tau)P(t-\tau)s^2 + \right.$$

$$\left. + \frac{1}{2}\sum_{m=1}^{r}\int_0^t d\Lambda_m^{(0)}(\tau)s'p_m(t-\tau)p'_m(t-\tau)s - \frac{1}{2}s'\int_0^t P'(t-\tau)\Delta[d\Lambda^{(0)}(\tau)]P(t-\tau)s\right\} =$$

$$= \exp\left\{ -\frac{1}{2}s' \int_0^t [\Delta[d\Lambda^{(0)'}(\tau)P(t-\tau)] - P'(t-\tau)\Delta[d\Lambda^{(0)}(\tau)]P(t-\tau)]s - \right.$$

$$\left. -\frac{1}{2}s' \int_0^t P'(t-\tau)\Delta[d\Lambda^{(0)}(\tau)]P(t-\tau)s \right\}.$$

The limit is a characteristic function of $\xi^{(1)}(t) + \xi^{(2)}(t)$. Thus, the convergence of one-dimensional distributions is proved.

Now we consider the two-dimensional distributions. Under the fixed path of input flow the distribution of $(Q(t_1), Q(t_2))$, $0 < t_1 < t_2$, coincides with the distribution of

$$\sum_{m=1}^{r} \left(\sum_{k=1}^{\nu_m(t_1)} \chi^{(m,k)}(t_1 - \tau_k^{(m)}), \sum_{k=1}^{\nu_m(t_1)} \chi^{(m,k)}(t_2 - \tau_k^{(m)}) + \sum_{k=\nu_m(t_1)+1}^{\nu_m(t_2)} \chi^{(m,k)}(t_2 - \tau_k^{(m)}) \right).$$

Using the formula (2) for $N = 2$ a joint generating function $\Phi(t_1, t_2, z(1), z(2))$ of vectors $Q(t_1), Q(t_2)$ can be represented as follows:

$$\Phi(t_1, t_2, z(1), z(2)) = E\left\{ \prod_{m=1}^{r} \prod_{k=1}^{\nu_m(t_1)} [1 + p'_m(t_1 - \tau_k^{(m)})(z(1) - \bar{1}) + \right.$$

$$\left. + p'_m(t_1 - \tau_k^{(m)})\Delta[z(1)]P(\Delta t_2)(z(2) - \bar{1})] \prod_{k=\nu_m(t_1)+1}^{\nu_m(t_2)} [1 + p'_m(t_2 - \tau_k^{(m)})(z(2) - \bar{1})] \right\}.$$

From here we find the limit for the joint characteristic function $\phi^{(n)}(s(1), s(2))$, $s(1), s(2) \in R_r$, of vectors $\xi^{(n)}(t_1)$ and $\xi^{(n)}(t_2)$.

Similarly we can check convergence of N-dimensional distributions for $N > 2$.

The resulting convergence of finite-dimensional distributions can be strengthened to convergence of $\xi^{(n)}(t)$ in the uniform topology. To do this it is necessary to use the convergence of normalized input flow $W^{(n)}(t)$ to the $W^{(0)}(t)$ in the uniform topology and the representation of the service process as the amount of indicator-type processes at the input flow.

Theorem is proved.

Part $\xi^{(1)}(t)$ of the limit process is associated with fluctuations of input flows and $\xi^{(2)}(t)$ with fluctuations of service times.

3 Limit Process as a Multidimensional Diffusion

In the one-dimensional case for Gaussian processes there is a criterion in terms of necessity and sufficiency to verify the Markov property ([7], Theorem 1 on page 115). In the multydimensional case the situation becomes more complicated and there is no general criterion. We will present a variant of the sufficient condition from [4] for the Markov property of r-dimensional Gaussian processes and apply this criterion to the limit process from the previous section.

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

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

Theorem 2. *Let for some matrix A and for all s, t $(0 \le s < t)$ the functions $R(s)$ and $R(s,t)$ be related by the following way:*

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

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

The set G_A of Gaussian processes for which the condition of Theorem 1 takes place and the corresponding matrices are the same (equal A) satisfies to the closure condition: a linear combination of two independent processes from G_A belongs G_A . Thus, as a consequence from Theorem 2 we obtain the following interesting fact: the sum of two independent Markov G_A-processes is a Markov process.

Note, that the many-dimensional Ornstein-Uhlenbeck process (see [6], page 166) satisfies the condition of Theorem 2.

Let us apply the criterion of Markov behaviour which is given by Theorem 2 to the limiting Gaussian process $\xi^{(1)}(t) + \xi^{(2)}(t)$.

Corollary 1. *If $\Lambda_i^{(0)}(t) = \int_0^t \lambda_i^{(0)}(u)du$, $\lambda_i^{(0)}(u) \in C[0,T]$, $i = 1, 2, \ldots, r$, then the limiting Gaussian process $\xi^{(1)}(t) + \xi^{(2)}(t)$, $t \in [0,T]$, is an r-dimensional diffusion with drift vector $A(x) = A'x$ and the diffusion matrix*

$$B(t) = \Delta[\lambda^{(0)'}(t) + q'(t)A] - A'\Delta[q(t)] - \Delta[q(t)]A,$$

where $A = \Delta(\mu)(P - I)$, $q'(t) = \int_0^t \lambda^{(0)'}(\tau)P(t-\tau)d\tau$, $\lambda^{(0)'}(\tau) = (\lambda_1^{(0)}(\tau), \ldots, \lambda_r^{(0)}(\tau))$.

If we denote as $R(t), R(s,t), s < t$, the corresponding correlation matrices of the process $\xi^{(1)}(t)+\xi^{(2)}(t)$ then for them the condition of Theorem 2 will be satisfied under $A = \Delta(\mu)(P - I)$ and $\xi^{(1)}(t) + \xi^{(2)}(t)$ will be Markov diffusion process. Drift vector and diffusion matrix are determined by the form of conditional distribution $P(\xi(t) \in B/\xi(s) = x)$, $B \in B_{R^r}$.

Thus, Corollary 1 relates to the method of the diffusion approximation of overloaded stochastic systems and networks. In this sense it extends the results of section 4.2 of [5] to the case of Poisson input flow with a varying rate.

The form of the limiting process as a many-dimensional diffusion is attractive in that the diffusion process is determined by only its local characteristics and we can use the developed tools of Markov diffusion processes for analysis of $\xi^{(1)}(t) + \xi^{(2)}(t)$. However, there is a loss because now the limiting process does not reflect in detail the structure of the prelimiting service process.

References

1. Gnedenko, B.V., Kovalenko, I.N.: An introduction to Queueing Theory, Moscow (2005)
2. Gihman, I.I., Skorohod, A.V.: Theory of stochastic processes, Moscow, Nauka, vol. 1 (1971)
3. Lebedev, E.A.: A limit theorem for stochastic networks and its applications. Theor. Probability and Math. Statist. 68, 81–92 (2003)
4. Lebedev, E.A.: On the Markov property of multivariate Gaussian processes. Physics-Mathematics, Bulletin of Kiev University, vol. 4, pp. 287–291 (2001)
5. Anisimov, V.V., Lebedev, E.A.: Stochastic service networks. Markov models. Kyiv, Lybid (1992)
6. Kovalenko, I.N., Kuznetsov, I.N., Shurenkov, N.Y.: Stochastic Processes. Kyiv, Naukova Dumka (1983)
7. Feller, W.: An introduction to probability theory and its applications, Moscow, Mir, vol. 2 (1984)

Performance Evaluation of Finite Buffer Queues through Regenerative Simulation

Oleg Lukashenko[1], Evsey Morozov[1], Ruslana Nekrasova[1], and Michele Pagano[2]

[1] Karelian Research Center RAS, Petrozavodsk State University
{lukashenko-oleg,ruslana.nekrasova}@mail.ru, emorozov@karelia.ru
[2] University of Pisa
m.pagano@iet.unipi.it

Abstract. In this paper we discuss the estimation of the loss probability in a queueing system with finite buffer fed by Brownian traffic, the Gaussian counterpart of the well-known Poisson process. The independence among arrivals in consecutive time slots allows the application of regenerative simulation technique, combined with the so-called Delta-method to construct confidence intervals for the stationary loss probability. Numerical simulation are carried out to verify the efficiency of the regenerative approach for different values of the queue parameters (buffer size and utilization) as well as simulation settings (digitization step and generalizations of the regeneration cycle).

1 Introduction

We consider a single server queue with finite buffer of size b, constant service rate C and cumulative input process

$$A(t) = mt + \sqrt{m}B(t), \tag{1}$$

given by the superposition of a deterministic linear term mt with positive drift $m > 0$ (corresponding to the mean arrival rate) and an adequately scaled version of the Brownian motion (BM) $\{B(t)\}$, with Var $A(1) = m$. The resulting process $A(t)$ is known in the literature as the Gaussian counterpart of the Poisson stream [6].

The workload process Q_n in this queueing system, which will be denoted in the following as Bi/D/1/b, is described by the well-known (discrete time) Lindley recursion:

$$Q_n = \min((Q_{n-1} - C + X_n)^+, b),\ n = 1, 2, \ldots, \tag{2}$$

where

$$X_n := A(n+1) - A(n) =_{st} m + N(0, m)$$

represents the amount of work arriving during the n^{th} time slot and $=_{st}$ means stochastic equality. The increments X_n are i.i.d. random variables and in the following X will denote the generic element.

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 131–139, 2013.

Denote by $L_b(T)$ the workload lost during the interval $[0, T]$, that is

$$L_b(T) := \sum_{k=1}^{T} (Q_{k-1} - C + X_k - b)^+.$$

The time average loss $\mathsf{P}_\ell(b, T)$ is defined as the ratio between the amount of lost workload and the arrived workload during $[0, T]$, that is

$$\mathsf{P}_\ell(b, T) := \frac{L_b(T)}{A(T)}. \tag{3}$$

Due to the finite buffer size, the system is stable and the loss ratio converges to the stationary loss probability $\mathsf{P}_\ell(b)$, that is

$$\mathsf{P}_\ell(b) := \lim_{T\to\infty} \mathsf{P}_\ell(b, T) = \frac{\mathsf{E}(Q + X - C - b)^+}{m}, \tag{4}$$

where Q is the stationary workload. The following heuristic expression given in [5]

$$\mathsf{P}_\ell(b) \approx \frac{\mathsf{P}_\ell(0)}{\mathsf{P}(Q > 0)} \mathsf{P}(Q > b), \tag{5}$$

allows us to calculate the loss probability provided there is an explicit formula (or a satisfactory approximation) for the overflow probability $\mathsf{P}(Q > b)$ in the associated infinite buffer system. In our case, it is possible to use the following continuous-time approximation (see [8]):

$$\mathsf{P}(Q > b) \approx \exp\left(-2 \cdot \frac{C - m}{m} \cdot b\right). \tag{6}$$

Moreover, it is easy to calculate $\mathsf{P}_\ell(0)$, namely,

$$\begin{aligned} \mathsf{P}_\ell(0) &= \frac{\mathsf{E}(X - C)^+}{m} \\ &= \frac{1}{m^{3/2}\sqrt{2\pi}} \int_C^\infty (x - C) e^{-(x-m)^2/2m^2} dx. \end{aligned} \tag{7}$$

Thus results (5)–(7) allow us to find an approximation of the overflow probability $\mathsf{P}_\ell := \mathsf{P}_\ell(b)$ in our model.

2 Regenerative Approach

In this section, we show how to estimate the steady-state loss probability P_ℓ using the regenerative approach. First we construct regeneration points for the workload process (see also [4]). Let $\beta_0 = 0$ and

$$\beta_{k+1} = \min\{n > \beta_k : Q_{n-1} = 0,\, Q_n > 0\},\ k \geq 0, \tag{8}$$

where Q_n is defined by recursion (2). Before estimating the stationary loss probability, we must be sure that the workload process is positive recurrent, i.e. the mean regeneration period must be finite: $\mathsf{E}\beta < \infty$, where β denotes the generic regeneration period.

To prove that $\mathsf{E}\beta < \infty$ in our finite buffer system we require that, starting in a compact set, the process Q_t, $t \geq 0$, hits the regeneration (zero) state in a finite time with a positive probability [7].

It is easy to see that this requirement holds if the traffic intensity $\rho := m/C < 1$. Indeed, for an arbitrary instant t, consider the event $\mathsf{D}(t) = \{Q_t \leq b\}$ and note that $\mathsf{P}(\mathsf{D}(t)) = 1$ regardless of t. Let $m - C = -\varepsilon < 0$ and note that during each slot the accumulated workload reduces *in average* by the quantity $\varepsilon > 0$ (provided the system is not empty). Also consider the i.i.d. sequence N_i, $i \geq 1$, where each N_i is distributed as standard normal variable $N(0, 1)$. Consider the events $\omega_i = \{N_i \leq 0\}$, so $\mathsf{P}(\omega_i) = 1/2$ for all i. Denote by $R = \lceil b/\varepsilon \rceil$ the smallest integer not less than b/ε; then, regardless of t, the workload process reaches the regeneration state in any interval $[t, t + R]$ with a probability q, which is lower bounded by a positive constant as follows

$$q \geq \left[\frac{1}{2}\right]^R > 0.$$

Thus, the regeneration condition is satisfied, and it implies positive recurrence of the process of regenerations for $\rho < 1$.

As the queue content is upper bounded by the buffer size b, the queueing system is stable also when $\rho \geq 1$. In this case $C \leq m$, and we can take into account the negative values of the BM to compensate the shortage of the server capacity $C - m \leq 0$. Indeed, during each slot $[t, t + 1)$ the absence of newly arrived workload has probability

$$\begin{aligned}&\mathsf{P}(m + \sqrt{m}B(1) \leq 0)\\ &= \mathsf{P}(N(0, 1) \leq -\sqrt{m}) = \frac{1}{2} - \Phi(\sqrt{m}) := \delta > 0,\end{aligned}$$

where Φ denotes the Laplace function. Introduce the events $\omega_i = \{N_i < -\sqrt{m}\}$ and realize $R_1 = \lceil b/C \rceil < \infty$ such events. Then we obtain (as above) that the workload process reaches zero during interval $[t, t + R_1]$ with a probability which is lower bounded by a positive constant $\delta^{R_1} > 0$.

Thus, we have established that the workload process indeed reaches regeneration in a finite interval with positive probability, that is regeneration condition holds. It means that the renewal process of regenerations is positive recurrent for all values of ρ.

Denote by L_i and A_i the workload lost and arrived during the i^{th} regeneration cycle, respectively. The regenerative method leads to the following representation of the steady-state loss probability

$$\mathsf{P}_\ell = \frac{\mathsf{E}L}{\mathsf{E}A}$$

where the unknown means $\mathsf{E}L$ and $\mathsf{E}A = m\mathsf{E}\beta < \infty$ (as before, L and A denote the corresponding generic elements) can be estimated from n i.i.d. replications $L_1, .., L_n$, $A_1, ..., A_n$:

$$\widehat{L} := \widehat{L}(n) = \frac{1}{n}\sum_{i=1}^{n} L_i, \quad \widehat{A} := \widehat{A}(n) = \frac{1}{n}\sum_{i=1}^{n} A_i, \quad \widehat{\mathsf{P}}_\ell := \widehat{\mathsf{P}}_\ell(n) = \frac{\widehat{L}}{\widehat{A}}. \tag{9}$$

Using the Delta-method [1,2], it is possible to construct confidence intervals for P_ℓ. Let Z_i, $i = 1, 2$ be some random variables. Actually we need to find an estimation for

$$f(z) = f(z_1, z_2), \ z_i = \mathsf{E}Z_i, \qquad i = 1, 2, \tag{10}$$

where f is a sufficiently smooth function (in our case $f(z_1, z_2) = \frac{z_1}{z_2}$). It is reasonable to set

$$f_n(\widehat{z}) = f(\widehat{z}_1, \widehat{z}_2), \ \widehat{z}_i = \frac{1}{n}\sum_{k=1}^{n} Z_i^{(k)} \qquad i = 1, 2, \tag{11}$$

where $\{Z_i^{(k)}\}$ are i.i.d. replications of Z_i. Using Taylor expansion

$$f_n(\widehat{z}) - f(z) = \nabla f(z)(\widehat{z} - z) + o(||\widehat{z} - z||), \ \text{ where } \nabla f = \left(\frac{\partial f}{\partial z_1}, \frac{\partial f}{\partial z_2}\right)$$

and $||x||$ is Euclidean norm. It is possible to show that

$$\sqrt{n}\,(f_n(\widehat{z}) - f(z)) \Rightarrow N(0, \sigma^2), \ n \to \infty, \tag{12}$$

where $\Rightarrow$ stands for weak convergence and

$$\sigma^2 = \nabla f \cdot \Sigma \cdot (\nabla f)', \ \Sigma = (\mathrm{Cov}(\mathrm{Z_i}, \mathrm{Z_j}))_{1 \le i,j \le 2}\,.$$

In particular, for the loss probability convergence (12) can be rewritten in the following form:

$$\sqrt{n}\left(\widehat{\mathsf{P}}_\ell - \mathsf{P}_\ell\right) \Rightarrow N(0, \eta^2), \ \ n \to \infty, \tag{13}$$

where

$$\eta^2 = \frac{\mathsf{E}\left[L - A \cdot \mathsf{P}_\ell\right]^2}{(\mathsf{E}A)^2}.$$

and, applying the standard sample estimator, we get

$$\widehat{\eta}^2 := \widehat{\eta}^2(n) = \frac{\frac{1}{n-1}\sum\limits_{i=1}^{n}(L_i - \widehat{\mathsf{P}}_\ell A_i)^2}{\left(\frac{1}{n}\sum\limits_{i=1}^{n} A_i\right)^2} \tag{14}$$

Based on (13), the $(1 - \gamma/2)\%$ confidence interval for P_ℓ is given by

$$\left[\widehat{\mathsf{P}}_\ell - \frac{t_\gamma \widehat{\eta}}{\sqrt{n}}, \ \widehat{\mathsf{P}}_\ell + \frac{t_\gamma \widehat{\eta}}{\sqrt{n}}\right], \tag{15}$$

where $t_\gamma = \Phi^{-1}\left(\frac{\gamma}{2}\right)$, $\Phi^{-1}(x)$ is the inverse of Laplace function and γ is a given confidence level.

Actually, the estimator (11) can be biased, and it is useful to estimate the possible bias. To this aim let us consider the second order Taylor expansion of f (assuming twice differentiability of f, which is automatically fulfilled in our case):

$$f(\widehat{z}) - f(z) = \nabla f(z)(\widehat{z} - z) + \frac{1}{2}(\widehat{z} - z)' H(z)(\widehat{z} - z) + o(||\widehat{z} - z||^2),$$

where $H(z) = (H_{ij}(z))_{1 \le i,j \le 2}$ is the Hessian matrix of f. Then

$$\mathsf{E} f(\widehat{z}) - f(z) = \frac{1}{2n} \sum_{i,j=1}^{2} \mathrm{Cov}(Z_i, Z_j) H_{ij}(z) + o(1/n), \ n \to \infty. \tag{16}$$

Hence it seems reasonable to use the modified estimator $f(\widehat{z}) - g(\widehat{z})$, where

$$g(\widehat{z}) = \frac{1}{2n} \sum_{i,j=1}^{2} \widehat{\mathrm{Cov}}(Z_i, Z_j) H_{ij}(\widehat{z}) \tag{17}$$

and $\widehat{\mathrm{Cov}}(Z_i, Z_j)$ is the sample covariance. The relevance of this bias depends on the length of the simulation and on the system parameters, but typically it is negligible (see the numerical results in the next section).

3 Numerical Results

The regenerative approach is applied to the above considered system $Bi/D/1/b$ in order to estimate its stationary loss probability. In more detail, regeneration points are constructed according to (8) and then confidence intervals for the probability P_ℓ are calculated according to (15).

Several simulation tests have been carried out in order to analyze the different issues raised in the previous sections. Unless otherwise stated, the following values of the system parameters are considered: mean arrival rate $m = 0.8$, service rate $C = 1$, buffer size $b = 4$, simulation length $T = 10^5$ and 95% confidence level.

The first set of simulation aimed at checking the relevance of the bias as a function of the simulation length. As highlighted by Fig. 1, $|g(\widehat{z})|$ rapidly decreases with the simulation length T and this justifies the value $T = 10^5$ chosen for the remaining sets of simulations (as can be seen in Fig. 2, $\mathsf{P}_\ell \approx 0.033$ for the chosen values of the queue parameters).

Another relevant issue in discrete-time simulation is the choice of the digitization step: indeed, we considered a continuous-time system only at discrete points in time and this typically introduces some kind of approximation in the estimated parameters (see, for instance, [2,3] and [6] for a more detailed analysis of the problem). To this aim, we simulated the queueing system over non

overlapping time slots of length $h = 1/N$ and Fig. 2 shows the behaviour of P_ℓ (with its 95% confidence interval) as a function of N, which is rather insensitive to the selection of the concrete value of step digitization h in a wide range of values of N. This remark permits to choose a relatively small value of N (in the following $N = 10$ will be considered), saving simulation time (indeed the number of simulated slots is given by NT).

As far as the simulation set-up is concerned, we finally studied the dependence of the estimates on the choice of the regeneration points. Namely, we built *subsequences* of regeneration points $\{\beta_k^s\}$ for several values of s as $\beta_k^s = \beta_{sk}$ in order to estimate the effect of cycle aggregation (for the same fixed length of the simulation interval, i.e. $T = 10^5$). As highlighted by Fig. 3, the estimation of P_ℓ very weakly depends on the parameter s, so it is reasonable to make use of the *standard* regenerative simulation (i.e., with $s = 1$).

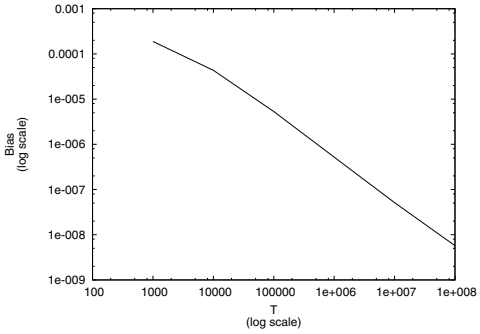


Fig. 1. Behaviour of $|g(\widehat{z})|$ in the Bi/D/1/4 queue

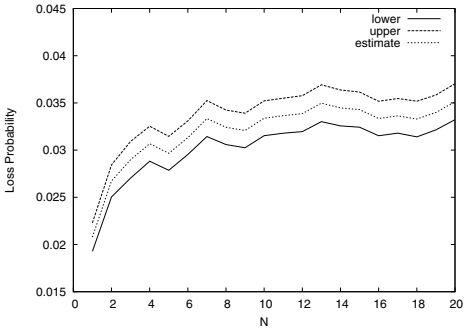


Fig. 2. Effect of the digitization step in the Bi/D/1/4 queue

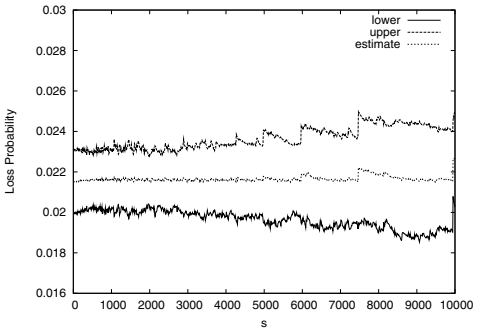


Fig. 3. Effect of alternative choices of the regeneration points in the Bi/D/1/4 queue

Fig. 4 compares the simulation results (considering the settings discussed above) with the analytical approximation (5) for different values of the buffer size b, confirming the goodness of the proposed approach.

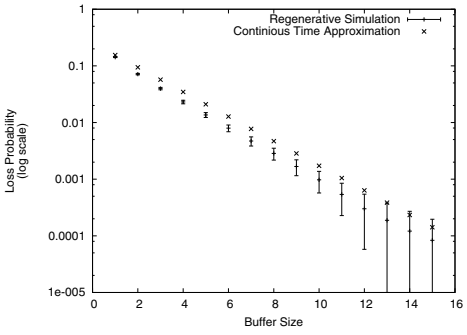


Fig. 4. Estimates of P_ℓ in the Bi/D/1/b queue for $m = 0.8$: regenerative method vs. approximation (5)

Due to the finite buffer size, regenerative simulation can also be applied when $\rho > 1$; in this case the average length of the regeneration cycles grows (as reported in Table 1 for different values of b in the two considered scenarios), but the estimation is still quite accurate as shown in Fig. 5 for $m = 1.2$. In this case it is not possible to compare the simulation results with the approximation (5) since the latter relies on the stability of the corresponding infinite-buffer system.

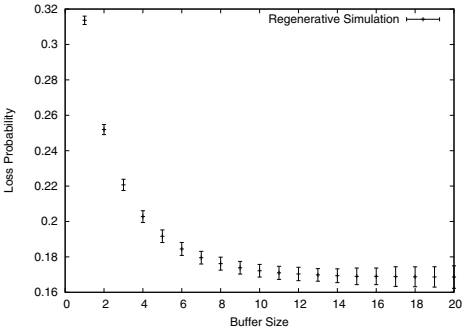


Fig. 5. Estimates of P_ℓ through regenerative simulation in the Bi/D/1/b queue for $m = 1.2$

Table 1. Average length of the regeneration cycles for Bi/D/1/b queue

Buffer Size	m=0.8	m=1.2
1	2.329	4.033
2	2.846	6.928
3	3.164	11.118
4	3.365	16.917
5	3.481	25.156
6	3.564	35.726
7	3.609	50.431
8	3.636	69.365
9	3.655	97.955
10	3.667	138.441
11	3.676	192.323
12	3.679	262.308
13	3.680	357.794
14	3.682	474.206
15	3.682	683.510

4 Conclusion

Loss probability is a relevant Quality of Service parameter in computer networks and simulation is a powerful tool for its estimation, provided that some information about its accuracy is available. In this work we considered a single-server queue with finite buffer fed by Brownian traffic, the Gaussian counterpart of the well-known Poisson process. The nature of the input process allowed us to construct confidence intervals for the stationary loss probability by combining regenerative simulation with the so-called Delta-method.

Through several sets of simulations, at first we discussed the effect of different parameters related to the simulation set-up, which may have a relevant impact on the simulation time. In more detail, we considered the entity of biasing, the effect of the digitization step and possible generalizations in the definition of

the regeneration cycles. Then we applied the regenerative approach for different values of the buffer size, confirming that the regenerative method efficiently works also when the corresponding infinite buffer system is unstable (i.e., for $\rho > 1$).

Finally, a known approximation of the loss probability via the overflow probability was used to verify the accuracy the estimation.

Acknowledgment. This work is partially supported by the strategic development program of Petrozavodsk State University for 2012-2016 and Russian Foundation for Basic research, project No 10-07-00017.

References

1. Asmussen, S.: Applied Probability and Queues. Springer (2002)
2. Asmussen, S., Glynn, P.: Stochactic Simulation: algorithms and analysis. Springer (2007)
3. Asmussen, S., Glynn, P., Pitman, J.: Discretization Error in Simulation of One-Dimensional Reflecting Brownian Motion. Ann. Appl. Probab. 5(4), 875–896 (1995)
4. Goricheva, R.S., Lukashenko, O.V., Morozov, E.V., Pagano, M.: Regenerative analysis of a finite buffer fluid queue. In: Proceedings of ICUMT, pp. 1132–1136 (2010)
5. Kim, H.S., Shroff, N.B.: Loss Probability Calculations and Asymptotic Analysis for Finite Buffer Multiplexers. IEEE/ACM Transactions on Networking 9, 755–768 (2001)
6. Mandjes, M.: Large Deviations of Gaussian Queues. Wiley, Chichester (2007)
7. Morozov, E., Delgado, R.: Stability analysis of regenerative queues. Automation and Remote Control 70(12), 1977–1991 (2009)
8. Takacs, L.: Combinatorial Methods in the Theory os Stochastic Processes. John Wiley&Sons (1967)

Finite Source Retrial Queues with State-Dependent Service Rate

Vadym Ponomarov and Eugene Lebedev

Taras Shevchenko National University of Kyiv

Abstract. This paper deals with the research of finitesource retrial queues whose service rate depends on queue length. Two- and three-dimensional models that describe threshold and hysteresis control policies are taken into account. Explicit vector-matrix representations of stationary distributions are main results in both cases.

Keywords: retrial queue, state-dependent service rate, stationary probabilities, optimization.

1 Introduction

Multi-channel systems with repeated calls and finite number of primary sources are an important class of the models in which arrival calls are not lost but repeat the attempt to receive service in case of failure. In contrast to the systems with Poisson input flow such models take into account the fact that the primary flow rate decreases with the growth of users in the system. Systems with the finite number of primary sources are often used in practice for cellular networks modeling (see [1], [2]). Additional cases of such models usage can be found in [3].

The first priority problem for retrial queues is the research of steady state of the process $X(t) = (C(t); N(t))$, where $C(t)$ - number of busy servers, $N(t)$ - number of retrials. For the classic Markov model (see, for example, [4], page 269) $X(t)$ is a Markov chain in state space $S(X) = \{0, 1, \ldots, c\} \times \{0, 1, \ldots, n - c\}$ (c - number of servers, n - number of primary sources), whose infinitesimal characteristics are build according to primary calls generation rate λ, service rate ν and repeated attempts Poisson flow rate μ related to the failed call.

In this paper the classic model becomes more complex as the service rate depends on the number of repeated calls. We consider two cases of such a dependency. In the first case the service process remains two-dimensional but in the second one the third component has to be added to save the Markov property. This generalization of the service process allows to model threshold and hysteresis control policies of the service rate, in order to formulate and solve optimization problems.

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 140–146, 2013.

2 Two-Dimensional Service Process for State-Dependent Service Rate

To define the service process for state-dependent service rate we will realize the following substitution in the classic model: $\nu = \nu_j > 0, j = 0, 1, \ldots, n-c$, where j – number of retrials. The structure of infinitesimal characteristics dependence on system's parameters remains the same.

Let us analyze steady state of the process $X(t) = (C(t), N(t))$. To formulate the main result we introduce the necessary notations.

Let $\pi_{ij}, (i,j) \in S(X)$ be stationary probabilities of the process $X(t)$, $\pi_j = (\pi_{0j}\ \pi_{1j}\ \ldots\ \pi_{c-1j})^T$, $e_i = (\delta_{i0}\ \delta_{i1}\ \ldots\ \delta_{ic-1})^T$, $\overline{1}$ – c-dimensional vector that consists of 1.

$A_j = \left\| a_{ik}^j \right\|_{i,k=0}^{c-1}$ – three-diagonal matrix with elements:

$$a_{ii}^j = (n-i-j)\lambda + j\mu + i\nu_j, i = 0, 1, \ldots, c-1,$$

$$a_{ii+1}^j = -(i+1)\nu_j, i = 0, 1, \ldots, c-2,$$

$$a_{ii-1}^j = -(n+1-i-j)\lambda, i = 1, 2, \ldots, c-1.$$

$B_j = \left\| b_{ik}^j \right\|_{i,k=0}^{c-1}$, where

$$b_{ii-1}^j = (j+1)\mu, i = 1, 2, \ldots, c-2,$$

$$b_{c-1i}^j = \frac{c(j+1)\mu\nu_j}{(n-c-j)\lambda}, i \neq c-2,$$

$$b_{c-1c-2}^j = \frac{(j+1)\mu\left((n-c-j)\lambda + c\nu_j\right)}{(n-c-j)\lambda},$$

all other elements are equal to 0.

$C = \|c_{ik}\|_{i,k=0}^{c-1}$, where $(c_{00}\ c_{10}\ \ldots\ c_{c-10})^T = e_0$ and $c_{ik} = a_{i-1k}^{n-c}$, when $i \neq 0$.

$$\Phi_j = \left(\prod_{i=j}^{n-c-1} A_i^{-1} B_i \right) C^{-1} e_0, j = 0, 1, \ldots, n-c.$$

Theorem 1. *Steady state probabilities of the system are defined through system's parameters in the following form:*

$$\pi_j = \Phi_j \pi_{0n-c}, j = 0, \ldots, n-c,$$

$$\pi_{cj} = \frac{(j+1)\mu}{(n-c-j)\lambda} \overline{1}^T \Phi_{j+1} \pi_{0n-c}, j = 0, \ldots, n-c-1,$$

$$\pi_{cn-c} = \frac{\left[\lambda + (n-c)\mu + (c-1)\nu_{n-c}\right] e_{c-1}^T - 2\lambda e_{c-2}^T}{c\nu_{n-c}} C^{-1} e_0 \pi_{0n-c},$$

where

$$\pi_{0n-c} = \left\{ \overline{1}^T \sum_{j=0}^{n-c} \frac{(n-c+1-j)\lambda + j\mu}{(n-c+1-j)\lambda} \Phi_j + \right.$$

$$\left. \frac{[\lambda + (n-c)\mu + (c-1)\nu_{n-c}]\, e_{c-1}^T - 2\lambda e_{c-2}^T}{c\nu_{n-c}} \Phi_{n-c} \right\}^{-1}.$$

Proof. To find probabilities π_{ij} let us use the theorem about the equality of probability flows balance through closed path of the phase space in the steady state ([5] p.49). For each $j = 1, 2, \ldots, n-c$ we use decomposition of the phase space $S(X) = S_j^{(1)}(X) \cup \bar{S}_j^{(1)}(X)$, $S_j^{(1)}(X) = \{(p,q) \in S(X) : \ q \le j\}$. Equating probability flows through the border of subset $S_j^{(1)}(X)$, we obtain:

$$(n-c+1-j)\lambda\pi_{cj-1} = j\mu \sum_{i=0}^{c-1} \pi_{ij}, \quad j = 1, \ldots, n-c. \tag{1}$$

Then for $i = 0, 1, \ldots, c-1$, $j = 1, 2, \ldots, n-c+1$ we build decomposition of state space $S(X) = S_{ij}^{(2)}(X) \cup \bar{S}_{ij}^{(2)}(X)$, $S_{ij}^{(2)}(X) = \{(i,j)\}$. Equating probability flows through the border of subset $S_{ij}^{(2)}(X)$, we obtain the following system of equations:

$$[(n+1-i-j)\lambda + (j-1)\mu + i\nu_{j-1}]\, \pi_{ij-1} =$$

$$= j\mu\pi_{i-1j} + (n+2-i-j)\lambda\pi_{i-1j-1} + (i+1)\nu_{j-1}\pi_{i+1j-1}, \tag{2}$$

$$j = 1, \ldots, n-c+1, \ i = 0, \ldots, c-1,$$

$$\sum_{j=0}^{n-c} \sum_{i=0}^{c} \pi_{ij}(H) = 1. \tag{3}$$

The last equation is normalizing condition. Let us rewrite (2) in the following form:

$$-(n+1-i-j)\lambda\pi_{i-1j} + [(n-i-j)\lambda + j\mu + i\nu_j]\, \pi_{ij} - (i+1)\nu_j\pi_{i+1j} =$$

$$= (j+1)\mu\pi_{i-1j+1}, \ j = 0, ..., n-c-1, \ i = 0, ..., c-2. \tag{4}$$

From equation (1) we find probability π_{cj} and substitute it into (2) for $i = c-1$. We obtain:

$$-(n-c+2)\lambda\pi_{c-2j} + [(n-c+1-j)\lambda + j\mu + (c-1)\nu_j]\, \pi_{c-1j} =$$

$$= \frac{(j+1)c\mu\nu_j}{(n-c-j)\lambda} \sum_{i=0, i\neq c-2}^{c-1} \pi_{ij+1} + \frac{(j+1)\mu\,[(n-c-j)\lambda + c\nu_j]}{(n-c-j)\lambda} \pi_{c-2j+1},$$

$$j = 0, \ldots, n-c-1. \tag{5}$$

We can rewrite the system of equations (4), (5) in the following vector-matrix form:

$$A_j\pi_j = B_j\pi_{j+1},\ j = 0,\dots,n-c-1.$$

From the last equation:

$$\pi_j = \left(\prod_{i=j}^{n-c-1} A_i^{-1}B_i\right)\pi_{n-c},\ j = 0,\dots,n-c-1. \tag{6}$$

Supplementing the system (2) for $i = 0,1,\dots,c-2,\ j = n-c+1$ by the equation $\pi_{0n-c} = \pi_{0n-c}$ and writing it in vector-matrix form we obtain:

$$C\pi_{n-c} = e_0\pi_{0n-c}.$$

Which leads to the following result for the vector π_{n-c}:

$$\pi_{n-c} = C^{-1}e_0\pi_{0n-c}. \tag{7}$$

Finally from equations (6) and (7) we obtain:

$$\pi_j = \Phi_j\pi_{0n-c},\ j = 0,\dots,n-c.$$

By substitution of expression for probability π_{j+1} into equation (1), we find $\pi_{cj},\ j = 0,\dots,n-c-1$. From equation (2) we obtain probability π_{cn-c} when $i = c-1,\ j = n-c+1$.

Probability π_{0n-c} is found from the normalizing condition (3).

The essence of the above result is to provide an efficient algorithm for computing steady state probabilities and the performance characteristics of the service process in the steady regime. The obtained explicit formulas also allow to determinate the structure of dependency of stationary distribution on the model's parameters.

The special case of service rate control, corresponding to the threshold strategy is described below. We fix the threshold $H \in \{-1,0,1,\dots,n-c\}$. If $j \le H$, then $\nu_j = \nu^{(1)}$ and the service process operates in the first mode. If $j > H$, then $\nu_j = \nu^{(2)}$ and the process operates in the second mode. Theorem 1 gives us an efficient algorithm for finding an optimal threshold H^* that maximizes a quality functional. Numeric examples and other details can be found in [6].

3 Controll of Service Rate by Hysteresis Strategy

To define a strategy of the hysteresis type one should fix two thresholds $H_1, H_2 \in \{-1,0,1,\dots,n-c\}$, $H_1 \le H_2$. If the number of retrials $j \le H_1$ then the service rate is equal to $\nu_j^{(1)}$. When $j > H_2$ the service process operates in the second mode with the rate $\nu_j^{(2)}$. When $j \in (H_1, H_2]$, the system keeps its previous operating mode.

Such a delay in switching of service rate allows to lower a cost of this operation and to increase an efficiency of the control. In case of $H_1 = H_2 = H$, $\nu_j^{(1)} = \nu^{(1)}$, $\nu_j^{(2)} = \nu^{(2)}$ the hysteresis strategy turns into the threshold strategy. Single channel systems with retrials and controlled parameters of such type can be found in [7].

To provide a formal description of the service process while preserving the Markov property it is necessary to add the third component $R(t) \in \{1, 2\}$ which is the number of operating mode at time $t \geq 0$. As a result the service process turns into the three-dimensional Markov chain $Y(t) = (C(t), N(t), R(t))$ with a state space $S(Y) = S^{(1)}(Y) \cup S^{(2)}(Y)$, where

$$S^{(1)}(Y) = \{i = (i_1, i_2, 1) : i_1 = 0, \ldots, c, i_2 = 0, \ldots, H_2\},$$

$$S^{(2)}(Y) = \{i = (i_1, i_2, 2) : i_1 = 0, \ldots, c, i_2 = H_1 + 1, \ldots, n - c\},$$

$$S^{(1)}(Y) \cap S^{(2)}(Y) = \emptyset.$$

Local characteristics $\alpha_{ij}, i = (i_1, i_2, i_3), j = (j_1, j_2, j_3)$ of the chain $Y(t)$ remain similar to characteristics of two-dimensional process $X(t)$. We pay additional attention to transitions from layer $S^{(1)}(Y)$ to $S^{(2)}(Y)$ and vice versa:

1. if $i = (c, H_2, 1)$ then
   $\alpha_{ij} = (n - c - H_2)\lambda$, when $j = (c, H_2 + 1, 2)$,
   $\alpha_{ij} = c\nu_{H_2}^{(1)}$, when $j = (c - 1, H_2, 1)$,
   $\alpha_{ij} = -\left[(n - c - H_2)\lambda + c\nu_{H_2}^{(1)}\right]$, when $j = i$,
   otherwise $\alpha_{ij} = 0$.
2. if $[(i_1 = 0, \ldots, c - 1) \wedge (i_2 = H_1 + 1) \wedge (i_3 = 2)]$ then
   $\alpha_{ij} = (n - H_1 - 1 - i_1)\lambda$, when $j = (i_1 + 1, H_1 + 1, 2)$,
   $\alpha_{ij} = (H_1 + 1)\mu$, when $j = (i_1 + 1, H_1, 1)$,
   $\alpha_{ij} = i_1\nu_{H_1+1}^{(2)}$, when $j = (i_1 - 1, H_1 + 1, 2)$,
   $\alpha_{ij} = -\left[(n - H_1 - 1 - i_1)\lambda + (H_1 + 1)\mu + i_1\nu_{H_1+1}^{(2)}\right]$, when $j = i$,
   otherwise $\alpha_{ij} = 0$.

The main result of this section is to construct explicit vector-matrix formulas for steady state probabilities $\pi_{ij}^{(r)}, (i, j, r) \in S(Y)$ of the process $Y(t)$, $t \geq 0$.

Let us consider the matrices:

$$F_j^{(1)} = \left(\prod_{i=j}^{H_1} A_i^{-1}(1)B_i(1)\right)\left[E + \sum_{k=H_1+1}^{H_2}\left(\prod_{i=H_1+1}^{k-1} A_i^{-1}(1)B_i(1)\right)A_k^{-1}(1)D_k(1)\right],$$

$$j = 0, \ldots, H_1$$

$$F_j^{(1)} = \sum_{k=j}^{H_2}\left(\prod_{i=j}^{k-1} A_i^{-1}(1)B_i(1)\right)A_k^{-1}(1)D_k(1), j = H_1 + 1, \ldots, H_2;$$

(It is considered that $F^{(1)}_{H_2+1}$ is equal to zero matrix);

$$F^{(2)}_j = \left(E - \left[\sum_{k=j}^{H_2} \left(\prod_{i=j}^{k-1} A_i^{-1}(2) B_i(2) \right) A_k^{-1}(2) D_k(2) \right] \times \right.$$

$$\left. \times \left[E + \sum_{k=H_1+1}^{H_2} \left(\prod_{i=H_1+1}^{k-1} A_i^{-1}(2) B_i(2) \right) A_k^{-1}(2) D_k(2) \right]^{-1} \right) \times$$

$$\times \prod_{i=j}^{n-c-1} A_i^{-1}(2) B_i(2), j = H_1 + 1, \ldots, n-c-1;$$

$$\left(F^{(2)}_{n-c} = E = \|\delta_{ij}\|^{c-1}_{i,j=0} \right);$$

$$\widetilde{F}^{(1)}_j = \frac{(j+1)\mu}{(n-c-j)\lambda} F^{(1)}_{j+1} F^{(2)}_{H_1+1}, j = 0, \ldots, H_1 - 1,$$

$$\widetilde{F}^{(1)}_j = \frac{(j+1)\mu}{(n-c-j)\lambda} \left(F^{(1)}_{j+1} + \frac{H_1+1}{j+1} E \right) F^{(2)}_{H_1+1}, j = H_1, \ldots, H_2,$$

$$\widetilde{F}^{(2)}_j = \frac{(j+1)\mu}{(n-c-j)\lambda} \left(F^{(2)}_{j+1} - \frac{H_1+1}{j+1} F^{(2)}_{H_1+1} \right), j = H_1 + 1, \ldots, H_2,$$

$$\widetilde{F}^{(2)}_j = \frac{(j+1)\mu}{(n-c-j)\lambda} F^{(2)}_{j+1}, j = H_2 + 1, \ldots, n-c-1,$$

where matrices $A_j(r)$, $B_j(r)$ are defined similar to A_j, B_j; $D_j(r) = \left\| d^j_{ik}(r) \right\|^{c-1}_{i,k=0}$:

$$d^j_{c-1k}(r) = \frac{(H_1+1)\, c\mu \nu^{(r)}_j}{(n-c-j)\lambda}, k = 0, \ldots, c-1;$$

other elements of $D_j(r)$ are equal to 0.

Theorem 2. *Steady state probabilities of* $Y(t) = (C(t), N(t), R(t))$ *can be found in the following form:*

$$\pi^{(1)}_j = \pi^{(2)}_{0n-c} F^{(1)}_j F^{(2)}_{H_1+1} C^{-1} e_0, j = 0, \ldots, H_2,$$

$$\pi^{(2)}_j = \pi^{(2)}_{0n-c} F^{(2)}_j C^{-1} e_0, j = H_1 + 1, \ldots, n-c$$

$$\pi^{(1)}_{cj} = \pi^{(2)}_{0n-c} \overline{1}^T \widetilde{F}^{(1)}_j C^{-1} e_0, j = 0, \ldots, H_2,$$

$$\pi^{(2)}_{cj} = \pi^{(2)}_{0n-c} \overline{1}^T \widetilde{F}^{(2)}_j C^{-1} e_0, j = H_1 + 1, \ldots, n-c-1,$$

$$\pi^{(2)}_{cn-c} = \pi^{(2)}_{0n-c} \frac{\left[\lambda + (n-c)\mu + (c-1)\nu^{(2)}_{n-c} \right] e^T_{c-1} - 2\lambda e^T_{c-2}}{c\nu^{(2)}_{n-c}} C^{-1} e_0,$$

where $\pi_j^{(r)} = \left(\pi_{0j}^{(r)}\ \pi_{1j}^{(r)}\ \ldots\ \pi_{c-1j}^{(r)}\right)^T$, *matrix* C *is defined as above with the substitution of* $A_{n-c}(2)$ *instead of* A_{n-c},

$$\left(\pi_{0n-c}^{(2)}\right)^{-1} = \left(\overline{1}^T \left[E + \sum_{j=0}^{H_2}\left(F_j^{(1)}F_{H_1+1}^{(2)} + \widetilde{F}_j^{(1)}\right) + \sum_{j=H_1+1}^{n-c}\left(F_j^{(2)} + \widetilde{F}_j^{(2)}\right)\right] + \right.$$

$$\left. + \frac{\left[\lambda + (n-c)\mu + (c-1)\nu_{n-c}^{(2)}\right] e_{c-1}^T - 2\lambda e_{c-2}^T}{c\nu_{n-c}^{(2)}}\right) C^{-1} e_0.$$

The proof of theorem 2 is similar to such for theorem 1. We again use the balance equations of probability flows for special subset of the phase space to obtain the set of equations for stationary probabilities wich can be solved in the vector-matrix form.

Results of the theorem 2 are suitable and for solving optimization problems in case of control policy of hysteresis type.

References

1. Ajmone Marsan, M., De Carolis, G., Leonardi, E., Lo Cigno, R., Meo, M.: An approximate model for the computation of blocking probabilities in cellular networks with repeated calls. Telecommunication Systems 15, 53–62 (2000)
2. Artalejo, J.R., Gomez-Corral, A.: Modelling communication systems with phase type service and retrial times. IEEE Communications Letters 11, 955–957 (2007)
3. Falin, G.I., Artalejo, J.R.: A finite source retrial queue. European J. of Oper. Res. 108, 409–424 (1998)
4. Falin, G.I., Templeton, J.G.C.: Retrial Queues. Chapman&Hall (1997)
5. Warland, J.: An introduction to Queueing Networks, New Jersey (1993)
6. Lebedev, E.A., Ponomarov, V.D.: Optimization of retrial queues with finite population, Bulletin of Kiev University, vol. 2, pp. 91–97 (2008)
7. Dudin, A.N., Klimenok, V.I.: Optimization of dynamic control of input load in node of informational - computing network. Automation and Tehnology 3, 25–31 (1991)

Multidimensional Alternative Processes Reliability Models

Vladimir Rykov

Gubkin Russian State University of Oil and Gas, Moscow, Russia
vladimir_rykov@mail.ru

Abstract. Multidimensional alternative processes are introduced, their stationary and quasi-stationary probabilities are investigated, and their applications in reliability models are considered.

Keywords: alternative processes, reliability models, quasi-stationary probabilities.

1 Introduction and Motivation

Most of complex technical systems and biological objects have an hierarchical structure and are supported by inner control system. Therefore from reliability point of view they can be considered as renewable systems. Moreover these systems usually have high reliability that could be, for example, result of quick restoration.

Binary Markov models for renewable reliability systems are enough good studied. Multi-State Markov reliability models are in the focus of specialists during last time [1–4]. Some models of reliability control has been considered in [5, 6].

However elements life and restoration times not always exponentially distributed that leads to engage alternative processes for investigation of binary reliability models with general life and restoration times distributions. From another side because there are no infinitely long existing systems, the most interest represents an investigation of their behavior during their life time. The system behavior before its full failure is described with conditional probability distribution given life time didn't end.

Closed form representation of these probabilities in general case are hardly possible. However, because under quick restoration a system many times visit any of its not absorbing states an interesting problem is study limits of these probabilities for $t \to \infty$. The problems of these limits existence for Markov processes and especially for birth and death processes have been considered by several authors (see for example [7, 8] and the bibliography therein). Evaluation of the convergence rate to the quasi-limiting probabilities for queueing models in [9] has been studied. In [10] generalized birth and death processes as a model for systems degradation has been introduced and studied, where also the problem of quasi-limiting probabilities has been discussed.

In the paper multidimensional alternative processes are introduced and their steady state probabilities are calculated. It is shown that they are insensitive to

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 147–156, 2013.

the shape of process components sojourn time in their states distributions, and have a product form. For Markov case the life time distribution and problem of quasi-stationary distributions existence is considered. Some examples of the model applications in reliability are proposed.

2 The Problem Setting and Examples

Consider some system, consisting from n units, each of which can be in two states: "up" and "down" and suppose that simultaneously only one unit can change its state, being the sojourn time spent of any unit in any its state is a random variable (r.v.) with general distribution. The long time system behavior is usually described with steady state probabilities. However, because there is no infinitely long existing systems in most real situation it is necessary to study their life time (before entering in sone full failure subset of states) as well as their behavior during this time.

Therefore the problem consists in not only investigation of steady state probabilities of a system at infinity, but also in studying their life time i.e. distribution of some absorbing set destination as well as their behavior during this time. Consider some examples.

1. Heterogeneous Reliability System. Renewable reliability system from n heterogeneous elements can be described by n dimensional binary vector $\mathbf{j} = (j_1, \ldots, j_n)$, where

$$j_k = \begin{cases} 0, & \text{if } k\text{-th element is in up state,} \\ 1, & \text{if } k\text{-th element is in down state.} \end{cases}$$

The full set of states $E = \{\mathbf{j} = (j_1, \ldots, j_n),\ j_k = \{0, 1\},\ (k = \overline{1, n})\}$ should be divided into up E_0 and down E_1 subsets. For example the down subset for system in parallel is $E_1 = \{(1, \ldots, 1)\}$ and for system in line is $E_1 = E \setminus \{(1, \ldots, 1)\}$.

2. Homogeneous System. Homogeneous system is a special case of heterogeneous one for which failure and restoration parameters of elements are equal. In this case the model admit states enlarging by joining states with equal number of down elements, $j = \sum_{1 \le k \le n} j_k$ with set of states $E = \{0, 1, \ldots, n\}$.

3. Hierarchical System. Hierarchical system also can be modelled with multidimensional alternative process. In this case elements of the system should be numerated by vector indices $\mathbf{i} = (i_1, i_2, \ldots i_r)$, which components show numbers of sequence subsystems to which appropriate elements belongs, and index r denotes the hierarchical level of appropriate element. Thus, the system states are described with binary vectors $\mathbf{j_i}$, which indices show the elements position, and binary components denote the element state: 0–up state, 1–down state. The system structure function ϕ is a complex function, which should be constructed from structure functions $\phi_{\mathbf{i}_k}$ of subsystems $\mathbf{i}_k = (i_1, i_2, \ldots i_k)$ for $k = 1, 2, \ldots r$ up to elementary level. In this representation the subset $E_1 = \phi^{-1}(\mathbf{0})$ determines the system failure subset.

3 Definition. Main Relations

The above considered systems can be modelled by multidimensional alternative random process $\mathbf{J}(t) = (J_1(t), \dots J_n(t))$ with binary components $J_k(t) = \{0, 1\}$ $(k = \overline{1,n})$ and finite state space $E = \{\mathbf{j} = (j_1, \dots j_n),\ j_k = \{0, 1\}\ (k = \overline{1,n})\}$ consisting of $|E| = N = 2^n$ states. Accordingly to the system structure the transitions from state $\mathbf{j}$ are possible only to "neighboring" states $\mathbf{j} \to \mathbf{j} \pm \mathbf{e}_k$, where $\mathbf{e}_k$ is a vector all component of which are zero except of the k-th one which is equal to one, being the time spent by each component in its 0 and 1 states during any visit in it has general distribution.

In order to describe the system behavior with Markov process let us use the extended state space $\mathcal{E} = E \times R_+^n$ and consider multidimensional process $Z(t) = \{\mathbf{J}(t), \mathbf{X}(t)\}$ in which components $\mathbf{J}(t) \in E$ are the states of the initial process, and additional components $\mathbf{X}(t) \in R_+^n$ denote the time spent each of its first component in its state beginning from its last entering in it. Denote by

- $A_k(x),\ B_k(x)$ cumulative distribution functions (c.d.f.) of the sojourn time k-th component of process $\mathbf{J}(t)$ in its state $j_k = 0$ or $j_k = 1$ beginning from the time of last entering in it;
- $a_k(x),\ b_k(x)$ appropriate probability density function (p.d.f.);
- $a_k = \int (1 - A_k(x))dx,\ b_k = \int (1 - B_k(x))dx$ their expectations; and
- $\alpha_k(x) = (1 - A_k(x))^{-1} a_k(x) \quad \beta_k(x) = (1 - B_k(x)^{-1}) b_k(x)$ transition intensities of the process $\mathbf{J}(t)$ from the state $\mathbf{j}$ to states $\mathbf{j} + e_k$ and $\mathbf{j} - e_k$ correspondingly under condition that the time spent of its k-th component beginning from the last entering in it equal to x.

Put also

$$\bar{\mathbf{j}}_k = (j_1, \dots j_{k-1}, 1 - j_k, j_{k+1}, \dots j_n), \quad \mathbf{x}_k(u) = (x_1, \dots x_{k-1}, u, x_{k+1}, \dots x_n),$$

$$c_{(k,j_k)}(x) = a_k^{1-j_k}(x)\, b_k^{j_k}(x),$$

$$\gamma_{(k,j_k)}(x) = \alpha_k^{1-j_k}(x) \beta_k^{j_k}(x),\ \gamma_{\mathbf{j}}(x) = \sum_{1 \le k \le n} \gamma_{(k,j_k)}(x).$$

Denote by $\pi_{\mathbf{j}}(t; \mathbf{x}) = \pi_{(j_1,\dots j_n)}(t;\, x_1, \dots x_n)$ the p.d.f. of the process $Z(t)$,

$$\pi_{\mathbf{j}}(t;\, \mathbf{x})\mathbf{dx} = \pi_{(j_1,\dots j_n)}(t;\, x_1, \dots x_n) dx_1 \dots dx_n =$$

$$= \mathbf{P}\{J_i(t) = j_i,\ X_i(t) \in dx_i,\ i = \overline{1,n}\},$$

Kolmogorov's system of equations for these p.d.f.'s in the set $0 \le x_j \le t < \infty,\ j = 1, 2, \dots n$ is

$$\frac{\partial \pi_{\mathbf{j}}(t, \mathbf{x})}{\partial t} + \sum_{1 \le k \le n} \frac{\partial \pi_{\mathbf{j}}(t, \mathbf{x})}{\partial x_k} = -\gamma_{\mathbf{j}}(x) \pi_{\mathbf{j}}(t, \mathbf{x}) \quad (\mathbf{j} \in E), \tag{1}$$

and boundary conditions are

$$\pi_{\mathbf{j}}(t; \mathbf{x}_k(0)) = \int_0^t \pi_{\bar{\mathbf{j}}_k}(t; \mathbf{x}_k(u)) \gamma_{(k,\, j_k)}(\mathbf{j};\, u) du \quad (\mathbf{j} \in E), \tag{2}$$

while the initial conditions in terms of Dirac δ-function are

$$\pi_{\mathbf{j}}(0;\mathbf{0}) = \delta_{(\mathbf{j},\,\mathbf{0})}(t) \quad (\mathbf{j} \in E). \tag{3}$$

Based on characteristic method for first order partial differential equations solution [11], general solution of this system should be find in the form

$$\pi_{\mathbf{j}}(t;\mathbf{x}) = h_{\mathbf{j}}(t - x_1, \ldots, t - x_n) \prod_{1 \le k \le n} (1 - A_k(x_k))^{1-j_k} (1 - B_k(x_k))^{j_k}, \tag{4}$$

where functions $h_{\mathbf{j}}(\ldots)$ accordingly to the boundary conditions (2) should be find from equations

$$h_{\mathbf{j}}(\mathbf{t} - \mathbf{x}_k(0)) = \int_0^t h_{\bar{\mathbf{j}}_k}(\mathbf{t} - \mathbf{x}_k(u)) c_{(k,\,j_k)}(\mathbf{j};\, u) du \quad (\mathbf{j} \in E). \tag{5}$$

One can see that these equations hold for the functions $h_{\mathbf{j}}(\ldots)$ in the form

$$h_{\mathbf{j}}(u_1, \ldots, u_n) = \prod_{1 \le k \le n} h_{(k,\,j_k)}(u_k). \tag{6}$$

From the other side the initial conditions (3) show that the functions $h_{(k,\,j_k)}(u)$ have to satisfy to the equations

$$h_{(k,\,j_k)}(t) = \delta_{j_k,\,0}(t) + \int_0^t h_{k,\,1-j_k}(t-u) c_{(k,j_k)}(u)\, du \quad (k = \overline{1,\,n}),$$

which in terms of Laplace Transforms (LT) can be represented as

$$\tilde{h}_{(k,\,j_k)}(s) = \delta_{j_k,\,0} + \tilde{h}_{k,\,1-j_k}(s) \tilde{c}_{(k,\,j_k)}(s) \quad (k = \overline{1,\,n}).$$

The last system has the following solution

$$\tilde{h}_{(k,\,j_k)}(s) = \frac{\tilde{a}_k^{j_k}(s)}{1 - \tilde{a}_k(s)\tilde{b}_k(s)}. \tag{7}$$

In general case closed form representation of time dependent probabilities is hardly possible. However for Markov case one can get them.

Example. Consider multidimensional Markov alternative process with transition intensities α_k and β_k for k-th element. In this case $\tilde{a}_k(s) = \alpha_k(s + \alpha_k)^{-1}$, $\tilde{b}_k(s) = \beta_k(s + \beta_k)^{-1}$. Thus for functions $\tilde{h}_{(k,\,j_k)}(s)$ from (7) one find

$$\tilde{h}_{(k,\,j_k)}(s) = \frac{(s + \alpha_k)^{1-j_k} \alpha_k^{j_k} (s + \beta_k)}{s(s + \alpha_k + \beta_k)} \tag{8}$$

At least using (4) and (6) one find $\tilde{\pi}_{\mathbf{j}}(s) = \prod_{1\le k\le n} \tilde{\pi}_{k,\,j_k}(s)$ with

$$\tilde{\pi}_{k,\,j_k}(s) = \frac{\tilde{h}_{(k,\,j_k)}(s)}{(s+\alpha_k)^{1-j_k}\left((s+\beta_k)^{j_k}\right)} =$$

$$= \frac{\alpha_k^{1-j_k}\beta_k^{j_k}}{s(\alpha_k+\beta_k)} + (-1)^{j_k}\frac{\alpha_k}{(\alpha_k+\beta_k)(s+\alpha_k+\beta_k)},$$

Inversion of these expressions give

$$\pi_{k,\,j_k}(t) = \frac{\alpha_k^{1-j_k}\beta_k^{j_k}}{\alpha_k+\beta_k} + (-1)^{j_k}\frac{\alpha_k}{\alpha_k+\beta_k}e^{-(\alpha_k+\beta_k)t},$$

and $\pi_{\mathbf{j}}(t) = \prod_{1\le k\le n} \pi_{k,\,j_k}(t)$ that also can be find with direct Markov approach.

4 Stationary Regime

The expression (7) shows that the functions $h_{(k,\,j_k)}(t)$ are renewal densities of a renewal process, generated by process $Z(t)$ components returning times to the zero state and therefore for $t \to \infty$ they have the limits

$$\lim_{t\to\infty} h_{(k,\,j_k)}(t) = \lim_{s\to 0} s\tilde{h}_{(k,\,j_k)}(s) = \frac{1}{a_k+b_k}.$$

It is impossible to find closed form expressions for time-dependent probabilities in general case, however taking into account (6) for $t \to \infty$ in (4) one find

$$\pi_{\mathbf{j}}(\mathbf{x}) = \lim_{t\to\infty} \pi_{\mathbf{j}}(t;\mathbf{x}) = \prod_{1\le k\le n} \frac{1}{a_k+b_k}(1-A_k(x_k))^{1-j_k}(1-B_k(x_k))^{j_k}.$$

From another side for stationary probabilities of macro-states using Smith's key theorem from (4) and (6) one has

$$\pi_{\mathbf{j}} = \lim_{t\to\infty} \int\limits_{0\le x_k\le t,\,(k=\overline{1,n})} \cdots \int \pi_{\mathbf{j}}(t;x_1,\ldots,x_n)dx_1,\ldots,dx_n =$$

$$= \lim_{t\to\infty} \prod_{1\le k\le n} \int_0^t h_{(k,\,j_k)}(t-x_k)(1-A_k(x_k))^{1-j_k}(1-B_k(x_k))^{j_k}dx_k =$$

$$= \prod_{1\le k\le n} \frac{1}{a_k+b_k} \int_0^\infty (1-A_k(x_k))^{1-j_k}(1-B_k(x_k))^{j_k}dx_k = \prod_{1\le k\le n} \frac{a_k^{1-j_k}b_k^{j_k}}{a_k+b_k}.$$

This result means insensitivity of stationary probabilities to the shape of c.d.f. sojourn time.

5 Life Time Distribution and System Behavior on It

For many applications especially for reliability models actual problems are both: to find the failure subset of states E_1 destination time, which can be considered as the system "life time", and to investigate the system behavior during this time. Denote by

$$T = \inf\{t : \mathbf{J}(t) \in E_1\}$$

the subset E_1 destination time and by $F(t) = \mathbf{P}\{T \le t\}$ its c.p.f. In order to find this distribution and the system state distribution during this time one should solve the system (1) with subset E_1 as an absorbing set. In this case the Kolmogorov's system of equations (1) holds for $\mathbf{j} \in E_0$ with the same boundary (2) and initial (3) conditions and absorbing conditions

$$\pi_{\mathbf{j}}(t) = \sum_{k:\,\bar{\mathbf{j}}_k \in E_0} \int_0^t \pi_{\bar{\mathbf{j}}_k}(t; \mathbf{x}_k(u))\gamma_{(k,\,j_k)}(u)du \quad (\mathbf{j} \in E_1). \tag{9}$$

The general solution of this system does not change in the set $\mathbf{j} \in E_0$, i.e. has a form

$$\pi_{\mathbf{j}}(t; \mathbf{x}) = h_{\mathbf{j}}(t - x_1, \ldots, t - x_n) \prod_{1 \le k \le n} (1 - A_k(x_k))^{1-j_k}(1 - B_k(x_k))^{j_k}, \tag{10}$$

At that functions $h_{\mathbf{j}}(\cdots)$ must satisfy to equations that appropriate to the boundary (2), initial (3) and absorbing (9) conditions, being the last one takes the form

$$h_{\mathbf{j}}(t) = \sum_{k:\,\bar{\mathbf{j}}_k \in E_0} \int_0^t h_{\bar{\mathbf{j}}_k}(\mathbf{t} - \mathbf{x}_k(u))\, c_{(k,\,j_k)}du \quad (\mathbf{j} \in E_1) \tag{11}$$

Now the components of the process $Z(t)$ are dependent and there is no simple closed form solution for this system. However, the equations (1) for $\mathbf{j} \in E_0$ give the possibility for numerical solution of the problem and calculation of the subset E_1 destination time distribution in the form

$$F(t) = \mathbf{P}\{T \le t\} = \sum_{\mathbf{j} \in E_1} \pi_{\mathbf{j}}(t). \tag{12}$$

In case of Markov alternative process it is possible to calculate appropriate functions in terms of their LT. Denote by Λ infinitesimal matrix of the process $\mathbf{J}(t)$ and by $\boldsymbol{\pi}'(t) = (\pi_{\mathbf{j}}(t), \mathbf{j} \in E)$ its probability states vector with $\pi_{\mathbf{j}}(t) = \mathbf{P}\{\mathbf{J}(t) = \mathbf{j}\}$, and by $\boldsymbol{e}'_{\mathbf{0}}$ vector each components of which equals 0, except of those that corresponds to the state $\mathbf{0}$, which equals to 1. Here and below the sign "prime" denotes transpose operation while derivatives are denoted with up dots. With this notations the Kolmogorov's system of equations in matrix form with initial conditions are

$$\dot{\boldsymbol{\pi}}'(t) = \boldsymbol{\pi}'(t)\Lambda, \quad \boldsymbol{\pi}'(0) = \boldsymbol{e}'_{\mathbf{0}} \tag{13}$$

that in terms of LT is

$$s\tilde{\boldsymbol{\pi}}' - \boldsymbol{e}'_{\mathbf{0}} = \tilde{\boldsymbol{\pi}}'(s)\Lambda. \tag{14}$$

Representing the infinitesimal matrix and probability state vector $\boldsymbol{\pi}'(t)$ in block form

$$\Lambda = \begin{bmatrix} \Lambda_{0,0} & \Lambda_{0,1} \\ \Lambda_{1,0} & \Lambda_{1,1} \end{bmatrix}, \quad \boldsymbol{\pi}'(t) = (\boldsymbol{\pi}'_{E_0}(t), \boldsymbol{\pi}'_{E_1}(t)), \quad \boldsymbol{e}'_{\mathbf{0}} = (\boldsymbol{e}'_{E_0}, \boldsymbol{e}'_{E_1}),$$

where matrix blocks with indices 0 and 1 correspond to system states subsets E_0 and E_1 and putting $\Lambda_{1,0} = \boldsymbol{e}'_{E_1} = 0$, reduce the system (14) to the form

$$s\tilde{\boldsymbol{\pi}}'_{E_0}(s) - \boldsymbol{e}'_{E_0} = \tilde{\boldsymbol{\pi}}'_{E_0}(s)\Lambda_{0,0}, \qquad s\tilde{\boldsymbol{\pi}}'_{E_1}(s) - \boldsymbol{e}'_{E_1} = \tilde{\boldsymbol{\pi}}'_{E_0}(s)\Lambda_{0,1}$$

that have the solution

$$\tilde{\boldsymbol{\pi}}'_{E_0}(s) = \boldsymbol{e}'_{E_0}(Is - \Lambda_{0,0})^{-1}, \quad \tilde{\boldsymbol{\pi}}'_{E_1}(s) = \frac{1}{s}\boldsymbol{e}'_{E_0}(Is - \Lambda_{0,0})^{-1}\Lambda_{0,1} \tag{15}$$

Taking into account that the life time c.p.f. has a form (12), its generating function (g.f.) $\tilde{f}(s) = s\tilde{F}(s)$ can be represented as

$$\tilde{f}(s) = s\tilde{\boldsymbol{\pi}}'_{E_1}(s)\mathbf{1} = \boldsymbol{e}'_{E_0}(Is - \Lambda_{0,0})^{-1}\Lambda_{0,1}\mathbf{1}. \tag{16}$$

This expression allows to calculate the life time moments and moreover being fractionally rational it admits by calculation of its inversion to find life time p.d.f. and reliability function. Some example of this approach see in [12].

6 Quasi-stationary Probabilities

The main characteristic of an object behavior at its life cycle is the conditional state probabilities given life time is not finished. Denote by $\bar{\pi}_{\mathbf{j}}(t)$ conditional probability of the object occurring in the state $\mathbf{j}$ at its life cycle,

$$\bar{\pi}_{\mathbf{j}}(t) = \mathbf{P}\{\mathbf{J}(t) = \mathbf{j} \mid t < T\}.$$

Because the system (13) solution $\{\pi_{\mathbf{j}}(t),\ \mathbf{j} \in E_0\}$ with absorbing state E_1 is a joint probability the process being at the state $\mathbf{j}$ jointly with life cycle isn't finished, $\pi_{\mathbf{j}}(t) = \mathbf{P}\{\mathbf{J}(t) = \mathbf{j},\ t < T\}$, so for $\bar{\pi}_{\mathbf{j}}(t)$ it is true the representation

$$\bar{\pi}_{\mathbf{j}}(t) \equiv \mathbf{P}\{\mathbf{J}(t) = \mathbf{j} \mid t < T\} = \frac{\pi_{\mathbf{j}}(t)}{R(t)},$$

where $R(t) = 1 - F(t) = \mathbf{P}\{T > t\}$ is a reliability (survival) function of the system. Calculation of these function in general case is hardly possible. However because under quick restoration the object many times occurs in its non-absorbing states, it is interesting to prove the existence and to calculate the limiting value of these conditional probabilities for $t \to \infty$, that could be considered as part

of time spending by the process in each of its states at its fife cycle (before absorbing).

Consider these problems for the case of Markov alternative process. Asymptotic behavior of probabilities $\bar{\pi}_{\mathbf{j}}(t)$ for $t \to \infty$ will be studied with the help of their LT, that accordingly to (15) have the form

$$\tilde{\pi}_{\mathbf{j}}(s) = \boldsymbol{e}'_{E_0}(Is - \Lambda_{0,0})^{-1}\mathbf{1} = \frac{\Delta_{\mathbf{j}}(s)}{\Delta(s)}, \qquad \mathbf{j} \in E_0$$

$$\tilde{\pi}_{\mathbf{j}}(s) = \frac{1}{s\Delta(s)} \sum_{\mathbf{i} \in E_0} \Delta_{\mathbf{i}}(s)\lambda_{\mathbf{i},\mathbf{j}} \qquad \mathbf{j} \in E_1,$$

where $\Delta(s)$ is determinant of matrix $(Is - \Lambda_{0,0})$ and $\Delta_{\mathbf{j}}(s)$ is algebraic adjunct of $\mathbf{j}$-th component its first row.

It is very known that the all roots of characteristic equation

$$\Delta(s) = \det(Is - \Lambda_{0,0}) = 0 \tag{17}$$

are negative. Denote them with s_i $(i = \overline{1,N})$, numerate them in order its decreasing $s_N < \cdots < s_2 < s_1$ and suppose that the maximal root s_1 has an order one. Thus in simple fractions representation of the functions $\tilde{\pi}_{\mathbf{j}}(s)$ for $\mathbf{j} \in E_0$ has a form

$$\tilde{\pi}_{\mathbf{j}}(s) = \frac{A_{\mathbf{j}1}}{s - s_1} + \sum_{2 \le k \le n} \frac{A_{\mathbf{j},k}}{(s - s_k)^{i_{\mathbf{j},k}}} \tag{18}$$

where $i_{\mathbf{j},k}$ is an appropriate root multiply, and $A_{\mathbf{j},k}$ are some coefficients, first of which is a residue of the function $\tilde{\pi}_{\mathbf{j}}(s)$ at the point s_1,

$$A_{\mathbf{j}1} = \lim_{s \to s_1} (s - s_1)\tilde{\pi}_{\mathbf{j}}(s) = \frac{\Delta_{\mathbf{j}}(s_1)}{\dot{\Delta}(s_1)}.$$

Inversion of (18) gives

$$\pi_{\mathbf{j}}(t) = A_{\mathbf{j}1}e^{s_1 t}(1 + f_{\mathbf{j}}(t)), \tag{19}$$

where

$$f_{\mathbf{j}}(t) = \sum_{2 \le k \le n} \frac{A_{\mathbf{j}k}}{A_{\mathbf{j}1}} e^{-(s_1 - s_k)t} \to 0 \quad \text{for} \quad t \to \infty.$$

Show that the reliability function $R(t)$ has the same asymptotic behavior at infinity as probabilities $\pi_{\mathbf{j}}(t)$. Really, from (12) and (15) it follows that $\tilde{F}(s)$ can be represented as

$$\tilde{F}(s) = \frac{A_{E_1,0}}{s} + \frac{A_{E_1,1}}{s - s_1} + \sum_{2 \le k \le n} \frac{A_{E_1,k}}{(s - s_k)^{i_{E_1,k}}}$$

with some coefficients $A_{E_1,k}$. Show that the first coefficient $A_{E_1,0}$ in this representation equals 1, $A_{E_1,0} = 1$. Indeed, $F(t)$ is the non degenerated c.p.f. and therefore $\lim_{t \to \infty} F(t) = 1$, which leads to

$$A_{E_1,0} = \lim_{s \to 0} s\tilde{F}(s) = 1.$$

Now, taking into account that $R(t) = 1 - F(t)$, for its LT $\tilde{R}(s)$ holds

$$\tilde{R}(s) = \frac{1}{s} - \tilde{F}(s) = \frac{A_{E_1,1}}{s - s_1} + \sum_{2 \le k \le n} \frac{A_{E_1,k}}{(s - s_k)^{i_{E_1,k}}}.$$

Passing to original in the last relation one find

$$R(t) = A_{E_1,1} e^{s_1 t}(1 + f_{E_1}(t)), \tag{20}$$

with

$$f_{E_1}(t) = \sum_{2 \le k \le n} \frac{A_{E_1,k}}{A_{E_1,1}} e^{-(s_1 - s_k)t} \to 0 \quad \text{for} \quad t \to \infty.$$

From (19) and (20) it follows that for $t \to \infty$ holds

$$\bar{\pi}_{\mathbf{j}} = \lim_{t \to \infty} \bar{\pi}_{\mathbf{j}}(t) = \lim_{t \to \infty} \frac{\pi_{\mathbf{j}}(t)}{R(t)} = \lim_{t \to \infty} \frac{A_{\mathbf{j},1}(1 + f_{\mathbf{j}}(t))}{A_{E_1,1}(1 + f_{E_1}(t))} =$$
$$= \frac{A_{\mathbf{j},1}}{A_{E_1,1}} = \lim_{s \to s_1} \frac{\tilde{\pi}_{\mathbf{j}}(s)}{\tilde{R}(s)} = \frac{\tilde{\pi}_{\mathbf{j}}(s_1)}{\tilde{R}(s_1)}.$$

The above argumentations are represented in the theorem

Theorem 1. *Asymptotic behavior of the functions $\pi_{\mathbf{j}}(t)$ and $R(t)$ for $t \to \infty$ coincide and it is determined by the maximal root of characteristic equation $\Delta(s) = 0$. Thus, the conditional probabilities $\bar{\pi}_{\mathbf{j}}(t)$ have limits for $t \to \infty$, which equal*

$$\bar{\pi}_{\mathbf{j}} = \lim_{t \to \infty} \bar{\pi}_{\mathbf{j}}(t) = \frac{\tilde{\pi}_{\mathbf{j}}(s_1)}{\tilde{R}(s_1)}. \tag{21}$$

Denote by

$$\bar{p}_{\mathbf{ij}}(t) = \mathbf{P}\{\mathbf{J}(t) = \mathbf{j},\ t < T \,|\, \mathbf{J}(0) = \mathbf{i}\} \equiv \mathbf{P}_{\mathbf{i}}\{\mathbf{J}(t) = \mathbf{j},\ t < T\} \quad (\mathbf{i}, \mathbf{j} \in E_0)$$

transition probabilities of the process $\mathbf{J}(t)$ reduced to the subset E_0. Due to Markov property of the process $\mathbf{J}(t)$ the following equality holds

$$\bar{\pi}_{\mathbf{j}}(s + t) \equiv \bar{p}_{\mathbf{0},\mathbf{j}}(s + t) = \sum_{\mathbf{k} \in E_0} \bar{p}_{\mathbf{0},\mathbf{k}}(s) \bar{p}_{\mathbf{k},\mathbf{j}}(t).$$

Passing to the limit for $s \to \infty$ in the last equality one get that

$$\bar{\pi}_{\mathbf{j}} = \sum_{\mathbf{k} \in E_0} \bar{\pi}_{\mathbf{k}} \bar{p}_{\mathbf{k},\mathbf{j}}(t) \quad \text{for all } t. \tag{22}$$

This result can be represented as the following theorem

Theorem 2. *Quasi-stationary probabilities of the process $\mathbf{J}(t)$ exist, coincide with quasi-limiting ones, and satisfy to the equation (22).*

For the general case analogous theorems also hold. However its proof is more delicate and we'll leave it for the next time.

7 Conclusion

Multidimensional alternative processes are considered. It is shown that their steady state probabilities have a product form and unsensitive to the shape of its states sojourn time distributions. Distribution for some absorbing subset of states destination time is studied and existence of quasi-limiting probabilities is shown.

References

1. Rykov, V., Dimitrov, B.: On Multi-State Reliability Systems. Applied Stochastic Models and Information Processes. In: Proceedings of the International Seminar, Petrozavodsk, September 8-13, pp. 128–135 (2002), http://www.jip.ru/2002-2-2-2002.htm
2. Rykov, V., Dimitrov, B., Green Jr., D., Snanchev, P.: Reliability of complex hierarchical systems with fault tolerance units. In: Proceedings MMR 2004, Santa Fe (U.S.A.) (2004) (printed in CD)
3. Dimitrov, B., Green Jr., D., Rykov, V., Stanchev, P.: Reliability Model for Biological Objects. Longevity, Aging and Degradation Models. In: Antonov, V., Huber, C., Nikulin, M., Polischook, V. (eds.) Transactions of the First Russian-French Conference (LAD 2004), Saint Petersburg State Politechnical University, SPB, June 7-9, vol. 2, pp. 230–240 (2004)
4. Lisniansky, A., Levitin, G.: Multi-State System Reliability. Assessment, Optimization and Application, p. 358. World Scientific, New Jersey
5. Rykov, V., Efrosinin, D.: Reliability Control of Fault Tolerance Units. In: Proceedings MMR 2004, Santa Fe (U.S.A.) (2004) (published in CD)
6. Rykov, V., Efrosinin, D.: Reliability Control of Biological Systems with failures. Longevity, Aging and Degradation Models. In: Antonov, V., Huber, C., Nikulin, M., Polischook, V. (eds.) Transactions of the First Russian-French Conference (LAD 2004), Saint Petersburg State Politechnical University, SPB, vol. 2, pp. 241–255 (2004)
7. van Doorn, E.A.: Quasi-stationary distributions and convergence to quasi-stationarity of birth-death processes. Adv. Appl. Probab. 26, 683–700 (1991)
8. Kijima, M., Nair, M.G., Pollet, P.K., van Doorn, E.A.: Limiting conditional distributions for birth-death processes. Adv. Appl. Probab. 29, 185–204 (1997)
9. Granovsky, B.L., Zeifman, A.: Nonstationary queues: estimation of the rate of convergence. Queueing Systems 46, 363–388 (2004)
10. Rykov, V.: Generalized birth and death processes as degradation models. In: Vonta, F. (ed.) Proceedings of the International Conference Statistical Methods for Biomedical and Technical Systems, Limassol, Cyprus, Univ of Cyprus, Nicosia (2006)
11. Petrovsky, I.G.: Lections on theory usual differential equations. M.-L.: GITTL. p. 232 (1952) (in Russian)
12. Vishnevsky, V.M., Kozyrev, D.V., Rykov, V.V.: On reliability of hibrid system multimedia information thansmission. In: Proceedings BWWT 2013, Minsk (2013)

$BMAP/G/1$ Cyclic Polling Model with Binomial Disciplines

Zsolt Saffer

Department of Telecommunications
Budapest University of Technology and Economics (BUTE),
Hungary
safferzs@hit.bme.hu

Abstract. The paper deals with the analysis of $BMAP/G/1$ cyclic polling model with binomial-gated and binomial-exhaustive disciplines. The analysis relies on formerly applied methodology, in which the service discipline independent and service discipline dependent parts of the analysis are treated separately. In this work we complete the service discipline dependent part of the analysis for the binomial disciplines. This leads to a governing equation of the system in terms of the steady-state number of customers at the server arrival and departure epochs. A numerical procedure can be established based on the newly derived results together with formerly obtained service discipline independent results to determine the steady-state factorial moments of the number of customers in the system.

Keywords: queueing theory, polling model, BMAP, service discipline.

1 Introduction

This paper deals with the analysis of cyclic polling models with Batch Markovian Arrival Process (BMAP), which is the generalization of the classical cyclic polling model. The end of service at the given station is governed by the so-called service discipline, like e.g. gated, exhaustive, binomial-gated or binomial-exhaustive. For the analysis of classical cyclic polling systems see Takagi [7]. Polling models are effective instruments in modeling computer systems, manufacturing systems and telecommunication systems see e.g. in Takagi [8].

BMAP introduced by Lucantoni [3] is the natural generalization of the batch Poisson arrival process. The analysis method of BMAP queueing models with more stations, like priority models or polling models, can utilize the advantages of both the matrix analytic-method by Neuts [4] and factorization forms in probability-generating function (PGF) domain.

In this paper we apply the same analysis method for polling models with BMAP as in our previous work [6], in which the analysis is separated into two parts based on quantities at server arrival and departure epochs. In the service discipline independent part factorization forms are established in terms of

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 157–166, 2013.

quantities at server arrival and departure epochs, while in the service discipline dependent part these quantities are solved for the individual disciplines.

In *binomial-gated* discipline (Levy [2]) every customer present at the polling epoch is served with probability p. Under *binomial-exhaustive* discipline (Boxma [1]) every customer present at the polling epoch and arrived during its associated busy period is served with probability p.

In this work first we give an overview of the service discipline independent results including the steady-state vector factorial moments for the number of customers and then provide the discipline specific analysis for the binomial-gated and binomial-exhaustive disciplines.

The contribution of this paper is the service discipline specific analysis part of the $BMAP/G/1$ cyclic nonzero-switchover-times polling model with binomial-gated and binomial-exhaustive disciplines. We set up the governing equations of the system in terms of joint PGFs of the steady-state number of customers and the phases of the $BMAP$s at server arrival and departure epochs. They can be numerically solved by means of system of linear equations and afterwards the required quantities at the server arrival and departure epochs are computed.

The rest of this paper is organized as follows. In section 2 we introduce the model and the notations. In section 3 we give an overview of the former service discipline independent results, which we relay on. The analysis of the nonzero-switchover-times polling model with binomial-gated and binomial-exhaustive disciplines follows in section 4. Final remarks closes the paper in Section 5.

2 Model and Notation

2.1 *BMAP* Process

In this Section we give a brief summary on the BMAP related definitions and notations, which we use in this paper. For a more detailed description on BMAP see Lucantoni [3].

The number of BMAP phases is denoted by L. The $L \times L$ matrix $\mathbf{D}_0$ governs the transitions without any arrival. Similarly the $L \times L$ matrix $\mathbf{D}_k (k \geq 1)$ governs the transitions with batch arrivals, in which the batch size is k. The irreducible infinitesimal generator of the phase process is $\mathbf{D} = \sum_{k=0}^{\infty} \mathbf{D}_k$. Let $\boldsymbol{\pi}$ be the stationary probability vector of the phase process. Then $\boldsymbol{\pi}\mathbf{D} = \mathbf{0}$ and $\boldsymbol{\pi}\mathbf{e} = 1$ uniquely determine $\boldsymbol{\pi}$, where $\mathbf{e}$ is the column vector having all elements equal to one. The matrix generating function (matrix GF) of $\mathbf{D}_k$, $\widehat{\mathbf{D}}(z)$ is defined as

$$\widehat{\mathbf{D}}(z) = \sum_{k=0}^{\infty} \mathbf{D}_k z^k, \quad |z| \leq 1.$$

The stationary arrival rate of a BMAP is $\lambda = \boldsymbol{\pi} \left. \frac{d\widehat{\mathbf{D}}(z)}{dz} \right|_{z=1} \mathbf{e} = \boldsymbol{\pi} \sum_{k=0}^{\infty} k\mathbf{D}_k \mathbf{e}$.

So far we used the conventional BMAP notations, i.e. the station index i was suppressed throughout this Section. However from now on the first subscript stands for the station index, thus $\mathbf{D}_i$ denotes matrix $\mathbf{D}$ of the BMAP at station i.

2.2 The BMAP/G/1 Cyclic Polling Model

We consider a continuous-time asymmetric polling model with N stations. A single server attends the stations in a cyclic manner. Each station has an infinite buffer queue, which is served when the server attends that station. If no customer is present at a station at server arrival, the server immediately attends the next station. At each station batch of customers arrive according to BMAP process. We call the BMAP at station i as i-th BMAP and λ_i denotes its stationary arrival rate. The customer who arrives to station i is called i-customer. The customer service times at station i are general independent and identically distributed. B_i stands for the customer service time at station i and $\widetilde{B}_i(s)$, $B_i(t)$ and b_i denote its Laplace-Stieljes transform (LST), its cumulated distribution function and its mean, respectively. The model enables only nonzero-switchover-times. R_i denotes the switchover time from station i to the next one. The R_i switchover times of the consecutive cycles are general independent and identically distributed. $\widetilde{R}_i(s)$, $R_i(t)$ and r_i are its LST, cumulated distribution function and its mean, respectively.

The arrival of the server to a station and the departure of the server from a station are called *polling epoch* and *departure epoch*, respectively. We call the polling epoch of station i as i-polling epoch. Similarly the departure epoch of station i is an i-departure epoch. The *cycle time* of a given station is defined as the time elapsed from the server visit to station i in the actual cycle to the server visit to the same station in the next cycle. It is also called as polling cycle. The mean cycle time is denoted by c.

On the BMAP/G/1 cyclic polling model we impose the following assumptions:

A.1 At each station the phase process of the BMAP is irreducible and the stationary arrival rate is positive and finite, $0 < \lambda_i < \infty$.

A.2 The mean customer service time and the mean switchover time are positive and finite at each station, $0 < b_i < \infty$, $0 < r_i < \infty$.

A.3 The arrival processes, the service times and the switchover times are mutually independent.

The server utilization at station i and the overall utilization are $\rho_i = \lambda_i b_i$ and $\rho = \sum_{i=1}^{N} \rho_i$, respectively. We assume that all stations of the polling system are stable.

We define matrix $\mathbf{A}_i(k)$, whose (j,l)-th element, for $k \geq 0$, $1 \leq j,l \leq L$, is given as the conditional probability that during an i-customer service time the number of i-th BMAP arrivals is k and the final phase of the i-th $BMAP$ is l given that the initial phase of the i-th BMAP is j. Matrix GF $\widehat{\mathbf{A}}_i(z)$ is defined as $\widehat{\mathbf{A}}_i(z) = \sum_{k=0}^{\infty} \mathbf{A}_i(k) z^k, \quad |z| \leq 1$.

$adj\mathbf{Y}$ and $det\mathbf{Y}$ denote the adjugate and the determinant of matrix $\mathbf{Y}$, respectively. Furthermore $[\mathbf{Y}]_{j,l}$ stands for the j,l-th element of matrix $\mathbf{Y}$ and similarly $[\mathbf{y}]_j$ denotes the j-th element of vector $\mathbf{y}$. When $\widehat{\mathbf{Y}}(z)$, $|z| \leq 1$ is a matrix GF,

$\mathbf{Y}^{(k)}$ denotes its k-th ($k \geq 1$) factorial moment, i.e., $\mathbf{Y}^{(k)} = \frac{d^k}{dz^k}\widehat{\mathbf{Y}}(z)|_{z=1}$ and $\mathbf{Y}^{(0)}$ denote its value at $z = 1$, i.e., $\mathbf{Y}^{(0)} = \widehat{\mathbf{Y}}(1)$. The same notation convention is applied for any quantities having the form $\widehat{\mathbf{y}}(z)$, $det\widehat{\mathbf{T}}_i\,(z)$ and $adj\widehat{\mathbf{T}}_i\,(z)$ for $|z| \leq 1$. Hence, for $k \geq 1$, $\mathbf{y}^{(k)} = \frac{d^k}{dz^k}\widehat{\mathbf{y}}(z)|_{z=1}$, $\mathbf{y}^{(0)} = \widehat{\mathbf{y}}(1)$, $[det\mathbf{T}_i]^{(k)} = \left.\frac{d^k\left(det\widehat{\mathbf{T}}_i(z)\right)}{dz^k}\right|_{z=1}$, $[det\mathbf{T}_i]^{(0)} = det\widehat{\mathbf{T}}_i\,(1)$, $[adj\mathbf{T}_i]^{(k)} = \left.\frac{d^k\left(adj\widehat{\mathbf{T}}_i(z)\right)}{dz^k}\right|_{z=1}$ and $[adj\mathbf{T}_i]^{(0)} = adj\widehat{\mathbf{T}}_i\,(1)$.

3 Service Discipline Independent Results

In this section we recall the former service discipline independent results which we rely on in the analysis.

Let $N_i(t)$ and $J_i(t)$ be the right continuous number of i-customers in the system at time t, for $t \geq 0$, and the phase of the i-th $BMAP$ at time t, respectively. Furthermore, $t_i^f(\ell)$ and $t_i^m(\ell)$ denote the i-polling epoch and the i-departure epoch in the ℓ-th cycle, for $\ell \geq 1$, respectively.

We define the $1 \times L$ vector GFs of the stationary number of i-customers at arbitrary epoch, $\widehat{\mathbf{q}}_i\,(z)$, at i-polling epoch, $\widehat{\mathbf{f}}_i\,(z)$, and at i-departure epoch, $\widehat{\mathbf{m}}_i\,(z)$, by their j-th element, $j = 1 \ldots L$ as

$$[\widehat{\mathbf{q}}_i\,(z)]_j = \lim_{t\to\infty}\sum_{n=0}^{\infty} Pr\{N_i(t) = n, J_i(t) = j\}z^n \quad |z| \leq 1.$$

$$[\widehat{\mathbf{f}}_i\,(z)]_j = \lim_{\ell\to\infty}\sum_{n=0}^{\infty} Pr\{N_i(t_i^f(\ell)) = n, J_i(t_i^f(\ell)) = j\}z^n \quad |z| \leq 1,$$

$$[\widehat{\mathbf{m}}_i\,(z)]_j = \lim_{\ell\to\infty}\sum_{n=0}^{\infty} Pr\{N_i(t_i^m(\ell)) = n, J_i(t_i^m(\ell)) = j\}z^n \quad |z| \leq 1.$$

Theorem 1. *(Expression of $\widehat{\mathbf{q}}_i\,(z)$.) In the stable $BMAP/G/1$ cyclic polling model with a set of disciplines including the binomial-gated and binomial-exhaustive ones the following relation holds for steady-state vector GF of the number of i-customers:*

$$\widehat{\mathbf{q}}_i\,(z)\,\widehat{\mathbf{D}}_i\,(z)\left(z\mathbf{I} - \widehat{\mathbf{A}}_i(z)\right) = \frac{1}{c}\,(z-1)\left(\widehat{\mathbf{f}}_i\,(z) - \widehat{\mathbf{m}}_i\,(z)\right)\widehat{\mathbf{A}}_i(z). \tag{1}$$

Proof. The proof of the theorem can be found in [6]. Here we presented an equivalent form of the constant on the r.h.s. of (1), which can be obtained by applying equilibrium arguments (see [6]).

Let matrix $\widehat{\mathbf{T}}_i\,(z)$ be defined as:

$$\widehat{\mathbf{T}}_i\,(z) = \widehat{\mathbf{D}}_i\,(z)\left(z\mathbf{I} - \widehat{\mathbf{A}}_i(z)\right).$$

Theorem 2. *(The vector factorial moment formula for* $\mathbf{q}_i^{(n)}$*.) In the stable* $BMAP/G/1$ *cyclic polling model satisfying assumptions* **A.1** - **A.3** *with a set of disciplines including the binomial-gated and binomial-exhaustive ones the recursive formula for computing the factorial moments of the stationary number of* i*-customers at an arbitrary instant is given by:*

$$\begin{aligned}\mathbf{q}_i^{(n)} = & \frac{1}{c}\sum_{l=0}^{n+1}\sum_{k=0}^{n+1-l}\binom{n+2}{1,l,n+1-k-l,k}\left(\mathbf{f}_i^{(l)}-\mathbf{m}_i^{(l)}\right)\mathbf{A}_i^{(n+1-k-l)} \\ & \times\frac{[adj\mathbf{T}_i]^{(k)}}{(1+2n+1_{(n\geq 2)}\binom{n}{2})[det\mathbf{T}_i]^{(2)}} \\ & -\boldsymbol{\pi}_i\frac{[det\mathbf{T}_i]^{(n+2)}}{(1+2n+1_{(n\geq 2)}\binom{n}{2})[det\mathbf{T}_i]^{(2)}} \\ & -\left(\left(1_{(n\geq 3)}\sum_{k=1}^{n-2}\binom{n}{k+2}+1_{(n\geq 2)}\sum_{k=1}^{n-1}\left(\binom{n}{k+1}+\binom{n+1}{k+1}\right)\right)\mathbf{q}_i^{(n-k)}\right) \\ & \times\frac{[det\mathbf{T}_i]^{(k+2)}}{(1+2n+1_{(n\geq 2)}\binom{n}{2})[det\mathbf{T}_i]^{(2)}} \quad n\geq 1, \end{aligned} \tag{2}$$

where $1_{(con)}$ *denotes the indicator of condition "con".*

Proof. The proof of the theorem can be found in [5]. Here we presented an equivalent form of the constant on the r.h.s. of (1), which can be obtained by applying equilibrium arguments (see [6]).

4 Discipline Dependent Analysis of the Model with Binomial Disciplines

In this section we provide the service discipline specific solution for nonzero-switchover-times polling model with binomial-gated and binomial-exhaustive disciplines. In order to determine the discipline specific quantities $\mathbf{f}_i^{(n)}$ and $\mathbf{m}_i^{(n)}$ $(n\geq 1)$ used by the formula (2) first we determine the joint probabilities of the steady-state number of customers and the phases of the $BMAP$s at i-polling and i-departure epochs.

4.1 Notations

The joint probabilities of the steady-state number of customers and the phases of the BMAPs at i-polling and i-departure epochs are described as hypervectors. Notation $\otimes$ stands the Kronecker product and $\mathbf{e}_j=(0,\ldots,1,\ldots,0)$ denotes the $1\times L$ vector with 1 at the j-th position. The $1\times L^N$ stationary probability hypervector $\mathbf{p}_i^f(n_1,\ldots,n_N)$ is defined as

$$\mathbf{p}_i^f\,(n_1,\ldots,n_N)=\lim_{\ell\to\infty}\sum_{j_1=1}^{L}\ldots\sum_{j_N=1}^{L}\mathbf{e}_{j_1}\otimes\ldots\otimes\mathbf{e}_{j_N}$$
$$Pr\big\{N_1(t_i^f(\ell))=n_1,\ldots,N_N(t_i^f(\ell))=n_N,J_1(t_i^f(\ell))=j_1,\ldots,J_N(t_i^f(\ell))=j_N\big\},$$
$$n_1,\ldots,n_N\in\{0,1,\ldots\};\quad i=1,\ldots,N.$$

Similarly the $1\times L^N$ steady-state probability hypervector $\mathbf{p}_i^m\,(n_1,\ldots,n_N)$ is defined as:

$$\mathbf{p}_i^m\,(n_1,\ldots,n_N)=\lim_{\ell\to\infty}\sum_{j_1=1}^{L}\ldots\sum_{j_N=1}^{L}\mathbf{e}_{j_1}\otimes\ldots\otimes\mathbf{e}_{j_N}$$
$$Pr\big\{N_1(t_i^m(\ell))=n_1,\ldots,N_N(t_i^m(\ell))=n_N,J_1(t_i^m(\ell))=j_1,\ldots,J_N(t_i^m(\ell))=j_N\big\},$$
$$n_1,\ldots,n_N\in\{0,1,\ldots\};\quad i=1,\ldots,N.$$

Based on these quantities we define the steady-state hypervector GFs of the number of customers at i-polling and i-departure epochs as

$$\widehat{\mathbf{f}}_i(z_1,\ldots,z_N)=\sum_{n_1=0}^{\infty}\ldots\sum_{n_N=0}^{\infty}\mathbf{p}_i^f\,(n_1,\ldots,n_N)\,z_1^{n_1}\ldots z_N^{n_N},$$
$$\widehat{\mathbf{m}}_i(z_1,\ldots,z_N)=\sum_{n_1=0}^{\infty}\ldots\sum_{n_N=0}^{\infty}\mathbf{p}_i^m\,(n_1,\ldots,n_N)\,z_1^{n_1}\ldots z_N^{n_N},$$
$$i=1,\ldots,N;\quad |z_1|\le 1,\ldots,|z_N|\le 1.$$

We define the homogenous bivariate Markov chain $\big\{(N_i(t_i^d(n)),J_i(t_i^d(n)));n\in\{1,\ldots\}\big\}$ on the state space $\{0,1,\ldots\}\times\{1,2,\ldots,L\}$, where $t_i^d(n)$ denotes the n-th i-customer departure epoch during the same server visit at station i for $n\ge 1$. We define matrix $\mathbf{G}_i(t)$, $t\ge 0$, whose (j,ℓ)-th element is given as the probability that the first passage starting from state $(n+1,j)$ in the Markov chain to the level n, $n\in 0,1,2,\ldots$, $1\le j,\ell\le L$, occurs no later than time t, and the first state visited in level n is (n,ℓ).

We use notation $\oplus$ for the Kronecker sum and $\oplus_{k=1}^{N}\widehat{\mathbf{D}}_k(z_k)$ stands for $\widehat{\mathbf{D}}_1(z_1)\oplus\ldots\oplus\widehat{\mathbf{D}}_N(z_N)$. Additionally we introduce several further notations as follows:

$$\widehat{\mathbf{A}}_i(z_1,\ldots,z_N)=\int_0^{\infty}e^{t\oplus_{k=1}^{N}\widehat{\mathbf{D}}_k(z_k)}dB_i(t),$$
$$\widehat{\mathbf{U}}_i(z_1,\ldots,z_N)=\int_0^{\infty}e^{t\oplus_{k=1}^{N}\widehat{\mathbf{D}}_k(z_k)}dR_i(t).$$
$$\widehat{\mathbf{H}}_i(z_1,\ldots,z_{i-1},z_{i+1},\ldots,z_N)=\int_0^{\infty}e^{t\oplus_{k=1}^{i-1}\widehat{\mathbf{D}}_k(z_k)}\otimes d\mathbf{G}_i(t)\otimes e^{t\oplus_{k=i+1}^{N}\widehat{\mathbf{D}}_k(z_k)}.$$

Note that $\widehat{\mathbf{A}}_i(z_1,\ldots,z_N)$, $\widehat{\mathbf{U}}_i(z_1,\ldots,z_N)$ and $\widehat{\mathbf{H}}_i(z_1,\ldots,z_{i-1},z_{i+1},\ldots,z_N)$ are all $L^N \times L^N$ hypermatrices.

4.2 Polling Model with Binomial-Gated Discipline

Theorem 3. *(Governing equations of the system.)The governing equations of the stable $BMAP/G/1$ cyclic nonzero-switchover-times polling model satisfying assumptions* **A.1** - **A.3** *and with binomial-gated service discipline are given in terms of the hypervector GFs $\widehat{\mathbf{f}}_i(z_1,\ldots,z_N)$ and $\widehat{\mathbf{m}}_i(z_1,\ldots,z_N)$ for $i=1,\ldots,N$ as*

$$\widehat{\mathbf{m}}_i\,(z_1,\ldots,z_N)=\widehat{\mathbf{f}}_i\left(z_1,\ldots,z_{i-1},\left(p\widehat{\mathbf{A}}_i(z_1,\ldots,z_N)+(1-p)\mathbf{I}z_i\right),z_{i+1},\ldots,z_N\right),$$
$$\widehat{\mathbf{f}}_{i+1}\,(z_1,\ldots,z_N)=\widehat{\mathbf{m}}_i\,(z_1,\ldots,z_N)\,\widehat{\mathbf{U}}_i(z_1,\ldots,z_N). \tag{3}$$

Proof. The hypermatrix GF of the number of simultaneously arriving k-customers for $k=1,\ldots,N$ during the interval $(0,t)$, where t is independent of the arrival processes, is given as $e^{\widehat{\mathbf{D}}_1(z_1)t}\otimes\ldots\otimes e^{\widehat{\mathbf{D}}_N(z_N)t}=e^{(\widehat{\mathbf{D}}_1(z_1)\oplus\ldots\oplus\widehat{\mathbf{D}}_N(z_N))t}$. It follows that the hypermatrix GF of the number of simultaneously arriving k-customers for $k=1,\ldots,N$ during the service of one i-customer can be expressed as $\int_0^\infty\left(e^{t\oplus_{k=1}^N\widehat{\mathbf{D}}_k(z_k)}\right)dB_i(t)=\widehat{\mathbf{A}}_i(z_1,\ldots,z_N)$.

Under binomial-gated discipline every customer which is present at the polling epoch is served with probability p $(0<p\leq 1)$ independently from the other ones. Let us condition on the number of j-customers, for $j=1,\ldots,N$, present at the i-polling epoch, $n_1,\ldots,n_N$, for $n_1,\ldots,n_N\geq 0$. Then the probability that $0\leq k\leq n$ i-customers get service is $\binom{n}{k}p^k(1-p)^{n-k}$. Each of the k i-customers getting service generates a random population of simultaneously arriving k-customers for $k=1,\ldots,N$ arriving during its service time, whose hypermatrix GF is $\widehat{\mathbf{A}}_i(z_1,\ldots,z_N)$. Each of the other $n-k$ i-customers does not cause any change in the number of j-customers, for $j=1,\ldots,N$, thus this transition can be described by a hypermatrix $\mathbf{I}z_i$. Since $\mathbf{I}$ and $\widehat{\mathbf{A}}_i(z_1,\ldots,z_N)$ commute, for any selection order of the customers the hypermatrix GF of the number of j-customers, for $j=1,\ldots,N$, at the next i-departure epoch is $z_1^{n_1}\ldots z_{i-1}^{n_{i-1}}\widehat{\mathbf{A}}_i^k(z_1,\ldots,z_N)(\mathbf{I}z_i)^{n-k}z_{i+1}^{n_{i+1}}$. Using it and applying the binomial theorem yields to the hypermatrix GF of the number of customers at next i-departure epoch, given that there is $n_1,\ldots,n_N$ j- customers, for $j=1,\ldots,N$, present at the previous i-polling epoch, as

$$\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}z_1^{n_1}\ldots z_{i-1}^{n_{i-1}}\widehat{\mathbf{A}}_i^k(z_1,\ldots,z_N)(\mathbf{I}z_i)^{n-k}z_{i+1}^{n_{i+1}}\ldots z_N^{n_N}$$
$$=z_1^{n_1}\ldots z_{i-1}^{n_{i-1}}\left(p\widehat{\mathbf{A}}_i(z_1,\ldots,z_N)+(1-p)\mathbf{I}z_i\right)^{n_i}z_{i+1}^{n_{i+1}}\ldots z_N^{n_N}. \tag{4}$$

Unconditioning (4) yields

$$\begin{aligned}
&\sum_{n_1=0}^{\infty} \dots \sum_{n_N=0}^{\infty} \mathbf{p}_i^f\,(n_1,\dots,n_N)\, z_1^{n_1} \dots z_{i-1}^{n_{i-1}} \left(p\widehat{\mathbf{A}}_i(z_1,\dots,z_N) + (1-p)\mathbf{I}z_i\right)^{n_i} \\
&\times z_{i+1}^{n_{i+1}} \dots z_N^{n_N}, \\
&= \sum_{n_1=0}^{\infty} \dots \sum_{n_N=0}^{\infty} \mathbf{p}_i^m\,(n_1,\dots,n_N)\, z_1^{n_1} \dots z_N^{n_N}, \quad i = 1,\dots,N.
\end{aligned} \tag{5}$$

Applying similar arguments as before yields the relation for the transition $m_i \to f_{i+1}$ of the binomial-gated polling model as

$$\begin{aligned}
&\sum_{n_1=0}^{\infty} \dots \sum_{n_N=0}^{\infty} \mathbf{p}_i^m\,(n_1,\dots,n_N)\, z_1^{n_1} \dots z_N^{n_N}\, \widehat{\mathbf{U}}_i(z_1,\dots,z_N) \\
&= \sum_{n_1=0}^{\infty} \dots \sum_{n_N=0}^{\infty} \mathbf{p}_{i+1}^f\,(n_1,\dots,n_N)\, z_1^{n_1} \dots z_N^{n_N}, \quad i = 1,\dots,N.
\end{aligned} \tag{6}$$

The theorem comes by applying the compact notation for the steady-state hypermatrix GFs in (5) and (6).

The steady-state probability hypervectors $\mathbf{p}_i^f\,(n_1,\dots,n_N)$ and $\mathbf{p}_i^m\,(n_1,\dots,n_N)$, for $i = 1,\dots,N$, can be determined from the equations, which can be obtained by setting an upper limit X for $n_1,\dots,n_N$ in (5) and (6) and taking their x_1-th, …, x_N-th derivatives ($x_1,\dots,x_N \in \{0,\dots,X\}$) at $z_1 = \dots = z_N = 1$, respectively. This results in the following system of linear equations for $i = 1,\dots,N$ and $x_1,\dots,x_N \in \{0,\dots,X\}$:

$$\begin{aligned}
&\sum_{n_1=0}^{X} \dots \sum_{n_N=0}^{X} \mathbf{p}_i^f\,(n_1,\dots,n_N) \\
&\times \left.\frac{d^{x_1}\dots d^{x_N}\left(z_1^{n_1}\dots z_{i-1}^{n_{i-1}}\left(p\widehat{\mathbf{A}}_i(z_1,\dots,z_N)+(1-p)\mathbf{I}z_i\right)^{n_i} z_{i+1}^{n_{i+1}}\dots z_N^{n_N}\right)}{dz_1{}^{x_1}\dots dz_N{}^{x_N}}\right|_{\mathbf{z}=1} \\
&= \sum_{n_1=x_1}^{X} \dots \sum_{n_N=x_N}^{X} \mathbf{p}_i^m\,(n_1,\dots,n_N)\, \frac{n_1!}{(n_1-x_1)!} \dots \frac{n_N!}{(n_N-x_N)!},
\end{aligned} \tag{7}$$

$$\begin{aligned}
&\sum_{n_1=0}^{X} \dots \sum_{n_N=0}^{X} \mathbf{p}_i^m\,(n_1,\dots,n_N) \left.\frac{d^{x_1}\dots d^{x_N}\left(z_1^{n_1}\dots z_N^{n_N}\,\widehat{\mathbf{U}}_i(z_1,\dots,z_N)\right)}{dz_1{}^{x_1}\dots dz_N{}^{x_N}}\right|_{\mathbf{z}=1} \\
&= \sum_{n_1=x_1}^{X} \dots \sum_{n_N=x_N}^{X} \mathbf{p}_{i+1}^f\,(n_1,\dots,n_N)\, \frac{n_1!}{(n_1-x_1)!} \dots \frac{n_N!}{(n_N-x_N)!},
\end{aligned} \tag{8}$$

where $\mathbf{z} = 1$ stands for $z_1 = \dots = z_N = 1$.

4.3 Polling Model with Binomial-Exhaustive

Theorem 4. *(Governing equations of the system.)The governing equations of the stable $BMAP/G/1$ cyclic nonzero-switchover-times polling model satisfying assumptions* **A.1** - **A.3** *and with binomial-exhaustive service discipline are given in terms of the hypervector GFs $\widehat{\mathbf{f}}_i(z_1,\ldots,z_N)$ and $\widehat{\mathbf{m}}_i(z_1,\ldots,z_N)$ for $i=1,\ldots,N$ as*

$$\begin{aligned}
&\widehat{\mathbf{m}}_i\left(z_1,\ldots,z_{i-1},1,z_{i+1},\ldots,z_N\right)\\
&=\widehat{\mathbf{f}}_i\left(z_1,\ldots,z_{i-1},\left(p\widehat{\mathbf{H}}_i(z_1,\ldots,z_{i-1},z_{i+1},\ldots,z_N)+(1-p)\mathbf{I}z_i\right),z_{i+1},\ldots,z_N\right),\\
&\widehat{\mathbf{f}}_{i+1}\left(z_1,\ldots,z_N\right)=\widehat{\mathbf{m}}_i\left(z_1,\ldots,z_{i-1},1,z_{i+1},\ldots,z_N\right)\widehat{\mathbf{U}}_i(z_1,\ldots,z_N). \qquad (9)
\end{aligned}$$

Proof. Under binomial-exhaustive discipline each customer present at the polling epoch is handled together with the newly arrived ones during its associated busy period. Every such customer group is served with probability p. Based on this and applying similar argument as before for the model with the binomial-gated discipline results in the statement.

The stationary probability hypervectors $\mathbf{p}_i^f\left(n_1,\ldots,n_N\right)$ and $\mathbf{p}_i^m\left(n_1,\ldots,n_N\right)$, for $i=1,\ldots,N$, can be determined again from a system of linear equations which is given for $i=1,\ldots,N$ and $u_1,\ldots,u_N\in\{0,\ldots,U\}$ as

$$\begin{aligned}
&\sum_{n_1=0}^{X}\cdots\sum_{n_N=0}^{X}\mathbf{p}_i^f\left(n_1,\ldots,n_N\right)\frac{d^{x_1}\ldots d^{x_{i-1}}d^{x_{i+1}}\ldots d^{x_N}}{dz_1{}^{x_1}\ldots dz_{i-1}{}^{x_{i-1}}\ dz_{i+1}{}^{x_{i+1}}\ldots dz_N{}^{x_N}}\\
&\times\left(z_1^{n_1}\ldots z_{i-1}^{n_{i-1}}\left(p\widehat{\mathbf{H}}_i(z_1,\ldots,z_{i-1},z_{i+1},\ldots,z_N)+(1-p)\mathbf{I}z_i\right)^{n_i}z_{i+1}^{n_{i+1}}\ldots z_N^{n_N}\right)\Big|_{\mathbf{z}=1} \qquad (10)\\
&=\sum_{n_1=x_1}^{X}\cdots\sum_{n_{i-1}=x_{i-1}}^{X}\sum_{n_{i+1}=x_{i-1}}^{X}\cdots\sum_{n_N=x_N}^{X}\mathbf{p}_i^m\left(n_1,\ldots,n_{i-1},0,n_{i+1},\ldots,n_N\right)\\
&\times\frac{n_1!}{(n_1-x_1)!}\cdots\frac{n_{i-1}!}{(n_{i-1}-x_{i-1})!}\frac{n_{i+1}!}{(n_{i+1}-x_{i+1})!}\cdots\frac{n_N!}{(n_N-x_N)!},
\end{aligned}$$

$$\begin{aligned}
&\sum_{n_1=0}^{X}\cdots\sum_{n_{i-1}=0}^{X}\sum_{n_{i+1}=0}^{X}\cdots\sum_{n_N=0}^{X}\mathbf{p}_i^m\left(n_1,\ldots,n_{i-1},0,n_{i+1},\ldots,n_N\right)\\
&\quad\times\frac{d^{x_1}\ldots d^{x_N}\left(z_1^{n_1}\ldots z_{i-1}^{n_{i-1}}z_{i+1}^{n_{i+1}}\ldots z_N^{n_N}\widehat{\mathbf{U}}_i(z_1,\ldots,z_N)\right)}{dz_1{}^{x_1}\ldots dz_N{}^{x_N}}\Bigg|_{\mathbf{z}=1} \qquad (11)\\
&=\sum_{n_1=x_1}^{X}\cdots\sum_{n_N=x_N}^{X}\mathbf{p}_{i+1}^f\left(n_1,\ldots,n_N\right)\frac{n_1!}{(n_1-x_1)!}\cdots\frac{n_N!}{(n_N-x_N)!}.
\end{aligned}$$

5 Final Remarks

The numerical procedure for computation of the vector factorial moments for the model with the binomial-gated (binomial-exhaustive) discipline consists of the same steps as for the model with the gated (exhaustive) disciplines. For details see [5].

The total number of operations required by the whole numerical procedure for the model with the binomial-gated discipline is in the magnitude of $N^2\ L^{3N}(X+1)^{3N}$, while it is $N^2\ L^{3N}(X+1)^{3N-3}$ for the system with binomial-exhaustive discipline. The total number of required elementary computational steps increases with X, L and with N. Hence the numerical solution becomes computationally intensive when the server utilization is high, the number of BMAP phases is high or the system is large.

Setting $p = 1$ in the model with binomial-gated discipline yields the model with gated discipline as special case. Similarly the model with exhaustive discipline can be obtained as special case by setting $p = 1$ in the model with binomial-exhaustive discipline.

References

1. Boxma, O.J.: Workloads and waiting times in Single-server systems with multiple customer classes. Queueing Systems 5, 185–214 (1989)
2. Levy, H.: Analysis of cyclic polling systems with binomial-gated service. In: Hasegawa, T., Takagi, H., Takahashi, Y. (eds.) Performance of Distributed and Parallel Systems, pp. 127–139. Elsevier Science Publishers, North-Holland (1989)
3. Lucantoni, D.L.: New results on the single server queue with a batch markovian arrival process. Stochastic Models 7, 1–46 (1991)
4. Neuts, M.F.: Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. The John Hopkins University Press, Baltimore (1981)
5. Saffer, Z.: Unified Analysis of Cyclic Polling Models with BMAP. Ph.D. thesis, Department of Telecommunications, Budapest University of Technology and Economics (2010)
6. Saffer, Z., Telek, M.: Unified analysis of BMAP/G/1 cyclic polling models. Queueing Systems 64(1), 69–102 (2010)
7. Takagi, H.: Analysis of Polling Systems. MIT Press (1986)
8. Takagi, H.: Analysis and Application of Polling Models. In: Reiser, M., Haring, G., Lindemann, C. (eds.) Dagstuhl Seminar 1997. LNCS, vol. 1769, pp. 423–442. Springer, Heidelberg (2000)

Analysis of Fluid Queues in Saturation with Additive Decomposition

Miklós Telek and Miklós Vécsei

Budapest University of Technology and Economics
Department of Telecommunications
1521 Budapest, Hungary
{telek,vecsei}@webspn.hit.bme.hu

Abstract. Fluid queueing models with finite capacity buffers are applied to analyze a wide range of real life systems. There are well established numerical procedures for the analysis of these queueing models when the load is lower or higher than the system capacity, but these numerical methods become unstable as the load gets close to the system capacity. One of the available numerical procedures is the additive decomposition method proposed by Nail Akar and his colleagues.

The additive decomposition method is based on a separation of the eigenvalues of the characterizing matrix into the zero eigenvalue, the eigenvalues with positive real part and the eigenvalues with negative real part. The major problem of the method is that the number of zero eigenvalues increases by one at saturation. In this paper we present an extension of the additive decomposition method which remain numerically stable at saturation as well.

Keywords: Markov fluid queue, additive decomposition method.

1 Introduction

Intuitively it is quite clear that infinite buffer queueing systems remain stable as long as the system load is below the system capacity. It is also widely accepted that finite buffer systems remain stable also when the system load is higher than the system capacity. This second statement suggests that finite buffer systems can be easily analyzed for any load level. In contrast, it turns out that standard solution methods suffer from severe numerical instabilities at the region where the load is close to the system capacity. It is interesting to note that analysis methods of finite buffer queueing systems used for the dimensioning of telecommunication network components are typically used for evaluating models close to saturation.

Apart of this practical issue, the common analysis approaches of finite buffer queueing systems exclude the case of saturation, because the discussion is restricted to the case when the load is below the system capacity and it is commonly left for the reader to invert the buffer content process if the load is higher than the system capacity. Unfortunately, this approach does not help when the load is equal to the system capacity.

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 167–176, 2013.

In this paper we consider Markov fluid queues (MFQs) with finite fluid buffers. There is a wide literature devoted to this subject (see e.g., [6,1,7,3,9,5]) for the case when the load is different from the capacity, but the case of saturation is considered only recently in [10] for the method proposed by Soares and Latouche in [7,8]. Here we investigate an other analysis method, the additive decomposition, which is proposed in [9], [5]. We propose an modification of the method which remains applicable in case of saturation.

The rest of the paper is organized as follows. Section 2 introduces Markov fluid queues (MFQs) with finite buffer and their analytical description. The next section discusses the additive decomposition method. The first subsection of Section 3 presents the solution method applicable for systems below and above saturation. The next subsection contains the proposed modification of the procedure for the case of saturation. Section 4 demonstrates the numerical properties of the standard and the proposed analysis methods. The paper is concluded in Section 5.

2 Markov Fluid Queue

The evolution of Markov fluid queue with single fluid buffer is determined by a discrete state of the environment and the continuous fluid level in the fluid buffer. The $Z(t) = \{M(t), X(t); t \geq 0\}$ process represents the state of the MFQ, where $M(t) \in \mathcal{S}$ is the (discrete) state of the environment and $X(t) \in [0, b]$ is the fluid level in the fluid buffer at time t, where b denotes the buffer size. The fluid level cannot be negative or greater than b. We define $\hat{\pi}_j(t,x)$, $\hat{p}_j(t,0)$ and $\hat{p}_j(t,b)$ to describe the transient fluid densities at fluid level x and the transient probability masses of the fluid distribution at idle and full buffer as follows

$$\hat{\pi}_j(t,x) = \lim_{\Delta \to 0} \frac{Pr(M(t) = j, x \leq X(t) < x + \Delta)}{\Delta} ,$$

$$\hat{p}_j(t,x) = Pr(M(t) = j, X(t) = x) \quad x = 0, b.$$

One of the main goal of the analysis of MFQ is to compute the stationary fluid density $\pi_j(x) = \lim_{t \to \infty} \hat{\pi}_j(t,x)$ and fluid mass at idle and full buffer $p_j(x) = \lim_{t \to \infty} \hat{p}_j(t,x)$, $x = 0, b$. The row vector $\pi(x) = \{\pi_j(x)\}$, satisfies [4]

$$\frac{d}{dx}\pi(x)\mathbf{R} = \pi(x)\mathbf{Q} , \tag{1}$$

where matrix $\mathbf{Q} = \{Q_{ij}\}$ is the transition rate matrix of the environment process, and the diagonal matrix $\mathbf{R} = \text{diag}\langle R_j \rangle$ is composed by the fluid rates $R_j, j \in \mathcal{S}$. R_j rate determines the rate at which the fluid level changes (increases when $R_j > 0$ or decreases when $R_j < 0$) when the environment is in state j. In this paper we assume that matrix $\mathbf{Q}$ determines an irreducible Markov chain and exclude the case of $R_j = 0$. If there are states in the model where the fluid level remains constant then a censored process needs to be defined and investigated

where sojourns in states with constant fluid level are excluded. Details of the censored analysis method can be find e.g. in [3]. A consequence of the exclusion of states with constant fluid level is that matrix $\mathbf{R}$ is non-singular. We denote the set of states with positive fluid rates by S^+ and the set of states with negative fluid rates by S^-.

Kulkarni investigated the properties of the characterizing matrix of (1) in [6]. First of all, he defined the stability condition of infinite buffer MFQs. Let γ be the stationary distribution of the CTMC with generator matrix $\mathbf{Q}$. γ is the solution of the linear system $\gamma\mathbf{Q} = 0, \gamma\mathbb{1} = 1$, where $\mathbb{1}$ is the column vector of ones of appropriate size. An infinite buffer MFQ is stable if it "drift" is negative, where the drift is $d = \gamma\mathbf{R}\mathbb{1}$.

Further more differential equation in (1) suggests to find the solution of the fluid density function in a matrix exponential form. To find the matrix exponential solution [6] defines the relation of the number of states with positive and negative fluid rates and the number of eigenvalues of matrix $\mathbf{QR}^{-1}$ with positive and negative real parts. These results are summarized in Table 1.

Table 1. Drift related properties of finite MFQs, where $|S^-|$ ($|S^+|$) is the number of states with negative (positive) fluid rate

	$d < 0$	$d = 0$	$d > 0$
positive eigenvalues	$\|S^-\| - 1$	$\|S^-\| - 1$	$\|S^-\|$
negative eigenvalue	$\|S^+\|$	$\|S^+\| - 1$	$\|S^+\| - 1$
zero eigenvalue	1	2	1

The initial vector of the matrix exponential solution is determined by the boundary conditions.

$$p_i(0) = 0 \text{ for } i \in S^+,\ p_i(b) = 0 \text{ for } i \in S^-, \tag{2}$$

and

$$-\pi_i(0)R_i + \sum_{j \in S^-} p_j(0)Q_{ji} = 0,\ \pi_i(b)R_i + \sum_{j \in S^+} p_j(b)Q_{ji} = 0. \tag{3}$$

(2) states that the fluid level cannot be 0 when the fluid rate is positive and it cannot be b when the fluid rate is negative. For $i \in S^-$ the first part of (3) means that the fluid level can be 0 due to a state transition of the environment from an other state with negative fluid rate or due to the fact that the fluid level reduced to 0 in a state with negative fluid rate. For $i \in S^+$ the first part of (3) represents that the fluid level can start increasing from 0 due to the fact that the process stayed in a state with negative fluid rate at level 0 and a state transition occurred to a state with positive fluid rate. The second part of (3) contains the counterpart statements for buffer level b.

3 The Additive Decomposition Method

A numerically stable approach to the analysis of MFQs is the additive decomposition method [5]. It will be summarized in the following section. It's stability is based on the separation of the eigenvalues of the matrices in equation 1. The original additive decomposition algorithm from [5] can not be applied for fluid queues at saturation directly.

3.1 Fluid Queues at Non-zero Drift

Due to the fact that states with constant fluid rates are excluded we can multiply both sides of (1) with $\mathbf{R}^{-1}$. If we denote $\mathbf{QR}^{-1}$ with $\mathbf{A}$, this will result in the following differential equation:

$$\frac{d}{dx}\pi(x) = \pi(x)\mathbf{A} \tag{4}$$

The usual way of solving equations like (4) is based on it's spectral representation:

$$\pi(x) = e^{\lambda x}\Gamma$$

λ is a scalar and Γ is a row vector. Substituting this form to (1), we find:

$$\lambda\Gamma = \lambda\mathbf{A} \tag{5}$$

After finding the eigenvalues λ_i and eigenvectors Γ_i one may search for $\pi(x)$ as the sum of the results of (4), with a_i parameters:

$$\pi(x) = \sum_i a_i e^{\lambda_i x}\Gamma_i \tag{6}$$

The limitation of this method may appear when we want to fit the formula to the boundary conditions at the buffer limit. The arising equations will define the a_i parameters in (6), hence they are crucial for solving the problem. If the buffer limit is large ($b \to \infty$), than if $\lambda_i > 0 \to e^{\lambda_i b} \to \infty$ moreover if $\lambda_j < 0$, then $e^{\lambda_j b} \to 0$. This will result in badly conditioned linear equations for a_i.

The additive decomposition method solves this problem by separating the eigenvalues based on their sings, and by handling them separately. In [5] a procedure is described with which one may transform $\mathbf{A}$ into a blockmatrix form. (It uses Schur-decomposition and solves a Lyapunov-equation in order to find it.)

$$\mathbf{Y}^{-1}\mathbf{AY} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \mathbf{A}_- & 0 \\ 0 & 0 & \mathbf{A}_+ \end{pmatrix} \tag{7}$$

$\mathbf{A}_-$ ($\mathbf{A}_+$) is a square matrix, and all of it's eigenvalues are negative (positive). Let us denote different parts of $\mathbf{Y}^{-1}$ with the following notations

$$\mathbf{Y} = \begin{pmatrix} \mathbf{L_0} \\ \mathbf{L}_- \\ \mathbf{L}_+ \end{pmatrix} \tag{8}$$

$\mathbf{L_0}$ is the first row of $\mathbf{Y}^{-1}$ while $\mathbf{L}_-$ and $\mathbf{L}_+$ have the same amount of rows as A_- and A_+ respectively. In [5] it is proven, that the following form is also a complete solution of the differential equation (4):

$$\pi(x) = a_0\mathbf{L_0} + a_-e^{\mathbf{A}_-x}\mathbf{L}_- + a_+e^{-\mathbf{A}_+(b-x)}\mathbf{L}_+$$

a_0 is a scalar and a_- and a_+ are row vectors with the same number of columns as $\mathbf{A}_-$ and $\mathbf{A}_+$ respectively. They are the parameters we need to define from the boundary conditions. The linear equations in this case are numerically stable, because all of the eigenvalues of $\mathbf{A}_-$ and $-\mathbf{A}_+$are negative.

3.2 Fluid Queues at Saturation

The additive decomposition method was developed for fluid queues with non-zero mean drift, but the procedure as it is described in the previous subsection does not work in case of saturation. A slight enhancement is needed in order to apply the procedure for MFQs at saturation.

Theorem 1. *In* $\mathbf{A}$*'s normal Jordan form, there is one Jordan-block belonging to the zero eigenvalue, and it's size is* 2×2*.*

Proof. The numbers of eigenvalues of different signs are given in [6] and are summarized in Table 1. The multiplicity of the zero eigenvalue is 2 in saturation. Now we need to show that there is a single (linear independent) eigenvector associated with the zero eigenvalue, because in this case the Jordan decomposition contains a Jordan block of size 2.

The left eigenvector associated with the zero eigenvalue satisfies

$$\alpha\mathbf{Q}\mathbf{R}^{-1} = 0$$

Multiplying both sides with $\mathbf{R}$ shows that α should also be the left eigenvector of $\mathbf{Q}$ associated with the zero eigenvalue. Due to the fact that $\mathbf{Q}$ defines an irreducible Markov chain it has only a single (linear independent) eigenvector associated with the zero eigenvalue and it is γ.

Corollary 1. *It is not possible to transform* $\mathbf{A}$ *to the same form as in* (7)*.*

In case of a MFQ in saturation, instead of having a single matrix element associated with the zero eigenvalue, we have a Jordan-block of size 2×2 in the similar decomposition of $\mathbf{A}$ as the one in (7). Hence one needs to modify the original method for MFQs at saturation. The proposed modification is to transform A to the following form

$$\mathbf{Y}^{-1}\mathbf{A}\mathbf{Y} = \begin{pmatrix} \mathbf{A_0} & 0 & 0 \\ 0 & \mathbf{A}_- & 0 \\ 0 & 0 & \mathbf{A}_+ \end{pmatrix}, \tag{9}$$

where $\mathbf{A_0}$ corresponds to the 0 eigenvalues. Consequently $\mathbf{L_0}$ will have two rows, a_0 will have two elements in (8), and the expression for $\pi(x)$ changes to

$$\pi(x) = a_0 e^{\mathbf{A_0}x}\mathbf{L_0} + a_- e^{\mathbf{A}_- x}\mathbf{L}_- + a_+ e^{-\mathbf{A}_+(b-x)}\mathbf{L}_+$$

Unfortunately, this formula is not stable for large buffer limits. This happens, because one the off-diagonal elements of the Jordan-block $\mathbf{A_0}$ are nonzero. For example, if it is an upper tridiagonal matrix then

$$\mathbf{A_0} = \begin{pmatrix} 0 & a_{12} \\ 0 & 0 \end{pmatrix} \rightarrow e^{\mathbf{A_0}x} = \begin{pmatrix} 1 & a_{12}x \\ 0 & 1 \end{pmatrix},$$

and $a_{12}x \rightarrow \infty$ as $x \rightarrow \infty$, therefore this matrix will be badly conditioned for large buffer limits. Thus one might experience numerical problems when fitting the parameters of the system to the boundary conditions.

4 Numerical Examples

We analyzed the numerical properties of the algorithms for finite buffer MFQs using our MATLAB implementations, which are parts of the BuTools package (available at http://webspn.hit.bme.hu/∼butools/). We compared the proposed procedure (Section 3.2), with the original additive decomposition method (Section 3.1) at two different drift values, one far from zero and one close to zero.

4.1 Comparison of Methods When the Drift Is Far from Zero

First we evaluated the MFQ with buffer size $b = 30$, generator matrix and fluid rate matrix

$$\mathbf{Q} = \left[\begin{array}{cc|ccc} -4 & 0 & 2 & 1 & 1 \\ 3 & -6 & 0 & 2 & 1 \\ \hline 1 & 3 & -5 & 1 & 0 \\ 3 & 1 & 1 & -7 & 2 \\ 1 & 1 & 0 & 1 & -3 \end{array}\right], \quad \mathbf{R} = \left[\begin{array}{cc|ccc} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{array}\right],$$

respectively. The stationary distribution of the CTMC characterized by $\mathbf{Q}$ is $\gamma = (0.314, 0.142, 0.154, 0.143, 0.247)$ and the drift is $d = -0.00933$. To quantify the difference between the results of the methods we used the following error measure:

$$\Delta = \sum_{i \in S} \int_0^b |\pi_i^O(x) - \pi_i^M(x)| dx + \sum_{i \in S} |p_i^O(0) - p_i^M(0)| + \sum_{i \in S} |p_i^O(b) - p_i^M(b)| \ ,$$

where $\pi_i^O(x)$ and $\pi_i^M(x)$ correspond to the fluid density for state i at level x for the original and the modified algorithms. $p_i^O(0)$ and $p_i^M(0)$ are the probabilities for the empty buffer and $p_i^O(b)$ and $p_i^M(b)$ are for the full buffer. The fluid density curves computed by the two methods are depicted in Figure 1.

We also calculated the difference between the methods for systems with state space cardinalities of 20 and 50. The results were similar. The average error was $\Delta \sim 10^{-5}$.

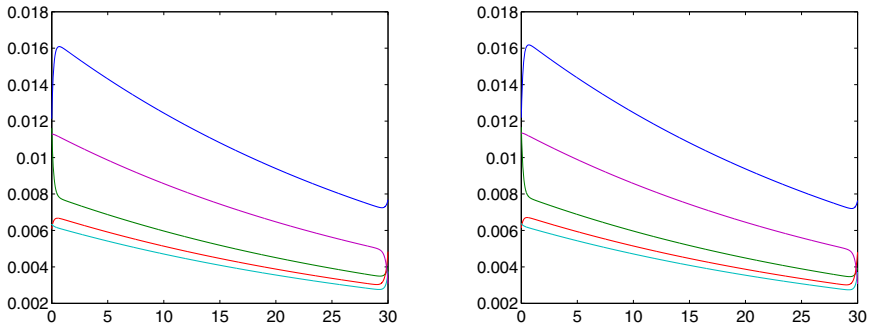


Fig. 1. The fluid density functions ($\pi_i(x)$ versus fluid level x) of the example with non-zero drift ($b = 30$, $d = -0.00933$). The left graph corresponds to the method proposed in [5], the right graph corresponds to the method proposed in Section 3.2.

4.2 Comparison of the Methods When the Drift Is Close to Zero

In our second example the buffer size is $b = 30$ the generator matrix and the fluid rate matrix are

$$\mathbf{Q} = \left[\begin{array}{cc|ccc} -5 & 3 & 1 & 0 & 1 \\ 5 & -8 & 0 & 2 & 1 \\ \hline 1 & 0 & -4 & 2 & 1 \\ 4 & 1 & 0 & -6 & 1 \\ 1 & 0 & 0 & 2 & -3 \end{array}\right], \quad \mathbf{R} = \left[\begin{array}{cc|ccc} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{array}\right].$$

The stationary distribution for this CTMC process is $\gamma = (0.349, 0.151, 0.087, 0.163, 0.250)$ and the drift is $d = -1.11 \cdot 10^{-16}$. The original method proposed in [5] and summarized in Section 3.1 failed in the phase of decomposition according to the signs of the eigenvalues using the standard numerical precision of MATLAB, while the modified method completes. The obtained fluid density curve is depicted in Figure 2. When the drift is close to zero the original procedure gets numerically instable as it is clearly visible on the figure.

4.3 Analysis of a Communication System with RED

We analyze a communication system using the proposed method. The fluid level represents the amount of data in the buffer, and the data arrival and service processes are modulated by an environmental Markov chain with generator Q. There are N identical users in the system. They are either in the ON or in the OFF state. In the ON state they transmit data at rate r, otherwise they do not. The sojourn time in state ON (OFF) is exponentially distributed with parameter α (β). The service speed of the server is c, and reject incoming data with probability $1-s$, consequently data arrive to the server at rate $r*s$. This last

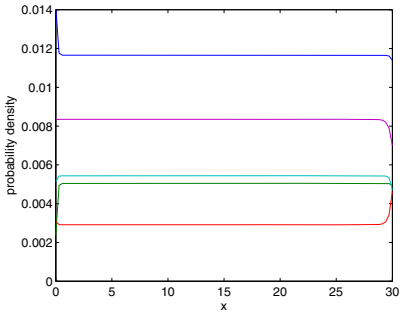


Fig. 2. The fluid density function ($\pi_i(x)$ versus fluid level x) for a queue with zero drift ($d = -1.11 \cdot 10^{-16}$, $b = 30$)

functional property is referred to as "random early detection" (RED) mechanism [2]. The RED method filters the input data as a function of the fluid level, namely $s(x)$ is a function of the fluid level x. Assuming that $s(x)$ is a piecewise constant function the multi region version of the adaptive decomposition method [5] and its modification for the case of zero drift in Section 3.2 allows to analyze the described communication systems. The limits of the constant regions of $s(x)$ are denoted by x_j $(j = 0, 1, \ldots, k)$, such that $x_0 = 0$ and $x_k = B$.

Due to the identity of the N users a MFQ with $N + 1$ states describe the system behavior with generator matrix

$$\mathbf{Q} = \begin{pmatrix} -N\beta & N\beta & 0 & 0 & 0 & 0 \\ \alpha & -\alpha-(N-1)\beta & (N-1)\beta & 0 & 0 & 0 \\ 0 & 2\alpha & -2\alpha-(N-2)\beta & (N-2)\beta & 0 & 0 \\ 0 & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & 0 & 0 & (N-1)\alpha & -(N-1)\alpha-\beta & \beta \\ 0 & 0 & 0 & 0 & N\alpha & -N\alpha \end{pmatrix},$$

and fluid rate matrix

$$\mathbf{R}(x) = \begin{pmatrix} -c & 0 & 0 & 0 & 0 & \\ 0 & rs(x)-c & 0 & 0 & 0 & 0 \\ 0 & 0 & 2rs(x)-c & 0 & 0 & 0 \\ 0 & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & 0 & 0 & 0 & (N-1)rs(x)-c & \\ 0 & 0 & 0 & 0 & 0 & Nrs(x)-c \end{pmatrix}.$$

One of the most important performance measure of this system is the loss. The loss is the amount of lost data. Loss may be caused by two phenomenons. The first is the filtering of the RED mechanism. When n users are ON the loss rate is $L_1 = (1 - s)nr$. The second reason for the loss is the finite buffer. The server may also loose data when the buffer is full. As the buffer is served with speed c, the loss rate due to the finite buffer capacity is $L_2 = snr - c$. These two parts of the loss can be computed as

$$L_1 = \int_0^B r_i(1 - s(x))f_i(x)dx + \sum_{j,k} p(x_j, k)r_k(1 - s(x_j),$$
$$L_2 = \sum_k p(B, k)(s(B)r_k - c)$$

where $f_i x$ is the stationary probability density for state i, and $p(x_j, k)$ is the probability at threshold level x_j for state k. Based on these loss rates the loss ratio is

$$L = \frac{L_1 + L_2}{\int_0^B r_i f_i(x)dx + \sum_{j,k} p(x_j, k)r_k}.$$

We analyze the performance measures of interest through the MFQ model and the additive decomposition method. The model parameters are $\alpha = 2/3\frac{1}{s}$, $\beta = 1\frac{1}{s}$, $r = 12.2kbps$, $N = 25$, $c = 190kbps$ and $B = 30kb$. Without RED filtering ($s(x) = 1$) the drift is $d = 183kbps$, and with decreasing RED acceptance probability the drift is decreasing to $d = -c$ at $s(x) = 0$. We considered 2 kinds of piecewise constant functions for $s(x)$. The $(0, B)$ interval was divided into 3 and 6 identical subintervals. E.g., in the first case $x_1 = 10$, $x_2 = 20$, $x_3 = 30$ and vector (s_1, s_2, s_3) contains the acceptance probabilities for the intervals $(0, 10)$, $(10, 20)$, $(20, 30)$, respectively. Figure 3 depicts the fluid density functions for different

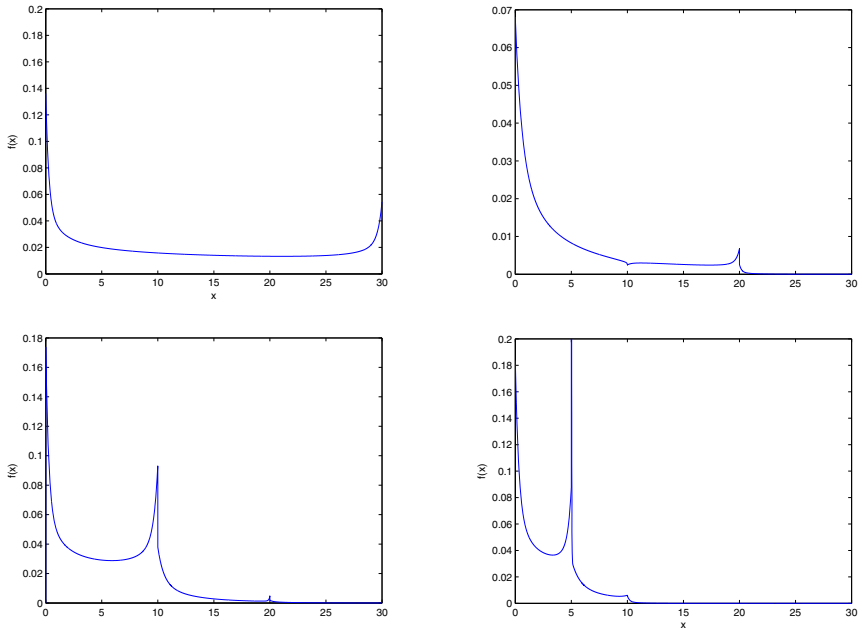


Fig. 3. Fluid density functions with $s(x) = 1$, $(s_1, s_2, s_3) = (0.905, 0.8041, 0.72)$, $(s_1, s_2, s_3) = (1, 0.8127, 0.76)$, $(s_1, s_2, s_3, s_4, s_5, s_6) = (0.908, 0.7936, 0.72, 0.69, 0.65, 0.54)$, respectively

$s(x)$ functions. In the first graph $s(x) = 1$, in the second graph $(s_1, s_2, s_3) = (0.905, 0.8041, 0.72)$, in the third graph $(s_1, s_2, s_3) = (1, 0.8127, 0.76)$, in the fourth graph $(s_1, s_2, s_3, s_4, s_5, s_6) = (0.908, 0.7936, 0.72, 0.69, 0.65, 0.54)$. The associated loss rations are 0.0121, 0.0995, 0.0492 and 0.100.

5 Conclusions

The problem of analyzing finite buffer MFQs in saturation has been considered recently in [10]. In that paper the numerical procedure by Soares and Latouche [7,8] was generalized for the case of saturation. In this paper we considered the additive decomposition procedure by Nail et al. [9,5] and generalized for the case of saturation.

The proposed modification seems to eliminate the numerical instabilities of the method for drift values close to zero and for moderate buffer sizes. The case of extremely large buffers still results in numerical problems, because in saturation a Jordan block of size 2×2 associated with the zero eigenvalue, which results in an exponentially increasing coefficient.

Acknowledgements. This work was supported by OTKA grant no. K-101150.

References

1. Ahn, S., Jeon, J., Ramaswami, V.: Steady state analysis of finite fluid flow models using finite qbds. Queueing Systems 49, 223–259 (2005)
2. Floyd, S., Jacobson, V.: Random early detection gateways for congestion avoidance. IEEE/ACM Transactions on Networking 1, 397–413 (1993)
3. Gribaudo, M., Telek, M.: Stationary analysis of fluid level dependent bounded fluid models. Performance Evaluation 65, 241–261 (2008)
4. Horton, G., Kulkarni, V., Nicol, D., Trivedi, K.: Fluid stochastic petri nets: Theory, applications, and solution techniques. European Journal of Operational Research 105, 184–201 (1998)
5. Kankaya, H., Akar, N.: Solving multi-regime feedback fluid queues. Stochastic Models 24(3), 425–450 (2008)
6. Kulkarni, V.G.: Fluid models for single buffer systems. In: Dshalalow, J.H. (ed.) Models and Applications in Science and Engineering. Frontiers in Queueing, pp. 321–338. CRC Press (1997)
7. da Silva Soares, A., Latouche, G.: Matrix-analytic methods for fluid queues with finite buffers. Perform. Eval. 63, 295–314 (2006)
8. da Silva Soares, A., Latouche, G.: Fluid queues with level dependent evolution. European Journal of Operational Research 196(3), 1041–1048 (2009)
9. Sohraby, K., Akar, N.: Infinite/finite buffer markov fluid queues: A unified analysis. Journal of Applied Probability 41(2), 557–569 (2004)
10. Telek, M., Vécsei, M.: Finite queues at the limit of saturation. In: 9th International Conference on Quantitative Evaluation of SysTems (QEST), pp. 33–42. Conference Publishing Services (CPS), London (2012)

Queue-Size Distribution in M/G/1-Type System with Bounded Capacity and Packet Dropping

Oleg Tikhonenko[1] and Wojciech M. Kempa[2]

[1] Czestochowa University of Technology, Institute of Mathematics
ul. Dabrowskiego 69, 42–201 Czestochowa, Poland
oleg.tikhonenko@gmail.com
[2] Silesian University of Technology, Institute of Mathematics
ul. Kaszubska 23, 44–100 Gliwice, Poland
wojciech.kempa@polsl.pl

Abstract. A single-server queueing system of $M/G/1$-type with bounded total volume is considered. It is assumed that volumes of arriving packets are generally distributed random variables. The AQM-type mechanism is used to control the actual buffer state: each of arriving packets is dropped with probability depending on its volume and the occupied volume of the system at the pre-arrival epoch. The explicit formulae for the stationary queue-size distribution and loss probability are found.

Keywords: AQM algorithms, finite buffer, loss probability, single-server queueing system, queue-size distribution.

1 Introduction

Wide applications of finite-buffer queueing systems in telecommunication, computer networks, management, transport and logistics are known. In particular, we can use them as mathematical models of the data packet traffic in the node (router) of the Internet network. A typical phenomenon which can be observed in the operation of the Internet network is the situation of buffer congestion causing losses of the arriving packets. It's clear that the enlarging the capacity of the buffer is not a good solution of this problem since it prolongs the sojourn time of the packet in the system. Another approach is based on using of Active Queue Management (AQM) algorithms. A basic algorithm, called Random Early Detection (RED) was proposed in [5]. In the RED scheme a drop function is defined that "controls" the input stream and rejects the arriving packet with probability dependent on the actual queue size at the pre-arrival epoch. Various types of drop functions were studied. In [2], [1, 10], [17] and [6] linear, exponential, quadratic and "gentle" linear drop functions were considered respectively. Some other results relating to the theoretical and practical aspects of using the AQM schemes can be found in [3, 4, 7, 8, 11–13, 16]. In [9] the $M/M/1/m$ system with single and batch arrivals, with the buffer "state" controlled by a drop function, was considered. The formulae for different stationary-state stochastic

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 177–186, 2013.

characteristics were derived there: the queue-size distribution, the number of packets (batches of packets) lost consecutively, and the time between two successive losses. In [14] the representation for the stationary queue-size distribution was obtained for the generalized $M/M/1/m$ system, in which the arriving packets have generally distributed volumes and the total their volume (capacity) in the system is bounded. The extension of results obtained in [14] can be found in [15] where the case of a multi-server system was investigated.

In the paper we generalize results from [14] for the case of a single-server system with Poisson arrivals and generally distributed service times. We replace the classical drop function by an "accepting" function that accepts the arriving packet with probability that depends on the actual occupied capacity of the system at the pre-arrival epoch, and on the volume of the arriving packet.

2 The Model and Auxiliary Results

Consider a single-server queueing system in which succesive packets arrive according to Poisson process with intensity a, and are characterized by their volumes which are generally distributed positive-valued random variables with a distribution function $L(x)$. Packets are served individually with a general-type service time distribution function $B(t)$, independently of their volumes. Sequences of successive interarrival and service times, and volumes of the arriving packets are supposed to be totally independent. The total volume of the system, i.e. the sum of volumes of all packets present in the system, is bounded by a non-random positive value (system capacity) V. We shall denote the system under consideration by $M/G/1/(\infty, V)$ (see e.g. [16, 17]). Let us note that the well-known "classical" $M/G/1/m$-type system, in fact, is a special case of the system described above, when $L(x) = 0$ for $x \leq 1$, $L(x) = 1$ for $x > 1$ and $V = m + 1$.

Let $\eta(t)$ be the number of packets present in the system at a fixed time t. Let $\xi^*(t)$ be the residual service time of the packet being in service at time t (if $\eta(t) > 0$).

Consider now the "classical" system $M/G/1/\infty$. Its evolution can be described by the following Markov process:

$$\big(\eta(t);\, \xi^*(t)\big). \tag{1}$$

For the system $M/G/1/(\infty, V)$ we need to supplement this process. Let $\zeta_i(t)$ be the volume of the ith packet present in the system at time t. Then $\sigma(t) = \sum_{i=1}^{\eta(t)} \zeta_i(t)$ is the "transient" volume of the system at time t, i.e. the sum of volumes of all packets present in the system at time instant t. Now the considered system can be described by the following Markov process:

$$\big(\eta(t);\, \zeta_i(t), i = \overline{1, \eta(t)}; \xi^*(t)\big). \tag{2}$$

Here we take the assumption that the arriving packets are numbered successively as they occur. Of course, if $\eta(t) = 0$ then also $\sigma(t) = 0$. Assume that there exists the stationary state of the system and introduce the following notations:

$$\lim_{t\to\infty} \sigma(t) = \sigma, \tag{3}$$

$$\lim_{t\to\infty} \xi^*(t) = \xi^*, \tag{4}$$

$$\lim_{t\to\infty} \zeta_i(t) = \zeta_i. \tag{5}$$

In the stationary state the stochastic process (1) can be characterized by the functions

$$w_k(y) = \mathbf{P}\{\eta = k;\, \xi^* < y\}, \tag{6}$$

where $k = 1, 2....$

Similarly, one can describe the stochastic process (2) using the following functions:

$$g_k(x, y)dx = \mathbf{P}\{\eta = k;\, \sigma \in [x, x + dx);\, \xi^* < y\}, \tag{7}$$

where k is defined as previously.

Let us note that for the process (2) we can write

$$w_k(y) = \int_0^V g_k(x, y)dx. \tag{8}$$

Define

$$P_k(t) = \mathbf{P}\{\eta(t) = k\}, \quad p_k = \mathbf{P}\{\eta = k\}, \tag{9}$$

where $k = 0, 1, ...$, and η stands for the number of packets present in the system in the stationary state.

In the stationary state we have evidently

$$p_k = \mathbf{P}\{\eta = k\} = w_k(\infty), \quad k = 1, 2, ... \tag{10}$$

Assume that $\beta_1 = \int_0^\infty tdB(t) < \infty$.

We end this section with supplementing necessary notations. Let us denote by p_{loss} the stationary loss probability i.e. the probability that the incoming packet is lost. Besides, let $\rho = a\beta_1$ be the traffic load of the system. Lastly, by $F_*^{(k)}(\cdot)$ we denote the k-fold Stieltjes convolution of any distribution function $F(\cdot)$ of non-negative random variable with itself i.e.

$$F_*^{(0)}(y) \equiv 1, \qquad F_*^{(k)}(y) = \int_0^y F_*^{(k-1)}(y - x)dF(x), \quad k = 1, 2, ... \tag{11}$$

3 Queue-Size Distribution in the Original System with Packet Dropping

Let us take into consideration the "classical" $M/G/1/\infty$-type queueing system without packet dropping. The stationary probabilities for this "clasical" system

can be obtained from the formula (10). Using the notations introduced in the previous section we can write the following system of differential equations for the unknown functions $w_k(y)$, where $k = 1, 2, ...$:

$$0 = -ap_0 + \frac{\partial w_1(y)}{\partial y}\Big|_{y=0}; \tag{12}$$

$$-\frac{\partial w_1(y)}{\partial y} + \frac{\partial w_1(y)}{\partial y}\Big|_{y=0} = ap_0 B(y) - aw_1(y) + \frac{\partial w_2(u)}{\partial u}\Big|_{u=0} B(y); \tag{13}$$

$$-\frac{\partial w_k(y)}{\partial y} + \frac{\partial w_k(y)}{\partial y}\Big|_{y=0} = aw_{k-1}(y) - aw_k(y) + \\ + \frac{\partial w_{k+1}(u)}{\partial u}\Big|_{u=0} B(y), \quad k = 2, 3, ... \tag{14}$$

For the system (12)–(14) the following boundary conditions hold true:

$$aw_k(y) = \frac{\partial w_{k+1}(u)}{\partial u}\Big|_{u=0}, \quad k = 1, 2, ... \tag{15}$$

Let us now take into consideration the original $M/G/1/(\infty, V)$-type queueing system, where the total volume of packets in the system is bounded by V. We introduce into the system the AQM algorithm defined as follows. Let $r(\cdot)$ be a right-hand continuous and nonincreasing function defined on the interval $[0, V]$, and satisfying conditions $r(0) \leq 1$ and $r(V) \geq 0$. If the volume of the arriving packet and the total volume of the system at the pre-arrival instant equal x and y respectively, then $r(x + y)$ is the probability that the arriving packet will be accepted for service. If the packet of the volume x, arriving at time t is dropped, we have $\eta(t) = \eta(t^-)$ and $\sigma(t) = \sigma(t^-)$. Similarly, in the case of acceptance the arriving packet we have $\eta(t) = \eta(t^-) + 1$ and $\sigma(t) = \sigma(t^-) + x$.

The system of Kolmogorov-type equations for the stationary state of the $M/G/1/(\infty, V)$ system with AQM takes the following form:

$$0 = -ap_0 \int_0^V r(v)dL(v) + \frac{\partial w_1(y)}{\partial y}\Big|_{y=0}; \tag{16}$$

$$-\frac{\partial w_1(y)}{\partial y} + \frac{\partial w_1(y)}{\partial y}\Big|_{y=0} = ap_0 B(y) \int_0^V r(v)dL(v) - \\ - a \int_0^V g_1(x, y) \int_0^{V-x} r(x + v)dL(v)dx + \frac{\partial w_2(u)}{\partial u}\Big|_{u=0} B(y); \tag{17}$$

$$-\frac{\partial w_k(y)}{\partial y} + \frac{\partial w_k(y)}{\partial y}\Big|_{y=0} = a \int_0^V g_{k-1}(x, y) \int_0^{V-x} r(x + v)dL(v)dx - \\ - a \int_0^V g_k(x, y) \int_0^{V-x} r(x + v)dL(v)dx + \frac{\partial w_{k+1}(u)}{\partial u}\Big|_{u=0} B(y), \quad k = 2, 3, ... \tag{18}$$

The boundary conditions are following:

$$a\int_0^V g_k(x,y)\int_0^{V-x} r(x+v)dL(v)dx = \frac{\partial w_{k+1}(u)}{\partial u}\Big|_{u=0} B(y), \quad k=1,2,.... \tag{19}$$

Let us introduce now the following notation:

$$R(z) = \int_0^z r(V-z+v)dL(v). \tag{20}$$

Putting (20) into the system (16)–(18) we get

$$0 = -ap_0R(V) + \frac{\partial w_1(y)}{\partial y}\Big|_{y=0}; \tag{21}$$

$$-\frac{\partial w_1(y)}{\partial y} + \frac{\partial w_1(y)}{\partial y}\Big|_{y=0} = ap_0B(y)R(V) -$$
$$-a\int_0^V g_1(x,y)R(V-x)dx + \frac{\partial w_2(u)}{\partial u}\Big|_{u=0} B(y); \tag{22}$$

$$-\frac{\partial w_k(y)}{\partial y} + \frac{\partial w_k(y)}{\partial y}\Big|_{y=0} = a\int_0^V g_{k-1}(x,y)R(V-x)dx -$$
$$-a\int_0^V g_k(x,y)R(V-x)dx + \frac{\partial w_{k+1}(u)}{\partial u}\Big|_{u=0} B(y), \quad k=2,3,.... \tag{23}$$

The boundary conditions (19) can be rewritten in the following form:

$$a\int_0^V g_k(x,y)R(V-x)dx = \frac{\partial w_{k+1}(u)}{\partial u}\Big|_{u=0} B(y), \quad k=1,2,.... \tag{24}$$

Systems (12)–(15) and (21)–(24) lead to the following theorem:

Theorem 1. *The stationary queue-size distribution p_k, $k=0,1,...$, in the $M/G/1/(\infty,V)$-type queueing system with packet dropping can be expressed as*

$$p_k = C\widehat{p}_k R_*^{(k)}(V), \quad k=0,1,..., \tag{25}$$

where

$$C = \left[\sum_{k=0}^{\infty} \widehat{p}_k R_*^{(k)}(V)\right]^{-1}, \tag{26}$$

$\widehat{p}_k$, $k=0,1,...$, are stationary probabilities in the "classical" $M/G/1/\infty$ system, and $R_*^{(k)}(\cdot)$ is the $k-$fold Stieltjes convolution of the function $R(\cdot)$ defined in (20) with itself.

Moreover, the loss probability p_{loss} is given by the formula

$$p_{\text{loss}} = 1 - \frac{1-p_0}{\rho}, \tag{27}$$

where $\rho = a\beta_1$, a is the arrival intensity, and β_1 is the first moment of the service time.

Proof. Assume that the number $\widehat{p}_0$ and functions $\widehat{w}_k(y)$, $k = 1, 2, ...$, satisfy the system of equations (12)–(15) for the "classical" $M/G/1/\infty$-type system, and besides the normalization condition

$$\widehat{p}_0 + \sum_{k=1}^{\infty} \widehat{w}_k(\infty_k) = 1. \tag{28}$$

Let C be a constant which will be found explicitly later, and denote by $R_*^{(k)}(\cdot)$ a $k-$fold Stieltjes convolution of the function $R(\cdot)$.

By a direct substitution, it is easy to verify that the number $p_0 = C\widehat{p}_0$ and the functions $g_k(x, y)$, such that

$$g_k(x, y)dx = C\widehat{w}_k(y)dR_*^{(k)}(x), \tag{29}$$

satisfy the system of equations (21)–(24).

Thus, if $\widehat{p}_k$ is the stationary probability that exactly k packets are present in the "classical" model, then - for the $M/G/1/(\infty, V)$-type system with packet dropping - the correspondent probability p_k can be found as $p_k = C\widehat{p}_k R^{k*}(V)$.

The normalization condition gives now

$$C = \left[\sum_{k=0}^{\infty} \widehat{p}_k R_*^{(k)}(V)\right]^{-1}.$$

that ends the proof of (25) and (26).

The formula (27) is a consequence of the stability condition. □

It is clear that in general case the formulae (25) are not convenient for calculations because of $\widehat{p}_k$ precise calculation impossibility and of Stieltjes convolutions presence. But we can calculate the probabilities p_k and p_{loss} for some special forms of the functions $B(\cdot)$ and $R(\cdot)$.

4 Some Special Cases

A. "Classical" AQM. Consider the classical system $M/G/1/m < \infty$ with drop function d_i, having the sense of probability that the arriving packet will be rejected, if there are i other packets in the system at the arriving epoch, $i = \overline{0, m}$.

It follows from the equations (25) and (26) that in this case we have for the probabilities p_k, that form the stationary queue-size distribution of the system,, the following representations:

$$p_0 = C\widehat{p}_0,\ p_k = C\widehat{p}_k \prod_{i=0}^{k-1}(1 - d_i),\ k = \overline{1, m+1},$$

where $\widehat{p}_k$ are the stationary probabilities that there are k packets present in the "classical" $M/G/1/\infty$ system, $k = \overline{0, m+1}$,

$$C = \left[\widehat{p}_0 + \sum_{k=1}^{m+1} \widehat{p}_k \prod_{i=0}^{k-1} (1 - d_i) \right]^{-1}.$$

In this case precise calculation of the probabilities p_k are evidently possible for service time having an exponential distribution (see [9]) or being a constant value. More widely, they are possible, if an algorithm for calculation of $\widehat{p}_k$ is known.

B. Constant Service Time. For service time $\xi \equiv t_0 = const$ (i.e. for the system $M/D/1/(\infty, V)$) we evidently obtain

$$p_1 = p_0(e^{\rho} - 1)R(V),$$

$$p_k = p_0 \left\{ e^{k\rho} + \sum_{i=1}^{k-1} (-1)^{k-i} e^{i\rho} \left[\frac{(i\rho)^{k-1}}{(k-i)!} + \frac{(i\rho)^{k-i-1}}{(k-i-1)!} \right] \right\} R_*^{(k)}(V), \ k \geq 2,$$

where

$$p_0 = \left\{ 1 + (e^{\rho} - 1)R(V) + \right.$$

$$\left. + \sum_{k=2}^{\infty} \left[e^{k\rho} + \sum_{i=1}^{k-1} (-1)^{k-i} e^{i\rho} \left(\frac{(i\rho)^{k-1}}{(k-i)!} + \frac{(i\rho)^{k-i-1}}{(k-i-1)!} \right) \right] R_*^{(k)}(V) \right\}^{-1}, \ \rho = at_0.$$

C. Exponentially Distributed Service Time. If $B(t) = 1 - e^{-\mu t}$, $\mu > 0$ (i.e. for the system $M/M/1/(\infty, V)$), we have evidently

$$p_k = p_0 \rho^k R_*^{(k)}(V), \ k = 1, 2, ...,$$

where

$$p_0 = C(1 - \rho) = \left[\sum_{k=0}^{\infty} \rho^k R_*^{(k)}(V) \right]^{-1}, \ \rho = a/\mu.$$

In Tab. 1 we present stationary probabilities p_k, for $k = 0, 1, ..., 16$, in the system in which $V = 10$ and volumes of packets are exponentially distributed with mean 2. The "accepting" function is defined as

$$r(x) = \frac{(x - 10)^2}{100}, \quad 0 \leq x \leq 10.$$

In computations we compare the cases of traffic loads $\rho = 1$ and $\rho = 4$. Results are presented geometrically in Fig. 1 (the case $\rho = 4$ in dark colour).

The values of loss probability for $\rho = 1$ and $\rho = 4$ equal 0.58391009 and 0.89597752 respectively.

Table 1. Stationary probabilities for exponential packet volume distribution, exponential service time, and $\rho = 1$ and $\rho = 4$

Queue size k	p_k for $\rho = 1$	p_k for $\rho = 4$
0	0.58391009	0.22623632
1	0.39020278	0.60473721
2	0.02544025	0.15770944
3	0.00044376	0.01100387
4	0.00000311	0.00030882
5	$1.08639396 \times 10^{-8}$	$4.31026173 \times 10^{-6}$
6	$2.40300481 \times 10^{-11}$	$3.81356307 \times 10^{-8}$
7	$1.34619664 \times 10^{-13}$	$8.54564379 \times 10^{-10}$
8	$2.50094105 \times 10^{-15}$	$6.35038026 \times 10^{-11}$
9	$3.70268432 \times 10^{-17}$	$3.76073692 \times 10^{-12}$
10	$3.90566084 \times 10^{-19}$	$1.58675833 \times 10^{-13}$
11	$3.06682151 \times 10^{-21}$	$4.98384757 \times 10^{-15}$
12	$1.88406302 \times 10^{-23}$	$1.22470550 \times 10^{-16}$
13	$1.00188522 \times 10^{-25}$	$2.60503884 \times 10^{-18}$
14	$5.91427732 \times 10^{-28}$	$6.15117253 \times 10^{-20}$
15	$4.91545009 \times 10^{-30}$	$2.04493499 \times 10^{-21}$
16	$4.87042159 \times 10^{-32}$	$8.10480858 \times 10^{-23}$

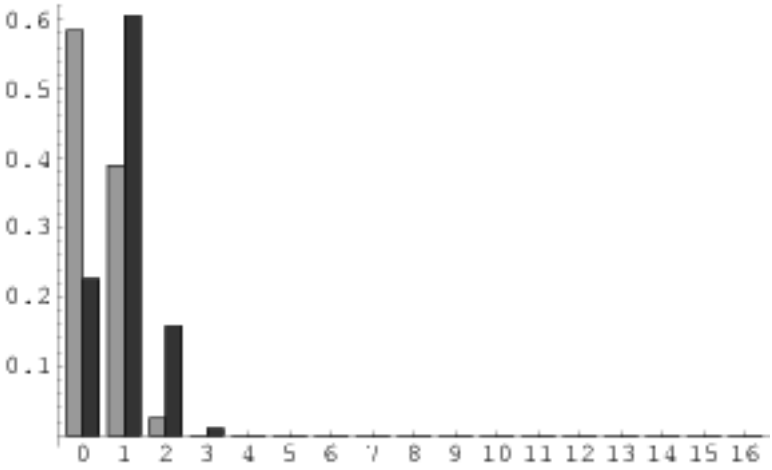


Fig. 1. Stationary probabilities for exponential packet volume distribution, exponential service time, and $\rho = 1$ and $\rho = 4$

In Tab. 2 we state stationary probabilities p_k for the system with $V = 10$ and volumes of packets having 2-Erlang distributions with parameter $\alpha = 1$ (so, with mean 2). The "accepting" function is defined as above. As previously, we compare the cases of $\rho = 1$ and $\rho = 4$. Results are presented in Fig. 2 (in dark colour the case $\rho = 4$).

Now the values of loss probability for $\rho = 1$ and $\rho = 4$ equal 0.53209755 and 0.79451408 respectively.

Table 2. Stationary probabilities for exponential packet volume distribution, 2-Erlang service time, and $\rho = 1$ and $\rho = 4$

Queue size k	p_k for $\rho = 1$	p_k for $\rho = 4$
0	0.53209755	0.17805630
1	0.42443864	0.56812119
2	0.04221294	0.22601210
3	0.00123538	0.02645745
4	0.00001539	0.00131870
5	$9.84474610 \times 10^{-8}$	$3.37342000 \times 10^{-5}$
6	$3.73550067 \times 10^{-10}$	$5.12005771 \times 10^{-7}$
7	$1.62007964 \times 10^{-12}$	$8.88223771 \times 10^{-9}$
8	$3.33422929 \times 10^{-14}$	$7.31208922 \times 10^{-10}$
9	$8.51994577 \times 10^{-16}$	$7.47382358 \times 10^{-11}$
10	$1.59676709 \times 10^{-17}$	$5.60283168 \times 10^{-12}$
11	$2.23845424 \times 10^{-19}$	$3.14176874 \times 10^{-13}$
12	$2.43726606 \times 10^{-21}$	$1.36832394 \times 10^{-14}$
13	$2.15496080 \times 10^{-23}$	$4.83933127 \times 10^{-16}$
14	$1.71000077 \times 10^{-25}$	$1.53603911 \times 10^{-17}$
15	$1.53884029 \times 10^{-27}$	$5.52916445 \times 10^{-19}$
16	$1.91507107 \times 10^{-29}$	$2.75239554 \times 10^{-20}$

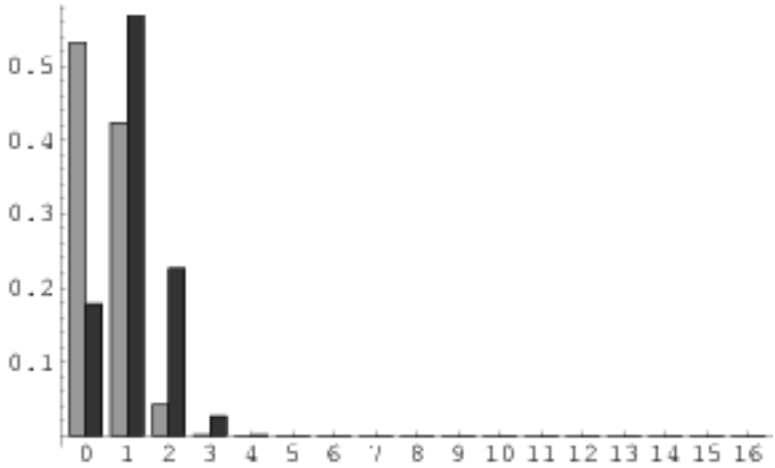


Fig. 2. Stationary probabilities for exponential packet volume distribution, 2-Erlang service time, and $\rho = 1$ and $\rho = 4$

5 Conclusion

In the paper stationary queue-size probabilities and loss probability for the system with bounded capacity $M/G/1/(\infty, V)$ and dropping packets mechanism are derived, under conditions that packet volume and service time are independent, and probability of dropping depends on the volume of the arriving packet and the total volume of packets present in the system at the pre-arrival epoch.

References

1. Athuralya, S., Low, S.H., Qinghe, Y.: REM: active queue management. IEEE Network 15(3), 48–53 (2001)
2. Bonald, T., May, M., Bolot, J.C.: Analytic evaluation of RED perfomance. In: Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 3, pp. 48–53 (2000)
3. Chydzinski, A., Chrost, L.: Analysis of AQM queues with queue size based packet dropping. International Journal of Applied Mathematics and Computer Science 21(3), 567–577 (2011)
4. Chydzinski, A.: Optimization problems in the theory of queues with dropping functions. In: HET-NETs, pp.121–132 (2011)
5. Floyd, S., Jacobson, V.: Random early detection gateways for congestion avoidance. IEEE ACM Telecommunication Network 1(4), 397–412 (1993)
6. Floyd, S.: Recomendations on using the gentle variant of RED (March 2000), http://www.aciri.org/floyd/red/gentle.html
7. Floyd, S.: Adaptive RED: an algoritm for increasing the robustness of RED's Active Queue Management (August 2001), http://www.aciri.org/floyd/papers/adaptiveRed.pdf
8. Hao, W., Wei, Y.: An Extended $GI^X/M/1/N$ Queueing Model for Evaluating the Performance of AQM Algorithms with Aggregate Traffic. In: Lu, X., Zhao, W. (eds.) ICCNMC 2005. LNCS, vol. 3619, pp. 395–404. Springer, Heidelberg (2005)
9. Kempa, V.M.: On main characteristics of the $M/M/1/N$ queue with single and batch arrivals and the queue size controlled by AQM algorithms. Kybernetika 47(6), 930–943 (2011)
10. Liu, S., Basar, T., Srikant, R.: Exponential RED: A stabilizing AQM scheme for low- and light-speed TCP protocols. IEEE/ACM ToN (2005)
11. Rosolen, V., Bonaventure, O., Leduc, G.: A RED discard strategy for ATM networks and its performance evaluation with TCP/IP traffic. Computer Communication Rewiew 29(3), 23–43 (1999)
12. Sun, L., Wang, L.: A novel RED scheme with preferential dynamic threshold deployment. In: Computational Intelligence and Security Workshops, pp. 854–857 (2007)
13. Suresh, S., Gol, O.: Congestion management of self similar IP traffic-application of the RED scheme. In: Wireless and Optical Communications Networks, Second IFIP International Conference, pp. 372–376 (2005)
14. Tikhonenko, O., Kempa, W.M.: The Generalization of AQM Algorithms for Queueing Systems with Bounded Capacity. In: Wyrzykowski, R., Dongarra, J., Karczewski, K., Waśniewski, J. (eds.) PPAM 2011, Part II. LNCS, vol. 7204, pp. 242–251. Springer, Heidelberg (2012)
15. Tikhonenko, O., Kempa, W.M.: On the queue-size distribution in the multi-server system with bounded capacity and packet dropping. Kybernetika (submitted)
16. Xiong, N., Yang, Y., Defago, X., He, Y.: LRC-RED: A self-tuning robust and adaptive AQM scheme. In: Sixth International Conference on Parallel and Distributed Computing Applications and Technologies, pp. 655–659 (2005)
17. Zhou, K., Jeung, K.L., Li, V.O.K.: Nonlinear RED: A simple yet efficient active queue management scheme. Computer Networks (18), 3784–3794 (2006)

Use of Time-Scale for Analysis of Data Source Traffic

Ivan Titov[1], Ivan Tsitovich[2], and Stoyan Poryazov[3]

[1] Moscow Technical University of Communications and Informatics, Moscow, Russia
[2] National Research University Higher School of Economics, Moscow, Russia
[3] Institute for Mathematics and Informatics, Sofia, Bulgaria
cito@iitp.ru

Abstract. In this paper, we consider the model of server traffic when the traffic is separated into several streams. The amount of transferred data differs for different streams. Based on real traffic measurements we proposed the server traffic model where traffic of every stream is described by the same independent processes, but each process has its own time scale. We show that for traffic analysis as well as for developing of the most effective methods of control of this traffic, it is necessary to correctly identify the time scale for each stream, as well as the time scale of traffic fluctuations those have a significant effect to QoS.

Keywords: communication system traffic, mathematical model, time scale, self-similar, Poisson arrival process.

1 Introduction

Recently a number of high-quality, high-resolution measurements of multimedia traffic in high-speed networks were carried out and analyzed. It is shown that the traffic in such networks is self-similar [1]. But mathematical analysis of models based on the self-similar processes is very difficult. On the other hand, the traditional traffic models such as the Poisson arrival process, the Markovian arrival process, etc., are well studied but do not give sufficient accuracy for a modern network traffic description, including a long-time dependence.

The modern networks traffic is described as a multi-stream traffic. The traffic is partitioned usually onto steams under the QoS conditions for different types of traffic: real-time audio-, video traffic, data, etc. [2]. In contrast, in this paper we analyze one type of traffic with respect to using a network resource by different types of requests, classified in [3]. It was found that some resources generate a family of streams which differ by a service time [4], [5]. Therefore, the total traffic can be described as linear combination of the flat-rate traffic with different time-scales. In contrast with the self-similar traffic, this approach gives possibilities to research processes with a long-time dependence using the classical teletraffic models if the components are describing by the classical models.

For traffic shaping, it is necessary to single out the significant stream for QoS characteristics calculating. When the significant stream is singled out then its

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 187–197, 2013.

influence onto the QoS characteristics may be investigated by using the classical teletraffic models.

Traffic of two real sources was examined: a multimedia resource and a musical portal [4], [5]. The multi-media resource provides access for users to files of various types. Most requests come for transfer of small files (HTML pages with images) when users search files and additional information. The number of user requests to transfer a specific mp3 file is substantially smaller, but the processing of such requests requires an essentially more time. The maximum service time is typical for requests for the transfer of archive files containing music albums and video files. The total traffic of the resource is self-similar (the estimation of the Hurst parameter is 0.88). It may be divided onto four streams with substantially different times of services (see Table below). We investigate properties of this streams and show that the traffic of every of them may be considered as a Poisson processes. Thus, we can say that the server generated several types of streams with different intensity and size of the requested files. Its time of service (or size of sent file) gives us the time-scale of this steam or its time unit. Analogous result was get for a musical portal [4].

In this paper, we propose the mathematical model of traffic based on the classical Poisson arrival processes but in case when the processes have different time-scales. This approach gives us possibility to use classical and more complicated theory (for example, [6], [7]). By the way, this approach gives us possibility to use the algorithms to estimate variations of the model calculated parameters under uncertainty in the model input parameters [8], [9].

In Section 2, we introduce the mathematical model of traffic generated by a source with an "unbounded" service time variation where every component of the total traffic has own time-scale unit. Based on this model, we discuss how to choose that component of the traffic which gives us the main influence on the interesting output parameter of the model. The choice is based on the time scale unit. In this case, the streams with more time service assume as non-random processes and the ones with less time service assume as processes with rapid random fluctuations. This unit gives us possibility to choose of the significant stream; this steam has the same time scale. Therefore, the time unit of the significant stream is the major characteristic for the total traffic model describing. In Section 3, we give an example of a server which generates four streams such that the average volume of a request differs from 7 up 20 times. We estimate auto-correlation functions (ACFs) of processes of the first steam input requests and total input requests and ACFs of the first steam transmitted data and the total transmitted data. In Section 4, we outline numerical simulation results of our model and some traffic control models. In this case, traffic shaping consists in regulation of a data transfer rate for different streams.

2 Mathematical Model

For simplicity of researches, we assume that the source of the load generates n Poisson streams of requests with intensities λ_i, $\Lambda = \sum_{i=1}^{n} \lambda_i$ is the summary

intensity of incoming requests, $p_i = \lambda_i/\Lambda$ is the probability that the request arrived from the i-th stream. A request from the i-th stream is serviced during a time units $\bar{a}_i$. Therefore, the average service time is $A = \sum_{i=1}^{n} p_i \bar{a}_i$. As a characteristic of load source's randomness it is considered the standard deviation of the service time $\sigma = \sqrt{D}$ where $D = \sum_{i=1}^{n} p_i \bar{a}_i^2 - A^2$.

We consider the situation when $\sigma/A \gg 1$. The mathematical approach of this condition consists of a model with an infinite variance of the service time.

For this model, the traffic generated by the data server can be portioned onto n independent stationary streams of requests and every stream generates the process of data level with the average $\lambda_i \bar{a}_i$ and the random component $Y_t^{(i)}$. The processes $Y_t^{(i)}$ for all i have the same structure and differ only by the time scale $\bar{a}_i$ and the variation α_i. Let X_t be a stationary stochastic process which describes the random component of the data level of a steam with servicing time value equals to 1 and the variation value of X_1 equals to 1. The deviation from the average of the total load generated by n streams can be found as

$$Y_t = \alpha_1 \cdot X^{(1)}_{t/\overline{a}_1} + \alpha_2 \cdot X^{(2)}_{t/\overline{a}_2} + ... + \alpha_n \cdot X^{(n)}_{t/\overline{a}_n} \tag{1}$$

where $X_t^{(i)}$ be independent copies of the process X_t. All the copies are considered as independent processes and α_i determine the proportion of the corresponding component in the random component of the total load.

We may suppose also that the service time of a request from the i-th stream is serviced during a random time a_i (independent of servicing times of others requests) with the mean value $\bar{a}_i$ and the variation σ_i and the values $\sigma_i/\bar{a}_i$ have the same order for all i.

It should be noted that server's traffic passes through the switch or router, which has a buffer on the output interface to compensate the fluctuations traffic and reduce losses. The buffer overflow probability is the main characteristic of QoS [10], [11]. Therefore, we are interested to know a maximum deviation for the time period $[0, T]$ where T is a time comparable with the router's buffer fill time and may give us the time-scale unit of the total traffic.

We consider the ratio of deviation from the average amount of data received from the server during the time period T:

$$\bar{Y} = \frac{1}{T}\int_0^T Y_t dt = \sum_{i=1}^{n} \frac{\alpha_i \bar{a}_i}{T} \int_0^{T/\bar{a}_i} X_t^{(i)} dt. \tag{2}$$

Since the streams are independent and $E\overline{Y} = 0$ the variance of the overall process is calculated as a sum of the variances [12]:

$$D\left[\overline{Y}\right] = \sum_{i=1}^{n} \left[\alpha_i^2 \left(1 - \frac{T}{3\bar{a}_i}\right) I\left(T \le \bar{a}_i\right) + \frac{\alpha_i^2 \bar{a}_i}{T} \left(1 - \frac{\bar{a}_i}{3T}\right) I\left(T > \bar{a}_i\right) \right] \tag{3}$$

where $I(B)$ is the indicator function of the event.

It follows from (3) that for $T \gg \bar{a}_i$ the variance of random deviations of the data generated by the i-th stream decreases in proportion to T, but for $T \ll \bar{a}_i$ the variance decreases with significantly lower rate.

It is shown on Fig. 1 the dependence of the variance $D\left[\overline{Y}\right]$ on T (the line D) with the following values of parameters n=3, $\bar{a}_1 = 1$, $\bar{a}_2 = 10$, $\bar{a}_3 = 100$, and $\alpha_1 = \alpha_2 = \alpha_3 = \sqrt{10/3}$ (named Example 1). For comparison, the plot $D_m(T)$ of dependence on the variance on T for the server with one Poisson stream intensity Λ, and constant service time $A = 10/3$ (named Example 2) is also shown on Fig. 1 (left), i.e. it is the case when we do not consider the structure of the source load and use its average characteristics only.

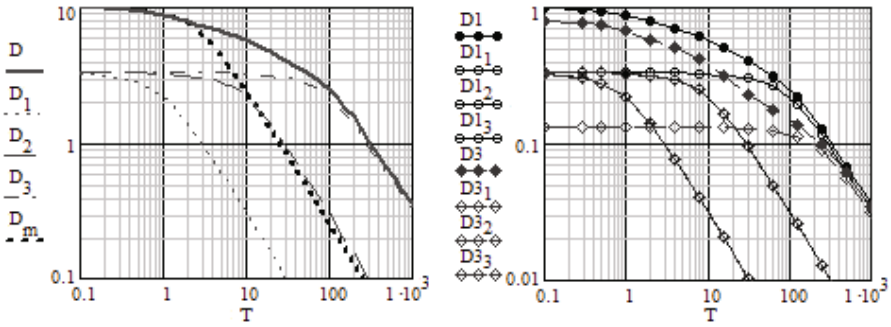


Fig. 1. Diagrams of the traffic variance with three and one stream (left) and diagrams of the variances for the 1-st and 3-rd systems from Section 4 (right)

It is shown that for $T < A$ the variance is approximately the same for both systems. Therefore, for systems without buffer the probability of loss will be the same. However, the traffic variance in the second example starts decreasing with a high speed at significantly lower values of T. For example, when $T = 100$ the variances $D(T)$ is 10 times greater than the variance $D_m(T)$. The dependence of the variance on T for each stream separately $D_i(T)$ also is shown on Fig. 1 (left). It follows from this dependence and the formula (3) that for small values of T the variances for all streams are the same. A linear decreasing of the variance of the i-th stream starts at values $\bar{a}_i$.

Therefore, if the average time of filling the buffer is more than A then it should take into account the structure of the source of the load, otherwise the assumptions about the buffer overflow probability will be too optimistic. For example, for the 1-st stream buffer with a specific size can be enough to provide a low level of data loss. Packets often enter in the buffer, but do not stay there for a long time because the system quickly changes its states; therefore, the amount of resources invested in the buffer reaches the maximum value rarely. At the same time, for the n-th stream this buffer may not be sufficient because the n-th stream has a different time scale. Therefore, the ratio between the time of filling the buffer and the stream time scale gives us possibility to understand how this stream may influence the probability of data loss. For example, for small values of T, all streams will make the same contribution into the buffer overflow probability. In contrast, if $T > \bar{a}_n$ then the probability of data loss will

be determined only by n-th stream. Therefore, the efficiency of control methods applied to different streams depends on T and the steam time scale.

3 Analysis of Web Server Traffic Properties

In this section, we analyze traffic coming from the real multimedia resource. This resource provides access for users to files of various types. Majority of requests consists in a transfer of a small file (HTML pages with images). The number of user's requests to transfer a specific mp3 file is substantially smaller, but the processing of such requests requires an essentially large time. The maximum service time is typical for requests for the transfer of archive files with music albums and video files. Thus, we can say that the server receives several types of streams of user requests with different intensity and size of the requested files.

Traces were collected at the edge of data-center where servers of this multimedia resource are located. Therefore we captured all traffic from these Web servers. We analyzed packet headers of network and transport layer. Selection criterion was a pool of IP addresses belonging to multimedia resource (primary and non-primary servers) and TCP source port 80 (HTTP). Traffic was analyzed from 10:00 to 13:30 on 23th of October 2010 (Saturday). During the observation period it was recorded 383000 TCP sessions and was transferred over 41 GB of data [5].

For analyzing the traffic, we recorded the times of start and finish of each TCP session. The histogram for time intervals between times of opening two consecutive TCP sessions, i.e. time intervals between two consecutive events of receiving user's request, is shown on Fig. 2 (left). It can be clearly seen that the probability distribution decreases exponentially (we use the logarithmic scale for x and the plot has a linear form). In addition, the mean value 39.4 ms and the variance 40.6 ms are almost equal. Such equality is typical for exponentially distributed random variables.

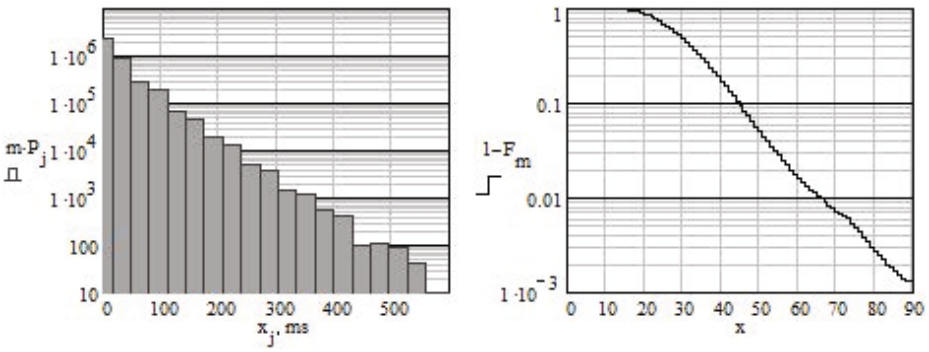


Fig. 2. Histogram of time intervals between user's requests (left) and $1 - F_m(x)$ (right)

For the long-time dependence investigating we considered another characteristic of incoming stream — the probability of receiving m requests for the time period t. Let $F_m(x)$ be the empirical distribution function of receiving m requests per 1 s. The dependence $1 - F_m(x)$ on x in a logarithmic scale is shown on Fig. 2 (right). The tail of this distribution also decreases exponentially.

In addition, we consider the sample estimate of the incoming stream ACF. ACF of number of requests received for the time period 1 s is shown on Fig. 3 (left). From this plot we can see that incoming stream does not possess a long-time dependence. The same results are obtained for time periods 10 and 100 s.

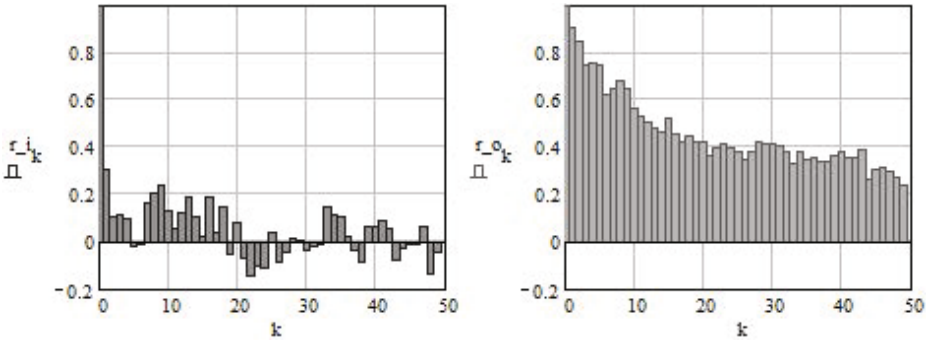


Fig. 3. ACF of the number of requests received in 1 s (left) and ACF of amount of data generated by the server in 1 s (right)

Thus, the arrival process of user's requests can be described as a Poisson process. Poisson processes are well studied and widely used as models of real streams. These results correspond to the classical view of the teletraffic theory on the structure of incoming streams of requests from a large number of independent sources.

For study the distribution of the amount of transmitted data, we estimate the amount data coming from the server to the client within a single TCP session.

The histogram for an amount of transferred data from Web server in a double logarithmic scale is shown on Fig. 4 (left). An amount of transmitted data in bytes is represented on the abscissa and the frequency of the corresponding 100-byte interval is represented on the ordinate.

From these plot it can be clearly seen that the probability distribution decreases nonmonotonically. For different values of the amount of transmitted data observed local maxima corresponding to the transfer of a large number of similar sized objects.

For a more detailed, we study the properties of the tail of the distribution. The dependence $1 - F_n(x)$ on x in a double logarithmic scale is shown on Fig. 4 (right).

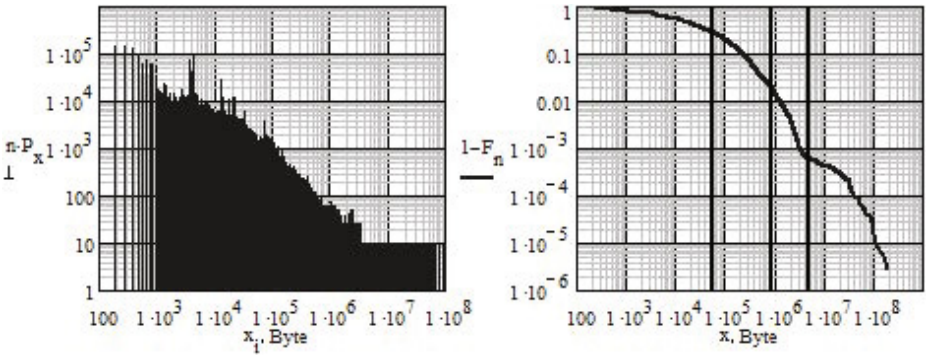


Fig. 4. The histogram of the amount of transmitted data (left) and $1 - F_n(x)$ as a function on the amount of transmitted data (right)

From this dependence it can be clearly seen that typical periods of slow decreasing of $1 - F_n(x)$ (i.e., at this period it was recorded a little number of sessions of the appropriate size) alternate with periods of rapid speed decreasing of $1 - F_n(x)$ (i.e. at this period it was observed a local maximum in the histogram). Thus, we can separate different streams of requests by depending on the rate of change of the empirical distribution function, as well as based on the character of changes in the histogram. Also there are shown on Fig. 4 (right) the three vertical lines corresponding to values of the volume of transferred data: $6 \cdot 10^4$, $9 \cdot 10^5$ and $5 \cdot 10^6$ bytes. These lines separate different streams of user's requests.

The main characteristics of data streams generated by the server during the observation time T_s are presented in Table: the number of sessions $\lambda \cdot T_s$, the average amount of transferred data within a single TCP session $\overline{V}$, and the total amount of information transmitted for each stream $\lambda \cdot T_s \cdot \overline{V}$.

Table 1. Characteristics of the data streams

No.	1	2	3	4
$\lambda \cdot T_s$	279490	96222	7019	229
$\overline{V}$, KB	11.09	213.2	1604	28812
$\lambda \cdot T_s \cdot \overline{V}$, GB	3.1	20.51	11.26	6.6

The first stream corresponds to the downloading of HTML pages containing images into JPEG and GIF formats with different sizes and scripts in Flash and JavaScript. The largest number of TCP sessions (73%) opens to download these types of files, but the amount of traffic generated by this stream is only 7.5% of the total traffic. The second stream consists of 30 seconds fragments of music and video files (preview). This stream generates almost half of total traffic. The third stream corresponds to the downloading of mp3-files and books in formats TXT, DOC and PDF. The fourth stream consists of video files and archived musical albums. The number of user requests for transfer files of third

and fourth types is only 1.9% of the total number but these streams create 43% of the total traffic.

The ACF of the total amount of data generated by the server in 1 s is shown on Fig. 3 (right). It is clearly seen that the ACF does not vanish during a longe time. Since, each of the streams corresponds to its time scale, the ACF of the amount of data from the specific stream needs significantly different time for decreasing to zero. For example, the ACF for the amount of data from the 1-st stream generated by the server in 1 s is shown on Fig. 5 (right). We can see that the ACF for the 1-st stream decreases rapidly. Results of research show that the ACF for other streams need much more time to vanish, and this time depends on the time scale corresponded to the stream.

In additional, we analyzed the ACF of numbers of requests received from a special stream. For example, ACF of number of requests from the 1-st stream received in 1 s is shown in Fig. 5 (left). For this stream (also as for other 3 streams) the ACF vanishes rapidly and it does not depend on a time period in which we measure a number of received requests.

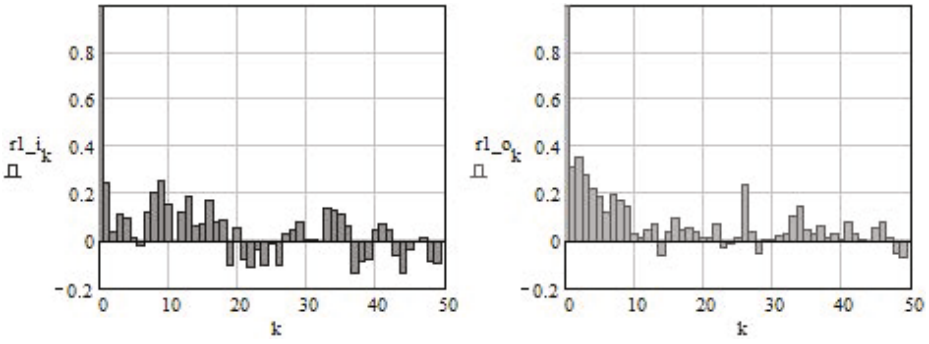


Fig. 5. ACF of the number of requests of the 1-st stream (left) and ACF of the amount of data of the 1-st stream generated by the server (right)

Thus, the analysis of the real Web server's traffic shows that this traffic can be divided into four Poisson streams with average times of service differ by more than tenfold. Therefore, each of the streams will correspond to its time scale that differs significantly for different streams.

4 Numerical Results

In this section, we outline results the numerical simulation of the Web server's data load when a traffic control is applied for different streams. For the numerical studies we use the mathematical model from Section 2 with the parameters $n = 3$, $\bar{a}_1 = 1$, $\bar{a}_2 = 10$, $\bar{a}_3 = 100$, and $\alpha_1 = \alpha_2 = \alpha_3 = \sqrt{1/3}$. All requests obtained the same rate C_m, consequently, the time of service is determined only by the size of requested file.

The traffic control means that we decrease rate of the data transfer in d times for one or several streams. Four systems with rate control were considered. The model without control is named the first system. The second system is one when the data transfer rate decreases to C_m/d for all requests; the 3-rd one is with the rate decreasing for requests with the maximum file size (the 3-rd stream) only; in the 4-th one the same control applies to requests of the 2-nd and 3-rd streams; and the 5-th one the control applies to requests with the minimum file size (the 1-st stream) only. For the 3-rd, 4-th and 5-th systems, the service time of requests is chosen in such a way that the average service time for all requests is the same for all four systems and, therefore, equals to $A \cdot d$.

The plots of the variance as a function on T for the total traffic of the 1-th system $D1$ as well as for each of streams $D1_i$ shown on Fig. 1 (right). In addition, analogous dependencies for the 3-rd system for $d = 1.5$ is shown in this figure ($D3$ and $D3_i$).

The comparison left and right parts of Fig. 1 shows that the changes of the variance for different streams obtained based on the analysis of the results of modeling of server traffic and obtained in the analytical form in Chapter 2 have the same character. As it is shown on Fig. 1 (right), for both systems the variance for the 1-st and the 2-nd streams changes identically. However, for the 3-rd system with the rate control, the variance of the 3-th stream becomes significantly lower. Consequently, a speed of decreasing of the total variance begins to increase for $T > 10$.

It is shown on Fig. 6 (left) the variance D_k (k is a number of the system) for 5 systems for $d = 1.5$.

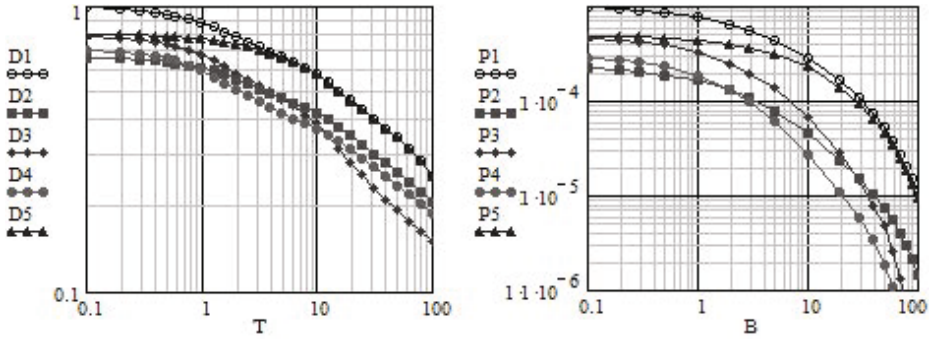


Fig. 6. Diagrams of the total variance for 5 systems (left) and of the loss probability as a function on the buffer size for $d = 1.5$ (right)

As it is followed from these plots, for different methods of the rate control correspond different ranges of T where the method gives the minimum values of the variance by k. Therefore, the effectiveness of a method of the rate control is differ in depending on the buffer size and on the bandwidth. For example, it is shown on Fig. 6 (right) the dependence of the loss probability (P_k, k is the

number of the system) on buffer size when the bandwidth $C = 120C_m$, $d = 1.5$, $\lambda_i = \lambda_1 \cdot 10^{(i-1)}$, $\lambda_1 = 33.3$. For these parameters each stream generates the same average load of data and the average total load of data is $100C_m$.

It is shown on these plots that the efficiency of the rate control applied only to requests for transferring files of maximum size (system 3) increases with increasing the buffer size. If the size of router's buffer is middle or big then it is not effective to decrease the rate for requests from the 1-st stream (system 5) [13]. The method of the rate control of the 4-th system is most effective and can significantly reduce the buffer overflow probability.

For example, for buffer size $B = 0.1C_m$, each stream makes equal contribution to the packets loss, therefore decreasing of the data transfer rate for all requests (system 5) is the most effective method (it decreases the loss probability in 4.5 times). However, if the router time scale is larger then an effectiveness of the rate control methods may be different. For example, for buffer size $B = 20C_m$, applying of the rate control to requests from the 2-nd and the 3-rd streams decreases the buffer overflow probability in 15 times, but applying of the rate control to requests from the 1-st stream decreases the buffer overflow probability only on 16%.

In addition, in [14] we analyzed effectiveness of rate control methods for system with distribution function of the amount of transmitted data corresponding to empirical distribution function of Web server considered in Chapter 3.

5 Conclusions

The data server traffic, which grants to users an access to files of different types, can be split into streams depending on the volume of a requested data. Analysis of the real Web server traffic shows that data volumes of requests of different streams may be substantially differed and the total traffic of Web server may be self-similar. But an appropriate splitting into streams based onto their time scales gives us possibility to use Poisson models or similar for QoS estimating or the Web server total traffic shaping.

References

1. Leland, W., Taqqu, M., Willinger, W., Wilson, D.: On the self-similar nature of ethernet traffic. IEEE/ACM Transactions on Networking 2(1), 1–15 (1994)
2. Cerqueira, E., Zeadally, S., Leszczuk, M., Curado, M., Mauthe, A.: Recent advances in multimedia networking. In: Multimedia Tools Appl., pp. 635–647 (2011)
3. ITU-T Recommendation Y.1541. Network performance objectives for IP-based services (February 2006)
4. Titov, I.: Characteristics of the data flows generated by a Web-server. T-Comm - Telecommunications and Transport (5), 30–33 (2010) (in Russian)
5. Titov, I.: Research of data server's traffic model by the results of multimedia resource's traffic measurement. T-Comm - Telecommunications and Transport (5), 46–49 (2011) (in Russian)

6. Anderson, A.T., Nielsen, B.F.: A Markovian approach for modeling packet traffic with long-range dependence. IEEE Journal on Selected Areas in Communications 16(5), 719–732 (1998)
7. Dudin, A.N., Klimenok, V.I., Tsarenkov, G.V.: A Single-Server Queueing System with Batch Markov Arrivals, Semi-Markov Service, and Finite Buffer: Its Characteristics. Automation and Remote Control 63(8), 1285–1297 (2002)
8. Seghaier, A.: Numerical Study of Loss Probability Calculation Algorithms for Multistreaming Models of Packet Networks. Journal of Communications Technology and Electronics 55(12), 1499–1513 (2010)
9. Seghaier, A., Tsitovich, I.: On the interval model of the birth–death process with a hysteresis. Information Processes 12(1), 117–126 (2012)
10. ITU-T Recommendation E.800. Definitions of terms related to quality of service (September 2008)
11. ITU-T Recommendation E.802. Framework and methodologies for the determination and application of QoS parameters (February 2007)
12. Tsitovich, I., Titov, I.: Time scale in mathematical model of source with unbounded service time variation. Information Processes 11(3), 369–377 (2011) (in Russian)
13. Tsitovich, I., Titov, I.: About characteristics of data server's traffic variation and its loss probability. In: Information and Telecommunication Technologies and Mathematical Modeling of High Technology Systems, Moscow, pp. 61–63 (2012) (in Russian)
14. Tsitovich, I., Titov, I.: Analysis of loss probability for multimedia resource's traffic. In: Information Technology and Systems Conference, Moscow, pp. 484–489 (2012) (in Russian)

On a Queueing Model with Group Services

Alexander Zeifman[1], Anna Korotysheva[2], Yakov Satin[2],
Galina Shilova[2], and Tatyana Panfilova[2]

[1] Vologda State Pedagogical University,
Institute of Informatics Problems RAS, and ISEDT RAS
[2] Vologda State Pedagogical University

Abstract. An analogue of $M_t/M_t/S/S$ Erlang loss system for a queue with group services is introduced and considered. Weak ergodicity of the model is studied. We obtain the bounds on the rate of convergence to the limiting characteristics and consider two concrete queueing models with finding of their main limiting characteristics.

Keywords: Markovian queueing models, group service, weak ergodicity, bounds.

1 Introduction

Stationary and nonstationary Erlang loss queueing model has been considered by a great number of authors, see [1]-[7],[9,11,16].

We believe the main reasons for the wide application of the model are simplicity and ease of application.

Since the 1970s, the most important problems have been connected with the bounds on the rate of convergence to the limiting regime [1,3,5,6,9,11,13,15,16].

Here we introduce and study a simplest analogue of $M_t/M_t/S/S$ queue for a queueing system with group services.

Namely, we suppose that an intensity of arrival of a customer to the queue is $\lambda(t)$, and an intensity of departure (servicing) of a group of k customers is $\mu_k(t) = \frac{\mu(t)}{k}$ if $1 \le k \le S$, and we also suppose that $X(t) \le S$, i.e. there are no waiting rooms.

Let $X = X(t)$, $t \ge 0$ be a queue-length process for the queue.

We suppose that intensities $\lambda(t)$ and $\nu(t)$ are locally integrable on $[0, \infty)$ nonnegative functions.

Then the probabilistic dynamics of the process is represented by the forward Kolmogorov differential system:

$$\frac{d\mathbf{p}}{dt} = A(t)\mathbf{p}(t). \tag{1}$$

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 198–205, 2013.

Here $A(t)$ is transposed intensity matrix,

$$A(t) = \begin{pmatrix} a_{00}(t) & \mu_1(t) & \mu_2(t) & \mu_3(t) & \mu_4(t) & \cdots & \mu_S(t) \\ \lambda(t) & a_{11}(t) & \mu_1(t) & \mu_2(t) & \mu_3(t) & \cdots & \mu_{S-1}(t) \\ 0 & \lambda(t) & a_{22}(t) & \mu_1(t) & \mu_2(t) & \cdots & \mu_{S-2}(t) \\ \cdots & & & & & & \\ 0 & 0 & 0 & \cdots & 0 & \lambda(t) & a_{SS}(t) \end{pmatrix}, \tag{2}$$

where $\mu_k(t) = \mu(t)/k$, and $a_{ii}(t)$ are such that all column sums in $A(t)$ equal zero for any $t \geq 0$.

We denote throughout the paper by $\| \bullet \|$ the l_1-norm, i. e. $\|\mathbf{x}\| = \sum |x_i|$, and $\|B\| = \max_j \sum_i |b_{ij}|$ for $B = (b_{ij})_{i,j=0}^{S}$.

Let Ω be a set all stochastic vectors, i. e. l_1 vectors with nonnegative coordinates and unit norm.

Let $E_k(t) = E\{X(t) \,|X(0) = k\}$ be the mean of the process at the moment t under initial condition $X(0) = k$, and $E_{\mathbf{p}}(t)$ be the mathematical expectation (the mean) at the moment t under initial probability distribution $\mathbf{p}(0) = \mathbf{p}$.

Recall that the process $X(t)$ is weakly ergodic if $\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\| \to 0$ as $t \to \infty$ for any initial conditions $\mathbf{p}^*(s), \mathbf{p}^{**}(s)$ and any $s \geq 0$. In particular, if the state space is finite, and the intensities are constant ($X(t)$ is stationary), then weak ergodicity is equivalent to ergodicity of the process, that is, the existence of steady-state distribution, say π.

2 Ergodicity Bounds

Theorem 1. *Queue-length process $X(t)$ is weakly ergodic if and only if the following assumption holds:*

$$\int_0^\infty (\lambda(t) + \mu(t))\, dt = +\infty. \tag{3}$$

Proof. Let firstly (3) do not satisfied. Then we have

$$\|A(t)\| = 2 \max |a_{ii}(t)| = 2 \left(\lambda(t) + \mu(t) \cdot \sum_{k=1}^{S} \frac{1}{k} \right) \leq 2 \left(\lambda(t) + (1 + \log S)\, \mu(t) \right), \tag{4}$$

and

$$\int_0^\infty \|A(t)\|\, dt < +\infty. \tag{5}$$

Therefore, $X(t)$ does not weakly ergodic by Theorem 3.3 [12].

Let now (3) hold. Using the method which was proposed by one of us, see [10], put $p_0 = 1 - \sum_{1 \leq i \leq S} p_i$. Then we obtain from (1) the following equation:

$$\frac{d\mathbf{z}}{dt} = B(t)\mathbf{z}(t) + \mathbf{f}(t), \tag{6}$$

where $\mathbf{f}(t) = (\lambda, 0, \cdots, 0)^T$,

$$B = \begin{pmatrix} a_{11}-\lambda & \mu_1-\lambda & \mu_2-\lambda & \mu_3-\lambda & \cdots & \cdots & \mu_{S-1}-\lambda \\ \lambda & a_{22} & \mu_1 & \mu_2 & \cdots & \cdots & \mu_{S-2} \\ 0 & \lambda & a_{33} & \mu_1 & \cdots & \cdots & \mu_{S-3} \\ \cdots & & & & & & \\ 0 & 0 & \cdots & \cdots & 0 & \lambda & a_{SS} \end{pmatrix}. \tag{7}$$

Employing the approach of [8,13,14], consider the triangular matrix

$$D = \begin{pmatrix} d_1 & d_1 & d_1 & \cdots & d_1 \\ 0 & d_2 & d_2 & \cdots & d_2 \\ \cdots & & & & \\ 0 & 0 & \cdots & 0 & d_S \end{pmatrix}, \tag{8}$$

and the respective vector norm $\|\mathbf{z}\|_{1D} = \|D\mathbf{z}\|_1$.

Then we obtain

$$DBD^{-1} = \begin{pmatrix} a_{11} & (\mu_1-\mu_2)\frac{d_1}{d_2} & (\mu_2-\mu_3)\frac{d_1}{d_3} & \cdots & (\mu_{S-1}-\mu_S)\frac{d_1}{d_S} \\ \lambda\frac{d_2}{d_1} & a_{22} & (\mu_1-\mu_3)\frac{d_2}{d_3} & \cdots & (\mu_{S-2}-\mu_S)\frac{d_2}{d_S} \\ 0 & \lambda\frac{d_3}{d_2} & a_{33} & \cdots & (\mu_{S-3}-\mu_S)\frac{d_3}{d_S} \\ \cdots & & & & \\ 0 & 0 & 0 & \cdots & a_{SS}-\lambda \end{pmatrix}. \tag{9}$$

Essential Service Rate. Let

$$\int_0^\infty \mu(t)\,dt = +\infty. \tag{10}$$

Put all $d_i = 1$. Then we have the following bound for the logarithmic norm of $B(t)$ in $D-$norm, see details in [5,13,14]:

$$\begin{gathered} \gamma\left(B(t)\right)_{1D} = \gamma\left(DB(t)D^{-1}\right)_1 = \\ \max\left(a_{SS}(t) - \lambda(t) + \sum_{k=1}^{S-1}(\mu_{S-k}(t) - \mu_S(t))\frac{d_k}{d_S}, \right. \\ \left. \max_{1\le i\le S-1}\left(a_{ii}(t) + \sum_{k=1}^{i-1}(\mu_{i-k}(t) - \mu_i(t))\frac{d_k}{d_i} + \lambda(t)\frac{d_{i+1}}{d_i}\right)\right) = \\ \max_{1\le i\le S-1}\left(-k\mu_k(t)\right) = -\mu(t). \end{gathered} \tag{11}$$

Then

$$\|\mathbf{z}^*(t) - \mathbf{z}^{**}(t)\|_{1D} \le e^{-\int_s^t \mu(u)du}\|\mathbf{z}^*(s) - \mathbf{z}^{**}(s)\|_{1D}, \tag{12}$$

for any $0 \le s \le t$ and any initial conditions $\mathbf{z}^*(s), \mathbf{z}^{**}(s)$.

We have $\|D\| = \sum_{i=1}^S d_i = S$, $\|D^{-1}\| = 2\max\frac{1}{d_k} = 2$.

Then the following bound in 'natural' l_1-norm holds:

$$\begin{gathered} \|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\| \le 2\|\mathbf{z}^*(t) - \mathbf{z}^{**}(t)\| = 2\|D^{-1}D\left(\mathbf{z}^*(t) - \mathbf{z}^{**}(t)\right)\| \le \\ 4\|\mathbf{z}^*(t) - \mathbf{z}^{**}(t)\|_{1D} \le 4e^{-\int_s^t \mu(\tau)\,d\tau}\|\mathbf{z}^*(s) - \mathbf{z}^{**}(s)\|_{1D} \le \\ 4Se^{-\int_s^t \mu(\tau)\,d\tau}\|\mathbf{z}^*(s) - \mathbf{z}^{**}(s)\| \le 4Se^{-\int_s^t \mu(\tau)\,d\tau}\|\mathbf{p}^*(s) - \mathbf{p}^{**}(s)\| \le 8Se^{-\int_s^t \mu(\tau)\,d\tau}, \end{gathered} \tag{13}$$

for any initial conditions $\mathbf{p}^*(s)$, $\mathbf{p}^{**}(s)$ and any $s, t,\ 0 \le s \le t$.

Finally, $X(t)$ is weakly ergodic and the following bound on the rate of convergence holds:

$$\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\| \le 8Se^{-\int_0^t \mu(\tau)\,d\tau}, \tag{14}$$

for any initial conditions $\mathbf{p}^*(0)$, $\mathbf{p}^{**}(0)$ and any $t \ge 0$.

Essential Arrival Rate. Let

$$\int_0^\infty \lambda(t)\,dt = +\infty. \tag{15}$$

Put all $d_k = \frac{1}{k}$. Hence in accordance with (11) we have the following bound for the logarithmic norm of $B(t)$:

$$\gamma\left(B(t)\right)_{1D} = -\frac{1}{S}\lambda(t) - \frac{1}{S-1}\mu(t) \le -\frac{\lambda(t)}{S}. \tag{16}$$

Then

$$\|\mathbf{z}^*(t) - \mathbf{z}^{**}(t)\|_{1D} \le e^{-\frac{1}{S}\int_s^t \lambda(u)du}\|\mathbf{z}^*(s) - \mathbf{z}^{**}(s)\|_{1D}, \tag{17}$$

for any $0 \le s \le t$ and any initial conditions $\mathbf{z}^*(s), \mathbf{z}^{**}(s)$.

We have $\|D\| = \sum_{i=1}^S d_i \le 1 + \log S$ and $\|D^{-1}\| = 2\max\frac{1}{d_k} = 2S$.

Therefore the following bound on the rate of convergence holds instead of (13) and (14) :

$$\begin{gathered}\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\| \le 2\|D^{-1}D\left(\mathbf{z}^*(t) - \mathbf{z}^{**}(t)\right)\| \le \\ 8S\left(1 + \log S\right)e^{-\frac{1}{S}\int_s^t \lambda(\tau)\,d\tau},\end{gathered} \tag{18}$$

for any initial conditions $\mathbf{p}^*(s)$, $\mathbf{p}^{**}(s)$ and any $s, t,\ 0 \le s \le t$.

Corollary 1. *Let (3) be fulfilled. Then queue-length process $X(t)$ is weakly ergodic and both bounds on the rate of convergence (14) and (18) hold.*

Bounds (14) and (18) are useful for estimation of the rate of convergence for the mean of the length of queue. Namely the next statement follows immediately from the inequality $\sum_{k=0}^S k|p_k| \le S\sum_{k=0}^S |p_k|$.

Corollary 2. *The following bounds on the rate of convergence for the mean hold:*

$$|E_{\mathbf{p}^*}(t) - E_{\mathbf{p}^{**}}(t)| \le 8S^2e^{-\int_0^t \mu(\tau)\,d\tau}, \tag{19}$$

and

$$|E_{\mathbf{p}^*}(t) - E_{\mathbf{p}^{**}}(t)| \le 8S^2\left(1 + \log S\right)e^{-\frac{1}{S}\int_s^t \lambda(\tau)\,d\tau}, \tag{20}$$

for any initial probability distributions $\mathbf{p}^*(0),\ \mathbf{p}^{**}(0)$, and any $t \ge 0$.

Corollary 3. *If λ and μ are constant, then queue-length process $X(t)$ is ergodic if and only if $\lambda + \mu > 0$. If $\mu > 0$ then the following bound on the rate of convergence holds:*

$$\|\mathbf{p}(t) - \pi\| \le 8Se^{-\mu t}, \tag{21}$$

$$|E_{\mathbf{p}}(t) - \phi| \le 8S^2 e^{-\mu t}, \tag{22}$$

where $\phi = \sum_{k=0}^{S} k\pi_k$, for any initial condition $\mathbf{p}(0) = \mathbf{p}$ and any $t \ge 0$. Similar bounds hold for $\lambda > 0$.

Corollary 4. *Let $\lambda(t)$ and $\mu(t)$ be $1-$periodic. Then queue-length process $X(t)$ is weakly ergodic if and only if*

$$\int_0^1 (\lambda(t) + \mu(t))\, dt > 0. \tag{23}$$

If $\int_0^1 \mu(t)\, dt > 0$ then the following bound on the rate of convergence holds:

$$\|\mathbf{p}(t) - \pi(t)\| \le 8Se^{-\int_0^t \mu(u)\, du}, \tag{24}$$

$$|E_{\mathbf{p}}(t) - \phi(t)| \le 8S^2 e^{-\int_0^t \mu(u)\, du}, \tag{25}$$

where $\pi(t)$ is the limiting $1-$periodic regime and $\phi(t) = \sum_{k=0}^{S} k\pi_k(t)$ is the respective $1-$periodic limiting mean. Similar bounds hold for $\int_0^1 \lambda(t)\, dt > 0$.

Remark 1. One can obtain perturbation bounds for the model using the results obtained in [8].

3 Examples

We consider two queueing models: ordinary $M_t/M_t/S/S$ Erlang loss system, and its analogue for a queue with group services with the same characteristics.

Put $S = 10^3$, $\lambda(t) = 1 + \sin 2\pi t$, $\mu(t) = 1 + \cos 2\pi t$.

Then we can apply estimates (24) and (25) for the *both* models, see Corollary 4 of the present paper, and Corollary 2 of [16].

We have $e^{-\int_0^t \mu(u)\, du} = e^{-t - \frac{\sin 2\pi t}{2\pi}} \le 1.2e^{-t}$ and therefore the following bounds hold for the *both* queueing models:

$$\|\mathbf{p}(t) - \pi(t)\| \le 10^4 e^{-t}, \tag{26}$$

$$|E_{\mathbf{p}}(t) - \phi(t)| \le 10^7 e^{-t}, \tag{27}$$

where $\pi(t)$ is the limiting $1-$periodic regime and $\phi(t) = \sum_{k=0}^{S} k\pi_k(t)$ is $1-$periodic limiting mean for the respective queue-length process.

Then $\mathbf{p}(t) \approx \pi(t)$ and $E_{\mathbf{p}}(t) \approx \phi(t)$, if $t \ge 23$ with an error less than 10^{-6} and 10^{-3} respectively. Hence we can calculate the limiting characteristics for the respective queue-length process by solving the Cauchy problem for the forward

Kolmogorov system (1) on the interval $[0, 24]$ with initial condition $X(0) = 0$ (i.e. $\mathbf{p}(0) = \mathbf{e}_0 = (1, 0, 0, \ldots, 0)^T$). Finally, we obtain (approximately) the limiting $1-$periodic regime $\pi(t)$ and $1-$periodic limiting mean $\phi(t)$ for both processes on the interval $[23, 24]$.

Remark 2. One can see the interesting fact that the limiting mathematical expectations are the same for both examples .

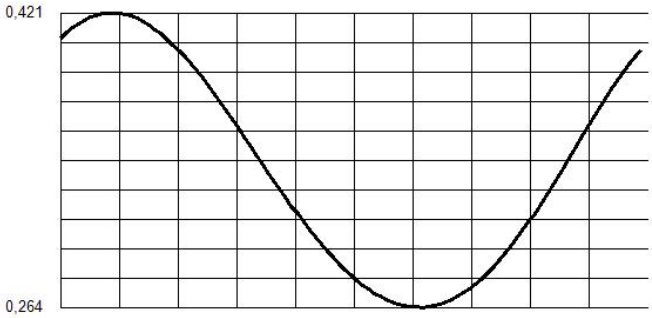


Fig. 1. Approximation of the limiting probability of the empty queue for Erlang model

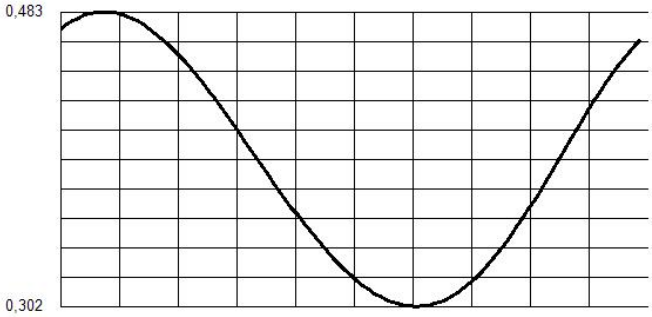


Fig. 2. Approximation of the limiting probability of the empty queue for analogue of Erlang loss system with group services

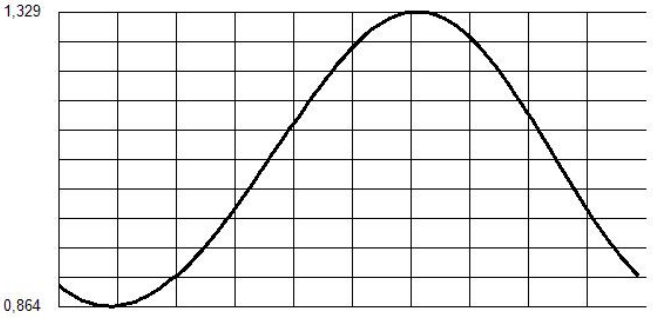


Fig. 3. Approximation of the limiting mean for Erlang model

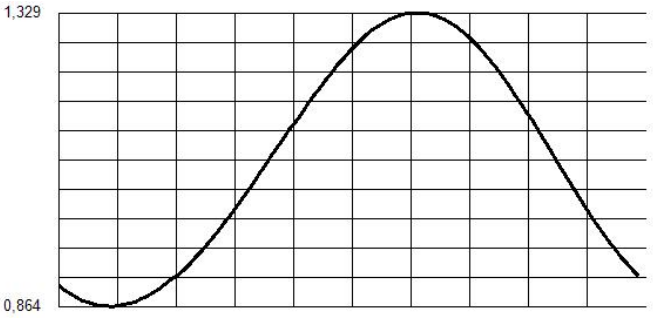


Fig. 4. Approximation of the limiting mean for analogue of Erlang loss system with group services

Acknowledgments. This work was supported by the Russian Foundation for Basic Research, projects no. 11-01-12026, 12-07-00115, 12-07-00109.

References

1. Van Doorn, E.A., Zeifman, A.I.: On the speed of convergence to stationarity of the Erlang loss system. Queueing Syst. 63, 241–252 (2009)
2. Erlang, A.K.: Løsning af nogle Problemer fra Sandsynlighedsregningen af Betydning for de automatiske Telefoncentraler. Elektroteknikeren 13, 5–13 (1917)
3. Fricker, C., Robert, P., Tibi, D.: On the rate of convergence of Erlang's model. J. Appl. Probab. 36, 1167–1184 (1999)
4. Gnedenko, B.V., Makarov, I.P.: Properties of a problem with losses in the case of periodic intensities. Diff. Equations 7, 1696–1698 (1971) (in Russian)
5. Granovsky, B., Zeifman, A.: Nonstationary queues: estimation of the rate of convergence. Queueing Syst. 46, 363–388 (2004)
6. Kijima, M.: On the largest negative eigenvalue of the infinitesimal generator associated with $M/M/n/n$ queues. Oper. Res. Let. 9, 59–64 (1990)
7. Massey, W.A., Whitt, W.: On analysis of the modified offered-load approximation for the nonstationary Erlang loss model. Ann. Appl. Probab. 4, 1145–1160 (1994)
8. Satin, Y.A., Zeifman, A.I., Korotysheva, A.V., Shorgin, S.Y.: On a class of Markovian queues. Informatics and Its Applications 5(4), 6–12 (2011) (in Russian)
9. Voit, M.: A note of the rate of convergence to equilibrium for Erlang's model in the subcritical case. J. Appl. Probab. 37, 918–923 (2000)
10. Zeifman, A.I.: Stability for contionuous-time nonhomogeneous Markov chains. Lect. Notes Math. 1155, 401–414 (1985)
11. Zeifman, A.I.: Properties of a System with Losses in the Case of Variable Rates. Autom. Remote Control (1), 82–87 (1989)
12. Zeifman, A.I., Isaacson, D.: On strong ergodicity for nonhomogeneous continuous-time Markov chains. Stoch. Proc. Appl. 50, 263–273 (1994)
13. Zeifman, A.I.: Upper and lower bounds on the rate of convergence for nonhomogeneous birth and death processes. Stoch. Proc. Appl. 59, 157–173 (1995)
14. Zeifman, A., Leorato, S., Orsingher, E., Satin, Y., Shilova, G.: Some universal limits for nonhomogeneous birth and death processes. Queueing Systems 52, 139–151 (2006)
15. Zeifman, A. I., Bening, V. E., Sokolov, I. A.: Markov Chains and Models in Continuous Time. Elex-KM, Moscow (2008) (in Russian)
16. Zeifman, A.I.: On the nonstationary Erlang loss model. Autom. Rem. Contr. 70, 2003–2012 (2009)

Study of Queues' Sizes in Tandem Intersections under Cyclic Control in Random Environment

Andrei Zorine*

N.I. Lobachevsky University of Nizhni Novgorod
Gagarina ave., 23, 603950 Nizhni Novgorod, Russia
zoav1602@gmail.com

Abstract. Tandem queueing systems under cyclic control with readjustments are investigated. Conflict input flows are formed in a random synchronious environment. Transition of customers from the first system to the second system occurs with random speeds. Two communicating intersections give an example of such a tandem. The blocks of the systems are described nonlocally. A mathematical model is constructed in form of a multidimensional denumerable discrete-time Markov chain. Limit behaviour of queues' sizes is studied.

Keywords: Conflict input flows, cyclic control, random environment, Markov chain, limit theorems.

1 Introduction

System of control for traffic flows attract researchers constantly [1,2,3,4]. Contemporary traffic flows in cities are highly intensive, have dependent inter-arrival intervals and variable structure. In these conditions the cyclic control with readjustments (yellow light) is quasi-optimal. So, construction and analysis of adequate stochastic models for tandem intersections with cyclic control algorithm at each intersection is tempting [5].

2 Problem Statement and Model Construction

Consider the following two single server queueing systems. Conflict input flows Π_1, Π_2 enter the first queueing system, while conflict flows Π_3, Π_4 enter the second. In essence, conflictness means that customers from different flows can not be serviced in the same queueing system simultaneously, and that the input flows can not be joined to reduce the number of input flows (and the dimensions of the resulting stochastic process). Input flows Π_1, Π_2, Π_4 are formed in a random external environment with a finite number d of states $e^{(1)}$, $e^{(2)}$, ..., $e^{(d)}$. The environment may change its state only at instants when one of the

* This research was supported by RFBR Grant 12-01-90409 "Modelling and analysis of controlling systems for interacting high intensity transport flows".

A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, pp. 206–215, 2013.

servers terminates work or readjustment. The probability of switch from $e^{(k)}$ to $e^{(l)}$ is $a_{k,l}$. In state $e^{(k)}$, $k = 1, 2, \ldots, d$, customers from Π_j, $j = 1, 2, 4$, arrive in batches so that the flow of batches is Poisson with parameter $\lambda_j^{(k)}$, and batch sizes are independent having b customers with probability $\pi(b; j, k)$. Input flow Π_3 consists of retrial customers from Π_1 and Π_2. Namely, after service, customers from Π_1 are directed into the second queueing system; they make input flow Π_5 of customers, for which the movement from the first queueing system to the second queueing system is another kind of service. Upon service termination each customer from Π_2, independently of the others, either joins Π_5 instantly with probability α and starts movement towards the second queueing system, or leaves tandem queueing systems with probability $1 - \alpha$ and joins the corresponding output flow. Thus one of the two input flows of the second queueing system consists of the flow of retrial customers from the first queueing system. It is assumed that transition from system to system takes time and occurs with random speed with unknown probability distribution. Moreover, the speeds of different customers are different and have different laws of probability distributions. Hence we assume that during a working act or a readjustment act each customer moving between the queueing systems either finishes with known probability, or keeps moving with complementary probability. Output flow of customers, whose service consisted in moving between two queueing systems, is the input flow Π_3 to the second queueing system. Customers from the flow Π_j, $j = 1, 2, \ldots, 5$, wait in a buffer O_j of unlimited capacity.

Service of conflict flows in each queueing system is done according to a cyclic algorithm with fixed durations. To resolve conflictness, after each service act a readjustment act is required. No customer is serviced during a readjustment act. It means that the server in the first queueing system has four possible states (regimes) $\Gamma^{(1,1)}$, $\Gamma^{(2,1)}$, $\Gamma^{(3,1)}$, $\Gamma^{(4,1)}$, and the duration for the state $\Gamma^{(s,1)}$ is non-random and equals $T_{s,1}$. In state $\Gamma^{(1,1)}$ only customers from Π_1 are serviced, in state $\Gamma^{(2,1)}$ only customers from Π_2 are serviced, and in states $\Gamma^{(2,1)}$ and $\Gamma^{(4,1)}$ readjustment is carried out. The states shift in the order $\ldots \to \Gamma^{(1,1)} \to \Gamma^{(2,1)} \to \Gamma^{(3,1)} \to \Gamma^{(4,1)} \to \Gamma^{(1,1)} \to \ldots$. The server in the second queueing system has also four possible states (regimes) $\Gamma^{(1,2)}$, $\Gamma^{(2,2)}$, $\Gamma^{(3,2)}$, $\Gamma^{(4,2)}$, and the duration for the state $\Gamma^{(s,2)}$ is non-random and equals $T_{s,2}$. In state $\Gamma^{(1,2)}$ only customers from Π_3 are serviced, in state $\Gamma^{(2,2)}$ only customers from Π_4 are serviced, and in states $\Gamma^{(2,2)}$ and $\Gamma^{(4,2)}$ a readjustment takes place. The states shift in the order $\ldots \to \Gamma^{(1,2)} \to \Gamma^{(2,2)} \to \Gamma^{(3,2)} \to \Gamma^{(4,2)} \to \Gamma^{(1,2)} \to \ldots$. Numbers $T_{1,1}, \ldots, T_{4,2}$ are assumed commensurable. In the remaining of this work it is convenient to consider the two servers as a new single server with some number n of cyclic states $\Gamma^{(1)}, \Gamma^{(2)}, \ldots, \Gamma^{(n)}$, and a fixed duration T_r for state $\Gamma^{(r)}$, $1 \le r \le n$. The new server changes its states exactly when one of the original servers changes its state. Here, number n and the durations T_r, $1 \le r \le n$ are uniquely determined by the initial states $\Gamma^{(r',1)}$, $\Gamma^{(r'',2)}$ at time 0 and by durations $T_{1,1}, \ldots, T_{4,2}$. The new states change by the following rule: $\Gamma^{(r)} \to \Gamma^{(r \oplus 1)}$ where $r \oplus 1 = r + 1$ for $1 \le r \le n - 1$, $n \oplus 1 = 1$. We write $u(\Gamma^{(r)}) = \Gamma^{(r \oplus 1)}$ then. Recall that besides ordinary service of flows Π_1, Π_2, Π_3, Π_4, the new server also delivers

service to customers from Π_5. So, state $\Gamma^{(r)}$ can belong to one of nine classes Γ^{I}, Γ^{II}, …, Γ^{IX}. For $\Gamma^{(r)} \in \Gamma^{\mathrm{I}}$ only customers from queue O_5 are serviced; for $\Gamma^{(r)} \in \Gamma^{\mathrm{II}}$ — from queues O_1, O_5; for $\Gamma^{(r)} \in \Gamma^{\mathrm{III}}$ — from queues O_2, O_5; for $\Gamma^{(r)} \in \Gamma^{\mathrm{IV}}$ — from queues O_3, O_5; for $\Gamma^{(r)} \in \Gamma^{\mathrm{V}}$ — from queues O_4, O_5; for $\Gamma^{(r)} \in \Gamma^{\mathrm{VI}}$ — from queues O_1, O_3, O_5; for $\Gamma^{(r)} \in \Gamma^{\mathrm{VII}}$ — from queues O_2, O_3, O_5; for $\Gamma^{(r)} \in \Gamma^{\mathrm{VIII}}$ — from queues O_1, O_4, O_5; finally, for $\Gamma^{(r)} \in \Gamma^{\mathrm{IX}}$ only customers from queues O_2, O_4, O_5 are serviced. In state $\Gamma^{(r)}$ of the server, during time interval of length T_r each customer from queue O_5 either finishes service with probability p_r, leaves O_5 and joins Π_3, or with probability $1 - p_r$ remains in O_5 for the next tact.

Service durations for single customers in each system are random, in general, dependent, and with different laws of probability distribution. Thus to define service processes we use saturation flows Π_j^{sat}, $j = 1, 2, \ldots, 5$, i.e. virtual output flows in conditions of highly loaded queues and maximal usage of server's resources. During time T_r let the saturation flow Π_1^{sat} contain nonrandom number $\ell_{r,1} \geq 1$ of customers in server state $\Gamma^{(r)} \in \Gamma^{\mathrm{II}} \cup \Gamma^{\mathrm{VI}} \cup \Gamma^{\mathrm{VIII}}$, and 0 customers in server state $\Gamma^{(r)} \notin \Gamma^{\mathrm{II}} \cup \Gamma^{\mathrm{VI}} \cup \Gamma^{\mathrm{VIII}}$, Π_2^{sat} contain nonrandom number $\ell_{r,2} \geq 1$ of customers in server state $\Gamma^{(r)} \in \Gamma^{\mathrm{III}} \cup \Gamma^{\mathrm{VII}} \cup \Gamma^{\mathrm{IX}}$ and 0 customers $\Gamma^{(r)} \notin \Gamma^{\mathrm{III}} \cup \Gamma^{\mathrm{VII}} \cup \Gamma^{\mathrm{IX}}$, Π_3^{sat} contain nonrandom number $\ell_{r,3} \geq 1$ of customers in server state $\Gamma^{(r)} \in \Gamma^{\mathrm{IV}} \cup \Gamma^{\mathrm{VI}} \cup \Gamma^{\mathrm{VII}}$ and 0 customers in server state $\Gamma^{(r)} \notin \Gamma^{\mathrm{IV}} \cup \Gamma^{\mathrm{VI}} \cup \Gamma^{\mathrm{VII}}$, Π_4^{sat} contain nonrandom number $\ell_{r,4} \geq 1$ of customers in server state $\Gamma^{(r)} \in \Gamma^{\mathrm{V}} \cup \Gamma^{\mathrm{VIII}} \cup \Gamma^{\mathrm{IX}}$ and 0 customers in server state $\Gamma^{(r)} \notin \Gamma^{\mathrm{V}} \cup \Gamma^{\mathrm{VIII}} \cup \Gamma^{\mathrm{IX}}$.

Tandem intersections with traffic flows Π_1 – Π_7 is a possible interpretation for tandem queueing systems in study (Fig. 1). Assuming flows Π_6, Π_7 have low intensity one can think that flow Π_6 passes through together with flow Π_3 of the second intersection and together with flow Π_1 of the first intersection, while traffic flow Π_7 is let through at the same time as flow Π_4 of the second intersection. Flow Π_5 is the total of serviced customers from Π_1, and serviced customers from Π_2 with certain thinning probability.

Let $\tau_0 = 0$, τ_1, τ_2, … be the instants of changes of server states. In what follows we need the following sets, random elements and randoms variables: the set $E = \{e^{(1)}, e^{(2)}, \ldots, e^{(d)}\}$ of environment states, the random element $\chi_i \in E$ describing environment state in the interval $(\tau_i, \tau_{i+1}]$, the size $\kappa_{j,i}$ of the queue O_j at instant τ_i, the set $\Gamma = \{\Gamma^{(1)}, \Gamma^{(2)}, \ldots, \Gamma^{(n)}\}$ of possible server states, server state $\Gamma_i \in \Gamma$ at instant τ_i, the number $\eta_{j,i}$ of customers from Π_j arrived during the time interval $(\tau_i, \tau_{i+1}]$, the number $\xi_{j,i}$ of customers in saturation flow $\Pi_j^{\mathbf{sat}}$ during $(\tau_i, \tau_{i+1}]$, the number $\bar{\xi}_{j,i}$ of customers of output flow $\Pi_j^{\mathbf{out}}$ during time interval $(\tau_i, \tau_{i+1}]$. Finally, denote by $\eta'_{5,i}$ the number of retrial customers from O_2 redirected to flow Π_5 after service.

To have a mathematical model for the input flows Π_1, Π_2, Π_4 consider a marked point process $\{(\tau_i, \eta_{1,i}, \eta_{2,i}, \eta_{4,i}, \nu_i); i = 0, 1, \ldots\}$ with a mark $\nu_i = (\Gamma_i, \chi_i)$ of customers arrived during $(\tau_i, \tau_{i+1}]$. Define $\varphi_{j,k}(x; t)$, $t > 0$, $j = 1, 2, 4$, $k = 1, 2, \ldots, d$, through the expansion

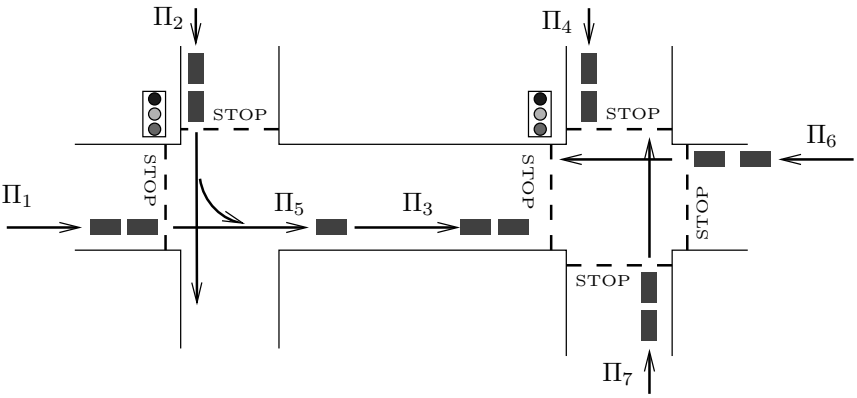


Fig. 1. Tandem intersections

$$q_{j,k}(z;t) = \sum_{x=0}^{\infty} z^x \varphi_{j,k}(x;t) = \exp\Big\{\lambda_j^{(k)} t\Big(\sum_{b=1}^{\infty} z^b \pi(b;j,k) - 1\Big)\Big\}.$$

Here $\varphi_{j,k}(x;t)$ is the probability of exactly $x = 0, 1, \ldots$ customers' arrivals from $\mathit{\Pi}_j$ in environment state $e^{(k)}$ during time t, and $q_{j,k}(z;t)$ is the corresponding probability generating function. Let the conditional probability distribution for the discrete selected component $\{(\eta_{1,i}, \eta_{2,i}, \eta_{4,i}, \nu_i); i = 0, 1, \ldots\}$ have the following form: for $b = 0, 1, \ldots$

$$\mathbf{P}(\{\omega\colon \eta_{j,i} = b\} | \{\omega\colon \mathit{\Gamma}_i = \mathit{\Gamma}^{(r)}, \chi_i = e^{(k)}\}) = \varphi_{j,k}(b; T_{r\oplus 1}).$$

For the saturation flows $\mathit{\Pi}_1^{\mathrm{sat}}$, $\mathit{\Pi}_2^{\mathrm{sat}}$, $\mathit{\Pi}_3^{\mathrm{sat}}$, $\mathit{\Pi}_4^{\mathrm{sat}}$ consider the marked point process $\{(\tau_i, \xi_{1,i}, \xi_{2,i}, \xi_{3,i}, \xi_{4,i}, \nu_i); i = 0, 1, \ldots\}$. Conditional probability distributions for the selected discrete component $\{(\xi_{1,i}, \xi_{2,i}, \xi_{3,i}, \xi_{4,i}, \nu_i); i = 0, 1, \ldots\}$ are given by

$$\begin{aligned}
\mathbf{P}(\{\omega\colon \xi_{1,i} = 0\} | \{\omega\colon \mathit{\Gamma}_i = \mathit{\Gamma}^{(r)}\}) &= 1 \quad \text{for } \mathit{\Gamma}^{(r\oplus 1)} \notin \mathit{\Gamma}^{\mathrm{II}} \cup \mathit{\Gamma}^{\mathrm{IV}} \cup \mathit{\Gamma}^{\mathrm{VIII}},\\
\mathbf{P}(\{\omega\colon \xi_{1,i} = \ell_{r\oplus 1,1}\} | \{\omega\colon \mathit{\Gamma}_i = \mathit{\Gamma}^{(r)}\}) &= 1 \quad \text{for } \mathit{\Gamma}^{(r\oplus 1)} \in \mathit{\Gamma}^{\mathrm{II}} \cup \mathit{\Gamma}^{\mathrm{IV}} \cup \mathit{\Gamma}^{\mathrm{VIII}},\\
\mathbf{P}(\{\omega\colon \xi_{2,i} = 0\} | \{\omega\colon \mathit{\Gamma}_i = \mathit{\Gamma}^{(r)}\}) &= 1 \quad \text{for } \mathit{\Gamma}^{(r\oplus 1)} \notin \mathit{\Gamma}^{\mathrm{III}} \cup \mathit{\Gamma}^{\mathrm{VII}} \cup \mathit{\Gamma}^{\mathrm{IX}},\\
\mathbf{P}(\{\omega\colon \xi_{2,i} = \ell_{r\oplus 1,2}\} | \{\omega\colon \mathit{\Gamma}_i = \mathit{\Gamma}^{(r)}\}) &= 1 \quad \text{for } \mathit{\Gamma}^{(r\oplus 1)} \in \mathit{\Gamma}^{\mathrm{III}} \cup \mathit{\Gamma}^{\mathrm{VII}} \cup \mathit{\Gamma}^{\mathrm{IX}},\\
\mathbf{P}(\{\omega\colon \xi_{3,i} = 0\} | \{\omega\colon \mathit{\Gamma}_i = \mathit{\Gamma}^{(r)}\}) &= 1 \quad \text{for } \mathit{\Gamma}^{(r\oplus 1)} \notin \mathit{\Gamma}^{\mathrm{IV}} \cup \mathit{\Gamma}^{\mathrm{VI}} \cup \mathit{\Gamma}^{\mathrm{VII}},\\
\mathbf{P}(\{\omega\colon \xi_{3,i} = \ell_{r\oplus 1,3}\} | \{\omega\colon \mathit{\Gamma}_i = \mathit{\Gamma}^{(r)}\}) &= 1 \quad \text{for } \mathit{\Gamma}^{(r\oplus 1)} \in \mathit{\Gamma}^{\mathrm{IV}} \cup \mathit{\Gamma}^{\mathrm{VI}} \cup \mathit{\Gamma}^{\mathrm{VII}},\\
\mathbf{P}(\{\omega\colon \xi_{4,i} = 0\} | \{\omega\colon \mathit{\Gamma}_i = \mathit{\Gamma}^{(r)}\}) &= 1 \quad \text{for } \mathit{\Gamma}^{(r\oplus 1)} \notin \mathit{\Gamma}^{\mathrm{V}} \cup \mathit{\Gamma}^{\mathrm{VIII}} \cup \mathit{\Gamma}^{\mathrm{IX}},\\
\mathbf{P}(\{\omega\colon \xi_{4,i} = \ell_{r\oplus 1,4}\} | \{\omega\colon \mathit{\Gamma}_i = \mathit{\Gamma}^{(r)}\}) &= 1 \quad \text{for } \mathit{\Gamma}^{(r\oplus 1)} \in \mathit{\Gamma}^{\mathrm{V}} \cup \mathit{\Gamma}^{\mathrm{VIII}} \cup \mathit{\Gamma}^{\mathrm{IX}}.
\end{aligned}$$

Input flow $\mathit{\Pi}_5$ is defined as a marked point process $\{(\tau_i, \eta'_{5,i}, \nu'_i); i = 0, 1, \ldots\}$ with a mark $\nu'_i = (\kappa_{2,i}, \eta_{2,i}, \xi_{2,i})$. Put $\psi(k; x, u) = C_x^k u^k (1-u)^{x-k}$ for $0 < u \le 1$ $0 \le k \le x$. Then for $0 \le b \le x$

$$\mathbf{P}(\{\omega\colon \eta'_{5,i} = b\} | \{\omega\colon \kappa_{2,i} = x, \eta_{2,i} = u, \xi_{2,i} = y\}) = \psi(b; \min\{y, x+u\}, \alpha).$$

Output flow Π_5^{out} is defined as a marked point process $\{(\tau_i, \bar{\xi}_{5,i}, \nu_i''); i = 0, 1, \ldots\}$ with a mark $\nu_i'' = (\Gamma_i, \kappa_{5,i})$ and conditional probability distributions

$$\mathbf{P}(\{\omega\colon \bar{\xi}_{5,i} = b\}|\{\omega\colon \Gamma_i = \Gamma^{(r)}, \kappa_{5,i} = x\}) = \psi(b; x, p_{r\oplus 1}) \qquad \text{for } 0 \le b \le x,$$

For non-negative integers i, x_1, x_2, x_3, x_4, x_5, $r = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, d$, introduce events

$$\begin{gathered} A_i(r, k, x_1, x_2, x_3, x_4, x_5) = \{\omega\colon \Gamma_i = \Gamma^{(r)}, \chi_i = e^{(k)}\} \\ \cap\{\omega\colon \kappa_{1,i} = x_1, \kappa_{2,i} = x_2, \kappa_{3,i} = x_3, \kappa_{4,i} = x_4, \kappa_{5,i} = x_5\}, \\ B_i(b_1, b_2, b_3, b_4, y_1, y_2, y_3, y_4, y_5) = \{\omega\colon \eta_{1,i} = b_1, \eta_{2,i} = b_2\} \\ \cap\{\omega\colon \eta_{4,i} = b_4, \xi_{1,i} = y_1, \xi_{2,i} = y_2, \eta'_{5,i} = b_3, \xi_{3,i} = y_3, \xi_{4,i} = y_4, \bar{\xi}_{5,i} = y_5\}. \end{gathered}$$

The problem statement suggests that

$$\begin{gathered} \mathbf{P}\Big(B_i(b_1, b_2, b_3, b_4, y_1, y_2, y_3, y_4, y_5) \Big| \bigcap_{\iota=0}^{i} A_\iota(r_\iota, k_\iota, x_{1,\iota}, x_{2,\iota}, x_{3,\iota}, x_{4,\iota}, x_{5,\iota})\Big) \\ = \mathbf{P}(\{\omega\colon \eta_{1,i} = b_1\}|\{\omega\colon \Gamma_i = \Gamma^{(r_i)}, \chi_i = e^{(k_i)}\}) \\ \times\mathbf{P}(\{\omega\colon \eta_{2,i} = b_2\}|\{\omega\colon \Gamma_i = \Gamma^{(r_i)}, \chi_i = e^{(k_i)}\}) \\ \times\mathbf{P}(\{\omega\colon \eta_{4,i} = b_4\}|\{\omega\colon \Gamma_i = \Gamma^{(r_i)}, \chi_i = e^{(k_i)}\}) \\ \times\mathbf{P}(\{\omega\colon \xi_{1,i} = y_1\}|\{\omega\colon \Gamma_i = \Gamma^{(r_i)}\})\mathbf{P}(\{\omega\colon \xi_{2,i} = y_2\}|\{\omega\colon \Gamma_i = \Gamma^{(r_i)}\}) \\ \times\mathbf{P}(\{\omega\colon \xi_{3,i} = y_3\}|\{\omega\colon \Gamma_i = \Gamma^{(r_i)}\})\mathbf{P}(\{\omega\colon \xi_{4,i} = y_4\}|\{\omega\colon \Gamma_i = \Gamma^{(r_i)}\}) \\ \times\mathbf{P}(\{\omega\colon \eta'_{5,i} = b_3\}|\{\omega\colon \kappa_{2,i} = x_2, \eta_{2,i} = b_2, \xi_{2,i} = y_2\}) \\ \times\mathbf{P}(\{\omega\colon \bar{\xi}_{5,i} = y_5\}|\{\omega\colon \Gamma_i = \Gamma^{(r_i)}, \kappa_{3,i} = x_5\}). \end{gathered}$$

The remaining blocks of the system, such as the server, output flows, queues' sizes are defined by functional dependencies

$$\begin{gathered} \bar{\xi}_{j,i} = \min\{\xi_{j,i}, \kappa_{j,i} + \eta_{j,i}\}, \quad \kappa_{j,i+1} = \max\{0, \kappa_{j,i} + \eta_{j,i} - \xi_{j,i}\} \quad j = 1, 2, 3, 4; \\ \Gamma_{i+1} = u(\Gamma_i),\ \eta_{3,i} = \bar{\xi}_{5,i},\ \eta_{5,i} = \bar{\xi}_{1,i} + \eta'_{5,i},\ \kappa_{5,i+1} = \kappa_{5,i} + \bar{\xi}_{1,i} + \bar{\xi}_{2,i} - \xi_{5,i}. \end{gathered} \tag{1}$$

3 Analysis of the Model

In the remainder of this paper we assume that Markov chain $\{\chi_i; i = 0, 1, \ldots\}$ is irreducible and aperiodic, and probability generating functions $q_{j,k}(z; T_r)$ are analytic in an open disk $|z| < 1 + \varepsilon$. Denote by $w = (w_1, w_2, w_3, w_4, w_5)$ an arbitrary element of the non-negative integer lattice $X = \{0, 1, \ldots\} \times \ldots \times \{0, 1, \ldots\}$, let

$$S_1' = \{(\Gamma^{(r)}, e^{(k)}, w)\colon \Gamma^{(r)} \in \Gamma^{\mathrm{II}} \cup \Gamma^{\mathrm{VI}} \cup \Gamma^{\mathrm{VIII}}, e^{(k)} \in E, \\ w \in X, w_1 > 0, w_5 < \ell_{r,1}\},$$

$$S_1 = \{(\gamma, e^{(k)}, w)\colon \gamma \notin \Gamma^{\mathrm{II}} \cup \Gamma^{\mathrm{VI}} \cup \Gamma^{\mathrm{VIII}}, e^{(k)} \in E, w \in X\},$$

$$S_2 = \{(\gamma, e^{(k)}, w)\colon \gamma \in \Gamma^{\mathrm{II}} \cup \Gamma^{\mathrm{VI}} \cup \Gamma^{\mathrm{VIII}}, e^{(k)} \in E, w \in X, w_1 = 0\},$$

$$S_3 = \{(\Gamma^{(r)}, e^{(k)}, w)\colon \Gamma^{(r)} \in \Gamma^{\mathrm{II}} \cup \Gamma^{\mathrm{VI}} \cup \Gamma^{\mathrm{VIII}}, e^{(k)} \in E, \\ w \in X, w_1 > 0, w_5 \geq \ell_{r,1}\}.$$

Recurrent equations (1) and the properties of conditional probabilities mentioned above permit to prove next statement.

Theorem 1. *The law of probability distribution for the vector*

$$(\Gamma_0, \chi_0, \kappa_{1,0}, \kappa_{2,0}, \kappa_{3,0}, \kappa_{4,0}, \kappa_{5,i})$$

given, the sequence

$$\{(\Gamma_i, \chi_i, \kappa_{1,i}, \kappa_{2,i}, \kappa_{3,i}, \kappa_{4,i}, \kappa_{5,i}); i = 0, 1, \ldots\} \tag{2}$$

is a Markov chain. The state space $\Gamma \times E \times X$ of Markov chain (2) is a union of unclosed set S_1' of nonessential states and closed set $S_1 \cup S_2 \cup S_3$ of essential periodic states with period n.

Let ${}^j\Gamma = \Gamma^{\mathrm{II}} \cup \Gamma^{\mathrm{VI}} \cup \Gamma^{\mathrm{VIII}}$ for $j = 1$, ${}^j\Gamma = \Gamma^{\mathrm{III}} \cup \Gamma^{\mathrm{VII}} \cup \Gamma^{\mathrm{IX}}$ for $j = 2$, and ${}^j\Gamma = \Gamma^{\mathrm{V}} \cup \Gamma^{\mathrm{VIII}} \cup \Gamma^{\mathrm{IX}}$ for $j = 4$. Denote by

$$\bar{\lambda}_j^{(k)} = \lambda_j^{(k)} \sum_{b=1}^{\infty} b\pi(b; j, k)$$

the expected number of arrivals from Π_j in environment state $e^{(k)}$ per time unit, A_k the stationary probability of the environment state $e^{(k)}$, $k = 1, 2, \ldots, d$ and put $\ell_j = \sum_{r \in {}^j\Gamma} \ell_{r,j}$. Then we have the following theorem.

Theorem 2. *Let $j = 1, 2, 4$. For expected sizes $\mathbf{E}\kappa_{j,i}$ of the queue O_j, $i = 0, 1, \ldots,$ to be bounded it is sufficient that the next inequality holds*

$$T \sum_{k=1}^{d} A_k \bar{\lambda}_j^{(k)} - \ell_j < 0. \tag{3}$$

Proof. Notice that the sequence

$$\{(\Gamma_i, \kappa_{j,i}, \chi_i); i = 0, 1, \ldots\} \tag{4}$$

is an irreducible periodic Markov chain as well as (2). Define the corresponding marginal laws of probability distribution

$$Q_{j,i}(r, k, x_j) = \sum Q_i(r, k, x_1, x_2, x_3, x_4, x_5),$$

where summation is done with respect to all variables $x_s = 0, 1, \ldots, 5$ with index $s \neq j$. For probability generating functions

$$\Psi_{j,i}(z;r,k) = \sum_{x_j=0}^{\infty} z^{x_j} Q_{j,i}(r,k,x_j) = \mathbf{E}\left(z^{\kappa_{j,i}} I(\{\omega\colon \Gamma_i = \Gamma^{(r)}, \chi_i = e^{(k)}\})\right)$$

one can establish the following recurrent equations, $i = 0, 1, \ldots$:

$$\Psi_{j,i+1}(z;r\oplus 1,l) = \sum_{k=1}^{d} a_{k,l} q_{j,k}(z;T_{r\oplus 1})\Psi_{j,i}(z;r,k), \quad \Gamma^{(r\oplus 1)} \notin {}^{j}\Gamma, \tag{5}$$

$$\Psi_{j,i+1}(z;r\oplus 1,l) = \sum_{k=1}^{d} a_{k,l} q_{j,k}(z;T_{r\oplus 1}) z^{-\ell_{r\oplus 1,j}} \Psi_{j,i}(z;r,k) + \sum_{k=1}^{d} a_{k,l}$$
$$\times \sum_{x_j=0}^{\ell_{r\oplus 1,j}-1} Q_{j,i}(r,k,x_j) \sum_{b=0}^{\ell_{r\oplus 1,j}-x_j} \varphi_{j,k}(b;T_{r\oplus 1})(1 - z^{x_j+b-\ell_{r\oplus 1,j}}), \ \Gamma^{(r\oplus 1)} \in {}^{j}\Gamma. \tag{6}$$

Equations (5), (6) show that the series $\Psi_{j,i}(z;r,k)$ converge inside the disk $|z| < 1+\varepsilon$ for all $i = 0, 1, \ldots$. Choose $z \in (1, 1+\varepsilon)$ and a positive integer g. Substitute index $i+1$ with $i+ng$, and i with $i+ng-1$ in equations (5), (6), then sum the obtained equations for k, r. Then apply equations equations (5), (6) again with change of index $i-1$ to $i+ng-1$, and so on. We get

$$\sum_{l=1}^{d} \Psi_{j,i+gn}(z;r,l) = \sum_{(k_1,k_2,\ldots,k_{gn})\in E^{gn}} a_{k_1,k_2} a_{k_2,k_3} \times \ldots \times a_{k_{gn-2},k_{gn-1}} z^{-g\ell_j}$$
$$\times q_{j,k_1}(z;T_{r\oplus 1}) q_{j,k_2}(z;T_{r\oplus 2}) \times \ldots \times q_{j,k_{gn}}(z;T_{r\oplus(gn)})\Psi_{j,i}(z;r,k_1) + B_{j,i}(z;r), \tag{7}$$

where the term $B_{j,i}(z;r) \geq 0$ includes probabilities $Q_{j,i}(r,k,x_j)$, $Q_{j,i+1}(r,k,x_j)$, $\ldots$, $Q_{j,i+gn-1}(r,k,x_j)$ only for $x_j = 0, 1, \ldots, \max\{\ell_{r,j}\colon \Gamma^{(r)} \in {}^{j}\Gamma\}$. Therefore it is possible to give an upper bound $\hat{B}_j(z;r) > 0$ independent of i. So we have

$$\sum_{l=1}^{d} \Psi_{j,i+gn}(z;r,l) \leq \sum_{(k_1,k_2,\ldots,k_{gn})\in E^{gn}} a_{k_1,k_2} \times \ldots \times a_{k_{gn-2},k_{gn-1}} z^{-g\ell_j}$$
$$\times q_{j,k_1}(z;T_{r\oplus 1}) \times \ldots \times q_{j,k_{gn}}(z;T_{r\oplus(gn)})\Psi_{j,i}(z;r,k_1) + \hat{B}_j(z;r), \tag{8}$$

The derivative at $z = 1$ of the multiplier in front of $\Psi_{j,i}(z;r,k_1)$ equals

$$\sum_{k=1}^{d} \bar{\lambda}_j^{(k)} \Big(T_{r\oplus 1}(a_{k_1,k}^{(0)} + a_{k_1,k}^{(n)} + \ldots + a_{k_1,k}^{(gn-g)}) + T_{r\oplus 2}(a_{k_1,k}^{(1)} + a_{k_1,k}^{(n+1)}$$
$$\ldots + a_{k_1,k}^{(gn-g+1)}) + \ldots + T_{r\oplus n}(a_{k_1,k}^{(n-1)} + a_{k_1,k}^{(2n-1)} + \ldots + a_{k_1,k}^{(gn-1)})\Big) - g\ell_j. \tag{9}$$

As $g \to \infty$ the Cesaro means converge [6],

$$g^{-1}(a_{k_1,k}^{(s)} + a_{k_1,k}^{(n+s)} + \ldots + a_{k_1,k}^{(gn-g+s)}) \to A_k, \qquad s = 0, 1, \ldots, n-1.$$

Hence for g large enough the sign of the derivative (9) coincides with that of (3). Since for $z = 1$

$$\sum_{(k_2,\ldots,k_{gn})\in E^{gn-1}} a_{k_1,k_2}a_{k_2,k_3}\times\ldots\times a_{k_{gn-2},k_{gn-1}}z^{-g\ell_j}$$
$$\times q_{j,k_1}(z;T_{r\oplus 1})q_{j,k_2}(z;T_{r\oplus 2})\times\ldots\times q_{j,k_{gn}}(z;T_{r\oplus(gn)})=1,$$

and the derivative (9) is negative, there exists z, $0<z<1$, such that

$$R_+=\max_{1\le k_1\le d}\Bigg\{\sum_{(k_2,\ldots,k_{gn})\in E^{gn-1}} a_{k_1,k_2}a_{k_2,k_3}\times\ldots\times a_{k_{gn-2},k_{gn-1}}z^{-g\ell_j}$$
$$\times q_{j,k_1}(z;T_{r\oplus 1})q_{j,k_2}(z;T_{r\oplus 2})\times\ldots\times q_{j,k_{gn}}(z;T_{r\oplus(gn)})\Bigg\}<1.$$

Define further a sequence by

$$\Psi^+_{j,0}=\sum_{k=1}^{d}\sum_{r=1}^{n}\Psi_{j,0}(z;r,k),\quad \Psi^+_{j,1}=\sum_{k=1}^{d}\sum_{r=1}^{n}\Psi_{j,1}(z;r,k),$$
$$\ldots,\quad \Psi^+_{j,ng-1}=\sum_{k=1}^{d}\sum_{r=1}^{n}\Psi_{j,ng-1}(z;r,k),$$
$$\Psi^+_{j,i+gn}=R_+\Psi_{j,i}+\hat{B}_j(z;r),\qquad i=0,1,\ldots.$$

The sequence $\{\Psi^+_i;i=0,1,\ldots\}$ thus defined is convergent, so it is bounded by some constant M. At the same time, $\Psi_i(z;r,k)\le\Psi^+_i\le M$. Finally, the Cauchy's integral formula

$$\mathbf{E}\kappa_{j,i}=\Big|\frac{1}{2\pi\sqrt{-1}}\int_{|z-1|=\rho}(z-1)^{-2}\sum_{k=1}^{d}\sum_{r=1}^{n}\Psi_{j,i}(z;r,k)\,dz\Big|\le\frac{ndM}{\rho}$$

ensures that the sequence $\mathbf{E}\kappa_{j,i}$, $i=0,1,\ldots$ is also bounded.

Theorem 3. *Let* $\kappa_{1,0}=\kappa_{2,0}=\ldots=\kappa_{5,0}=0$. *The sequence* $\{\mathbf{E}\kappa_{5,i};i=0,1,\ldots\}$ *is bounded.*

Proof. Consider probability generating functions

$$\Psi_i(z_1,z_2,z_3;r,k)=\sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty}\sum_{x_3=0}^{\infty}z_1^{x_1}z_2^{x_2}z_3^{x_3}Q_i(r,k,x_1,x_2,x_3),$$

which hold information about joint probability distribution

$$Q_i(r,k,x_1,x_2,x_3)=\mathbf{P}(\{\omega\colon \Gamma_i=\Gamma^{(r)},\kappa_{1,i}=x_1,\kappa_{2,i}=x_2,\kappa_{5,i}=x_5,\chi_i=e^{(k)}\})$$

of $\kappa_{1,i}$, $\kappa_{2,i}$, $\kappa_{5,i}$, Γ_i, and χ_i. Then for $\Gamma^{(r\oplus 1)} \in \Gamma^{\mathrm{I}}$

$$\Psi_{i+1}(z_1, z_2, z_3; r \oplus 1, l) = \sum_{k=1}^{d} a_{k,l} q_{1,k}(z_1, T_{r\oplus 1,1}) q_{2,k}(z_2, T_{r\oplus 1,1}) \times \Psi_i(z_1, z_2, p_{r\oplus 1} + (1 - p_{r\oplus 1}) z_3; r, k),$$

for $\Gamma^{(r\oplus 1)} \in \Gamma^{\mathrm{II}} \cup \Gamma^{\mathrm{VI}} \cup \Gamma^{\mathrm{VIII}}$

$$\Psi_{i+1}(z_1, z_2, z_3; r \oplus 1, l) = \left(\frac{z_3}{z_1}\right)^{\ell_{r\oplus 1,1}} \sum_{k=1}^{d} a_{k,l} q_{1,k}(z_1, T_{r\oplus 1,1}) q_{2,k}(z_2, T_{r\oplus 1,1}) \times \Psi_i(z_1, z_2, p_{r\oplus 1} + (1 - p_{r\oplus 1}) z_3; r, k) + \sum_{k=1}^{d} \sum_{x_1=0}^{\ell_{r\oplus 1,1}-1} \sum_{x_2=0}^{\infty} \sum_{x_3=0}^{\infty} Q_i(r, k, x_1, x_2, x_3) \times a_{k,l} z_2^{x_2} (p_{r\oplus 1} + (1 - p_{r\oplus 1}) z_3)^{x_3} q_{2,k}(z_2, T_{r\oplus 1}) \sum_{b=0}^{\ell_{r\oplus 1,1}-x_1-1} \varphi_{1,k}(b; T_{r\oplus 1}) \times (z_3^{x_1+b} - z_1^{x_1+b-\ell_{r\oplus 1,1}} z_3^{\ell_{r\oplus 1,1}}),$$

for $\Gamma^{(r\oplus 1)} \in \Gamma^{\mathrm{III}} \cup \Gamma^{\mathrm{VII}} \cup \Gamma^{\mathrm{IX}}$

$$\Psi_{i+1}(z_1, z_2, z_3; r \oplus 1, l) = \left(\frac{1 - \alpha + \alpha z_3}{z_1}\right)^{\ell_{r\oplus 1,2}} \sum_{k=1}^{d} a_{k,l} q_{1,k}(z_1, T_{r\oplus 1,1}) \times q_{2,k}(z_2, T_{r\oplus 1,1}) \Psi_i(z_1, z_2, p_{r\oplus 1} + (1 - p_{r\oplus 1}) z_3; r, k) + \sum_{k=1}^{d} a_{k,l} \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\ell_{r\oplus 1,2}-1} \sum_{x_3=0}^{\infty} Q_i(r, k, x_1, x_2, x_3) \times z_1^{x_1} (p_{r\oplus 1} + (1 - p_{r\oplus 1}) z_3)^{x_3} q_{1,k}(z_1, T_{r\oplus 1}) \sum_{b=0}^{\ell_{r\oplus 1,2}-x_2-1} \varphi_{2,k}(b; T_{r\oplus 1}) \times ((1 - \alpha + \alpha z_3)^{x_2+b} - z_2^{x_2+b-\ell_{r\oplus 1,2}} (1 - \alpha + \alpha z_3)^{\ell_{r\oplus 1,2}}).$$

Since $|p_{r\oplus 1} + (1 - p_{r\oplus 1}) z_3| < 1 + \varepsilon$ for $1 < z_3 < 1 + \varepsilon$, these equations determine functions $\Psi_i(z_1, z_2, z_3; r, k)$ analytic in the polydisk $\{|z_1| < 1 + \varepsilon, |z_2| < 1 + \varepsilon, |z_3| < 1 + \varepsilon\}$. Thus the derivatives

$$m_i(r) = \sum_{k=1}^{d} \frac{\partial}{\partial z_3} \Psi_i(1, 1, z_3; r, k)\big|_{z_3=1}$$

exist and satisfy equations

$$m_{i+1}(r\oplus 1)=(1-p_{r\oplus 1})m_i(r),\quad \Gamma^{(r\oplus 1)}\in\Gamma^{\mathrm{I}},$$

$$m_{i+1}(r\oplus 1)=(1-p_{r\oplus 1})m_i(r)+\ell_{r\oplus 1,1}\sum_{k=1}^{d}\Psi_i(1,1,1;r,k)$$
$$+\sum_{k=1}^{d}\sum_{x_1=0}^{\ell_{r\oplus 1,1}-1}\sum_{x_2=0}^{\infty}\sum_{x_3=0}^{\infty}Q_i(r,k,x_1,x_2,x_3)$$
$$\times\Bigg(\sum_{b=0}^{\ell_{r\oplus 1,1}-x_1-1}\varphi_{1,k}(b;T_{r\oplus 1})(x_1+b-\ell_{r\oplus 1,1})\Bigg),\quad \Gamma^{(r\oplus 1)}\in\Gamma^{\mathrm{II}}\cup\Gamma^{\mathrm{VI}}\cup\Gamma^{\mathrm{VIII}},$$

$$m_{i+1}(r\oplus 1)=(1-p_{r\oplus 1})m_i(r)+\alpha\ell_{r\oplus 1,2}\sum_{k=1}^{d}\Psi_i(1,1,1;r,k)$$
$$+\sum_{k=1}^{d}\sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\ell_{r\oplus 1,2}-1}\sum_{x_3=0}^{\infty}Q_i(r,k,x_1,x_2,x_3)$$
$$\times\Bigg(\sum_{b=0}^{\ell_{r\oplus 1,2}-x_2-1}\varphi_{2,k}(b;T_{r\oplus 1})\alpha(x_2+b-\ell_{r\oplus 1,2})\Bigg),\quad \Gamma^{(r\oplus 1)}\in\Gamma^{\mathrm{III}}\cup\Gamma^{\mathrm{VII}}\cup\Gamma^{\mathrm{IX}},$$

and are dominated by convergent sequences $m_i^+(r)=m_i(r)$, $0\le i\le n-1$, $m_i^+(r)=(1-p_1)\times\ldots\times(1-p_n)m_{i-n}^+(r)+\ell_1+\alpha\ell_2$, $i\ge n$. Hence they are bounded. We only have to recall that $\mathbf{E}\kappa_{5,i}=m_i(1)+m_i(2)+\ldots+m_i(n)$.

Theorems 2, 3 show that in tandem intersections the queue of retrial customers is always stable, and queues O_1, O_2, O_4 can be stable independently of each other.

References

1. Litvak, N.V., Fedotkin, M.A.: An adaptive control for conflicting flows: its probabilistic model. Automation and Remote Control 61(5), 67–76 (2000)
2. Litvak, N.V., Fedotkin, M.A.: An adaptive control for conflicting flows: a quantitative and numerical study of its probabilistic model. Automation and Remote Control 61(6), 69–78 (2000)
3. Proidakova, E.V., Fedotkin, M.A.: Control of output flows in the system with cyclic servicing and readjustments. Automation and Remote Control 69(6), 993–1002 (2008)
4. Fedotkin, M.A., Fedotkin, A.M.: Analysis and optimization of output processes of conflicting Gnedenko – Kovalenko traffic streams under cyclic control. Automation and Remote Control 70(12), 2024–2038 (2009)
5. Zorin, A.V.: Stability of a tandem of queueing systems with Bernoulli noninstantaneous transfer of customers. Theory of Probability and Mathematical Statistics 84, 173–188 (2012)
6. Doob, J.L.: Stochastic processes. Wiley, New York (1953)

Author Index