Advances in Intelligent Systems and Computing 383

Tien Van Do
Yutaka Takahashi
Wuyi Yue
Viet-Ha Nguyen *Editors*

Queueing Theory and Network Applications

Springer

Advances in Intelligent Systems and Computing

Volume 383

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

About this Series

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within "Advances in Intelligent Systems and Computing" are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Advisory Board

More information about this series at http://www.springer.com/series/11156

Tien Van Do · Yutaka Takahashi
Wuyi Yue · Viet-Ha Nguyen
Editors

Queueing Theory and Network Applications

Springer

Editors
Tien Van Do
Department of Networked Systems and Services
Budapest University of Technology and Economics
Budapest
Hungary

Yutaka Takahashi
Department of Systems Science
Kyoto University Graduate School of Informatics
Kyoto
Japan

Wuyi Yue
Faculty of Science and Engineering
Department of Information Science and Systems Engineering
Konan University
Kobe
Japan

Viet-Ha Nguyen
Faculty of Information Technology
VNU University of Engineering and Technology
Hanoi
Vietnam

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-3-319-22266-0 ISBN 978-3-319-22267-7 (eBook)
DOI 10.1007/978-3-319-22267-7

Library of Congress Control Number: 2015946102

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

This volume contains papers presented at the 10th International Conference on Queueing Theory and Network Applications (QTNA2015) held on 17–20 August, 2015 in Ha Noi and Ha Long, Vietnam. The conference is co-organized by Analysis, Design and Development of ICT systems (AddICT) Laboratory, Budapest University of Technology and Economics, Hungary, Vietnam National University, University of Engineering and Technology (VNU-UET) and Ha Long University.

The conference is a continuation of the series of successful QTNA conferences - QTNA2006 (Seoul, Korea), QTNA2007 (Kobe, Japan), QTNA2008 (Taipei, Taiwan), QTNA2009 (Singapore), QTNA2010 (Beijing, China), QTNA2011 (Seoul, Korea), QTNA2012 (Kyoto, Japan), QTNA2013 (Taichung, Taiwan) and QTNA2014 (Bellingham, USA).

The QTNA2015 conference is to promote the knowledge and the development of high-quality research on queueing theory and its applications in networks and other related fields. It brings together researchers, scientists and practitioners from the world and offers an open forum to share the latest important research accomplishments and challenging problems in the area of queueing theory and network applications.

The clear message of the proceedings is that the potentials of queueing theory are to be exploited, and this is an opportunity and a challenge for researchers. The intensive discussions have seeded future exciting applications. The works included in this proceedings can be useful for researchers, Ph.D. and graduate students in queueing theory. It is the hope of the editors that readers can find many inspiring ideas and use them to their research. Many such challenges are suggested by particular approaches and models presented in the proceedings.

We would like to thank all authors, who contributed to the success of the conference and to this book. Special thanks go to the members of Program Committees for their contributions to keeping the high quality of the selected papers. We would like to thank Dr. Vu Thi Thu Thuy (rector) and Dr. Bui Van Tan (vice-rector) of Ha Long University, who invited us to have sessions in Ha Long university. A special appreciation goes to the People's Committee of Quảng Ninh

Province and the President Board of Vietnam National University, Hanoi for their generous support. Cordial thanks are due to the Organizing Committee members for their efforts and the organizational work. Finally, we cordially thank Springer for supports and publishing this volume.

August 2015

Tien Van Do
Yutaka Takahashi
Wuyi Yue
Viet-Ha Nguyen

QTNA 2015 Organization

Honorary Chair

Viet Ha Nguyen	Vietnam National University, University of Engineering and Technology, Vietnam

General Chairs

Tien Van Do	Budapest University of Technology and Economics, Hungary
Yutaka Takahashi	Kyoto University, Japan
Nguyen Thanh Thuy	Vietnam National University, University of Engineering and Technology, Vietnam
Vu Thi Thu Thuy	Ha Long University, Vietnam
Bui Van Tan	Ha Long University, Vietnam

Program Chairs

Tien Van Do	Budapest University of Technology and Economics, Hungary
Yutaka Takahashi	Kyoto University, Japan
Wuyi Yue	Konan University, Japan

Local Organizing Committee

Tien Van Do	Budapest University of Technology and Economics, Hungary
Nam H. Do	Budapest University of Technology and Economics, Hungary

Pham Bao Son	Vietnam National University, University of Engineering and Technology, Vietnam
Tran Xuan Tu	Vietnam National University, University of Engineering and Technology, Vietnam
Le Anh Cuong	Vietnam National University, University of Engineering and Technology, Vietnam
Ha Quang Thuy	Vietnam National University, University of Engineering and Technology, Vietnam
Vu Duc Thi	Vietnam National University, University of Engineering and Technology, Vietnam
Nguyen Dai Tho	Vietnam National University, University of Engineering and Technology, Vietnam
Vu Anh Dung	Vietnam National University, University of Engineering and Technology, Vietnam
Tran Truc Mai	Vietnam National University, University of Engineering and Technology, Vietnam
Nguyen Hoai Son	Vietnam National University, University of Engineering and Technology, Vietnam
Tran Thi Thu Ha	Vietnam National University, University of Engineering and Technology, Vietnam
Le Dinh Thanh	Vietnam National University, University of Engineering and Technology, Vietnam
Nguyen Ngoc Hoa	Vietnam National University, University of Engineering and Technology, Vietnam

Steering Committee

Bong Dae Choi	Sungkyunkwan University, Korea
Yutaka Takahashi	Kyoto University, Japan
Wuyi Yue	Konan University, Japan
Hsing Paul Luh	National Chengchi University, Taiwan
Winston K.G. Seah	Victoria University of Wellington, New Zealand
Hideaki Takagi	Japan
Y.C. Tay	Singapore
Kuo-Hsiung Wang	Providence University, Taiwan
Jinting Wang	China
Deguan Yue	China
Zhe George Zhang	Western Washington University, USA

Program Committee

Sergey Andreev	Finland
Tien Van Do	Hungary, Vietnam
Qi-Ming He	Canada
Ganguk Hwang	Korea
Shoji Kasahara	Japan
Konosuke Kawashima	Japan
Bara Kim	Korea
Masahiro Kobayashi	Japan
Ho Woo Lee	Korea
Se Won Lee	Korea
Hiroyuki Masuyama	Japan
Agassi Melikov	Azerbaijan
Yoni Nazarathy	Australia
Yoshikuni Onozato	Japan
Tuan Phung-Duc	Japan
Wouter Rogiest	Belgium
Poompat Saengudomlert	Thailand
Zsolt Saffer	Hungary
Yutaka Sakuma	Japan
Winston Seah	New Zeland
Yang Woo Shin	Korea
Janos Sztrik	Hungary
Hideaki Takagi	Japan
Yutaka Takahashi	Japan
Y.C. Tay	Singapore
Jinting Wang	China
Sabine Wittevrongel	Belgium
Dequan Yue	China
Wuyi Yue	Japan
Yigiang Q. Zhao	Canada

Contents

Part I
Queueing Models I

Detailed Analysis of the Response Time and Waiting Time in the M/M/*m* FCFS Preemptive-Resume Priority Queue

Hideaki Takagi

Abstract We present a detail theoretical analysis of the response time and waiting time in the M/M/*m* FCFS preemptive-resume priority queueing system in the steady state by scrutinizing and extending the previous studies by Brosh (1969), Segal (1970), Buzen and Bondi (1983), Tatashev (1984), and Zeltyn et al. (2009). In particular, we analyze the durations of intermittent waiting times and service times during the response time of a tagged customer of each priority class that is preempted by the arrivals of higher-priority class customers. Numerical examples are shown in order to demonstrate the computation of theoretical formulas.

Keywords Priority queue · Multiserver · Preemptive-resume · Response time · Waiting time · First passage time

1 Introduction

We consider a queueing system with m servers and an infinite capacity of the waiting room with several priority classes of customers. Customers of class p arrive in a Poisson process with rate λ_p (> 0) independently of customers of all other classes. Every customer requests a service which has the exponential distribution with mean $1/\mu$ irrespective of his class. Classes are indexed $1, 2, \ldots$ such that customers of class p have preemptive priority for service over customers of class q if $p < q$.

There are three cases which may happen when a customer of class p arrives:

- Unless all servers are busy, his service is started immediately.
- If all servers are busy serving customers of classes not lower than p, he must wait at the tail of waiting customers of class p.

H. Takagi (✉)
Professor Emeritus, University of Tsukuba, Tsukuba Science City, Ibaraki 305-8573, Japan
e-mail: takagi@sk.tsukuba.ac.jp

T.V. Do et al. (eds.), *Queueing Theory and Network Applications*,
Advances in Intelligent Systems and Computing 383,
DOI: 10.1007/978-3-319-22267-7_1

– If all servers are busy serving customers, out of whom at least one of them is of class lower than p. Let q $(> p)$ be the lowest priority class of those customers being served. At this moment there are at most customers of classes $q, q+1, \ldots$ in the waiting room. In this case, the service to one of customers of class q is preempted and he is displaced from the service facility to the head of the waiting room. We select such a customer of class q for displacement that his service was started or resumed last among all the customers of class q in service. Then the service to the arriving customer of class p is started. This policy of selecting the customer to displace is assumed by Segal [4]. It is called *Last-Come, First-Displaced* (LCFD) by Fujiki [3].

As soon as a server becomes available, one of the customers of the highest priority class among those in the waiting room is called in for service. Within the same class, a customer is chosen on the first-come, first-served (FCFS) basis. When the service is resumed, a new sample of the service time is set up from the exponential distribution with mean $1/\mu$, irrespective of the amount of service given to him previously.

Thus we may call our system an "M/M/m preemptive-resume priority queue with FCFS and LCFD within the same class." The study of response times of customers in this model dates back to old days including Brosh [1], Segal [4], Buzen and Bondi [2], Tatashev [5], and Zeltyn et al. [6]. The purpose of this paper is to derive explicit formulas for the mean and second moment of the response time of a tagged customer of each priority class in the steady state.

We use the following notation for the analysis throughout this paper.

$$\rho_p := \frac{\lambda_p}{m\mu} \quad ; \quad \lambda_p^+ := \sum_{k=1}^{p} \lambda_k \quad ; \quad \rho_p^+ := \sum_{k=1}^{p} \rho_k = \frac{\lambda_p^+}{m\mu} \qquad p = 1, 2, \ldots$$

In the numerical examples in this paper, we assume that there are 4 classes of customers and that

$$m = 5 \quad ; \quad \mu = 1 \quad ; \quad \lambda_p = \frac{\lambda}{4} \quad (1 \le p \le 4).$$

For this setting, we will show several performance measures against λ for the range $0 \le \lambda \le 20$. Our formulas can be applied to systems with any number of servers, any number of classes, and any different distinct values of arrival rates. However we must assume that the service rates are identical for all customers of all classes and that the system is stable up to customers of class p $(\rho_p^+ < 1)$.

2 Mean Response Time and Mean Waiting Time

We first follow Buzen and Bondi [2] for the neat derivation of mean response time $E[T^{(p)}]$ for customers of each class p. Let us focus on customers of class p. Due to the service and preemption mechanism mentioned above, the behavior of a customer is never affected by customers of lower priority classes as well as customers of the

same class who arrive after him. Therefore, we have only to consider customers of classes $1, 2, \ldots, p$.

We denote by N_p^+ the number of customers of classes $1, 2, \ldots, p$ present in the system at an arbitrary time in the steady state and define

$$Q_{p,k}^+ := P\{N_p^+ = k\} \qquad k = 0, 1, 2, \ldots.$$

From the well-known analysis for the M/M/m queue with customers of classes $1, 2, \ldots, p$, we get

$$Q_{p,k}^+ = \begin{cases} Q_{p,0}^+ \dfrac{(m\rho_p^+)^k}{k!} & 1 \le k \le m, \\ Q_{p,m}^+ (\rho_p^+)^{k-m} & k \ge m+1, \end{cases} \tag{1}$$

where, from the normalization condition $\sum_{k=0}^{\infty} Q_{p,k}^+ = 1$, we have

$$\frac{1}{Q_{p,0}^+} = \sum_{k=0}^{m-1} \frac{(m\rho_p^+)^k}{k!} + \frac{(m\rho_p^+)^m}{m!(1-\rho_p^+)},$$

where we assume that $\rho_p^+ < 1$ for the system to be stable. Then we get

$$E[N_p^+] = \sum_{k=1}^{\infty} k Q_{p,k}^+ = m\rho_p^+ + \frac{\rho_p^+ C(m, m\rho_p^+)}{1-\rho_p^+},$$

where

$$C(m,a) := \frac{a^m}{m!} \Bigg/ \left[\left(1 - \frac{a}{m}\right) \sum_{k=0}^{m-1} \frac{a^k}{m!} + \frac{a^m}{m!} \right] \tag{2}$$

is the *Erlang's C formula*. In the present case, we have

$$C(m, m\rho_p^+) = \sum_{k=m}^{\infty} Q_{p,k}^+ = \frac{Q_{p,m}^+}{1-\rho_p^+} = \frac{Q_{p,0}^+}{1-\rho_p^+} \cdot \frac{(m\rho_p^+)^m}{m!}$$

as the probability that a customer of class p waits upon arrival.

We denote by N_p the number of customers of class p present in the system at an arbitrary time in the steady state. Then we get

$$E[N_p] = E[N_p^+] - E[N_{p-1}^+] = \frac{\lambda_p}{\mu} + \frac{\rho_p^+ C(m, m\rho_p^+)}{1-\rho_p^+} - \frac{\rho_{p-1}^+ C(m, m\rho_{p-1}^+)}{1-\rho_{p-1}^+}.$$

From Little's theorem $E[N_p] = \lambda_p E[T_p]$ for customers of class p, we obtain [2]

$$E[T_p] = \frac{E[N_p]}{\lambda_p} = \frac{1}{\mu} + \frac{\rho_p^+ C(m, m\rho_p^+)}{\lambda_p(1-\rho_p^+)} - \frac{\rho_{p-1}^+ C(m, m\rho_{p-1}^+)}{\lambda_p(1-\rho_{p-1}^+)}. \tag{3}$$

We denote by L_p the number of customers of class p present in the waiting room at an arbitrary time in the steady state. Then we have

$$E[L_p] = \frac{\rho_p^+ C(m, m\rho_p^+)}{1-\rho_p^+} - \frac{\rho_{p-1}^+ C(m, m\rho_{p-1}^+)}{1-\rho_{p-1}^+},$$

which gives the mean waiting time [6]

$$E[W_p] = \frac{E[L_p]}{\lambda_p} = \frac{\rho_p^+ C(m, m\rho_p^+)}{\lambda_p(1-\rho_p^+)} - \frac{\rho_{p-1}^+ C(m, m\rho_{p-1}^+)}{\lambda_p(1-\rho_{p-1}^+)} = E[T_p] - \frac{1}{\mu}. \tag{4}$$

We plot $E[W_p]$ and $E[T_p]$ in Figs. 1 and 2, respectively, for the numerical example described in Section 1.

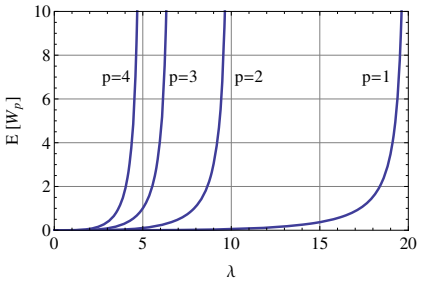


Fig. 1 Mean waiting time for a customer of class p in the M/M/m preemptive-resume priority queue

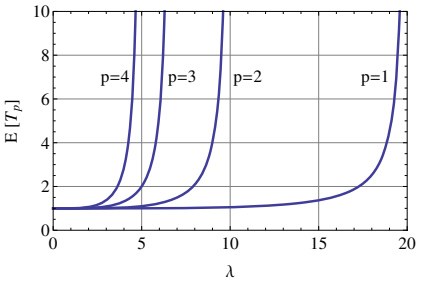


Fig. 2 Mean response for a customer of class p in the M/M/m preemptive-resume priority queue

3 Waiting Time

After a customer of class p enters service for the first time, his service may be preempted several times before completion when he is pushed out of the service facility by the arrivals of customers of classes $1, 2, \ldots, p-1$. He stays in the waiting room until he again enters service. The total amount of the time a customer spends in the waiting room is called the *waiting time*, which is the response time minus the service time.

Tatashev [5] derived the LST of the DF for the waiting time W_p for customers of class p. Later Zeltyn et al. [6] show the mean and the second moment of W_p. Their analysis and result are reviewed in this section.

Let $P_{p,k}\{\Pr\}$ be the probability that a tagged customer of class p competing for the servers with k other customers (they are all customers of classes $1, 2, \ldots, p-1$

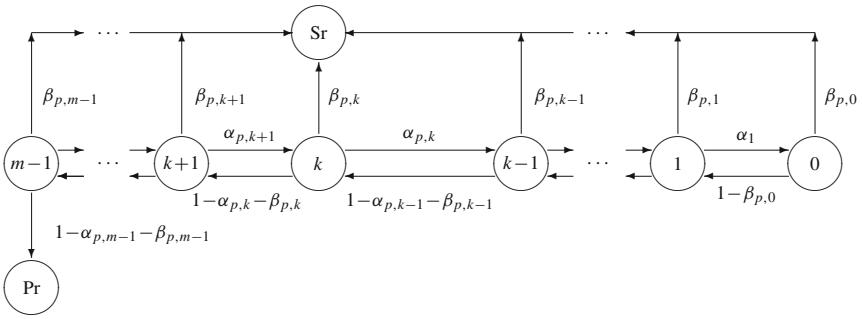


Fig. 3 State transition diagram for a customer of class p until service preemption or completion

and those customers of class p who have arrived before the tagged customer) is preempted, where $k = 0, 1, 2, \ldots .m-1$. The state transition diagram for our tagged customer is shown in Fig. 3. We consider the first passage time in this one-dimensional birth-and-death process with two absorbing states, namely "service preemption" denoted by "Pr" and "service completion" denoted by "Sr".

Referring to Figure 3, we have the complete set of equations for $\{P_{p,k}\{\Pr\}; 0 \leq k \leq m-1\}$ as follows:

$$P_{p,0}\{\Pr\} = (1-\beta_{p,0})P_{p,1}\{\Pr\},$$

$$P_{p,k}\{\Pr\} = (1-\alpha_{p,k}-\beta_{p,k})P_{p,k+1}\{\Pr\} + \alpha_{p,k}P_{p,k-1}\{\Pr\} \qquad 1 \leq k \leq m-2,$$

$$P_{p,m-1}\{\Pr\} = 1-\alpha_{p,m-1}-\beta_{p,m-1}+\alpha_{p,m-1}P_{p,m-2}\{\Pr\},$$

where $\alpha_{p,k}$ and $\beta_{p,k}$ are given by

$$\alpha_{p,k} = \frac{k\mu}{\lambda_{p-1}^{+}+(k+1)\mu} \quad ; \quad \beta_{p,k} = \frac{\mu}{\lambda_{p-1}^{+}+(k+1)\mu} \qquad 0 \leq k \leq m-1.$$

The solution is found to be [5]

$$P_{p,k}\{\Pr\} = \frac{B(m, m\rho_{p-1}^{+})}{B(k, m\rho_{p-1}^{+})} \qquad 0 \leq k \leq m-1 \tag{5}$$

with the well-known *Erlang's B formula*

$$B(m,a) := \frac{a^m}{m!} \Bigg/ \sum_{k=0}^{m} \frac{a^k}{k!} .$$

We note that

$$r_p := P_{p,m-1}\{\mathrm{Pr}\} = \frac{B(m, m\rho_{p-1}^+)}{B(m-1, m\rho_{p-1}^+)} = \rho_{p-1}^+ \left[1 - B(m, m\rho_{p-1}^+)\right]$$

is the probability that the service for a customer of class p started after waiting is preempted (note that $r_1 \equiv 0$). The probability that the service for a customer of class p started without waiting is found as follows. When his service is started there are k other customers of classes $1, 2, \ldots, p$ with probability

$$\frac{(m\rho_p^+)^k}{k!} \Bigg/ \sum_{j=0}^{m-1} \frac{(m\rho_p^+)^j}{j!} \qquad 0 \le k \le m-1.$$

Then his service is preempted with probability $P_{p,k}\{\mathrm{Pr}\}$. Thus we get ($q_1 \equiv 0$)

$$q_p = \sum_{k=0}^{m-1} \frac{B(m, m\rho_{p-1}^+)}{B(k, m\rho_{p-1}^+)} \cdot \frac{(m\rho_p^+)^k/k!}{\sum_{j=0}^{m-1}(m\rho_p^+)^j/j!} = \frac{\rho_{p-1}^+ \left[B(m, m\rho_p^+) - B(m, m\rho_{p-1}^+)\right]}{\rho_p \left[1 - B(m, m\rho_p^+)\right]}.$$

Similarly, let $P_{p,k}\{\mathrm{Sr}\}$ be the probability that a tagged customer of class p competing for the servers with k other customers is completed without preemption, where $k = 0, 1, 2, \ldots .m-1$. This is given by

$$P_{p,k}\{\mathrm{Sr}\} = 1 - P_{p,k}\{\mathrm{Pr}\} = 1 - \frac{B(m, m\rho_{p-1}^+)}{B(k, m\rho_{p-1}^+)} \qquad 0 \le k \le m-1. \tag{6}$$

Then we can numerically confirm the relation

$$\sum_{k=0}^{m-1} Q_{p,k}^+ P_{p,k}\{\mathrm{Sr}\} = [1 - C(m, m\rho_p^+)](1 - q_p)$$

as the probability that an arriving customer of class p is started service immediately upon arrival and his service is not preempted until completion. We can also confirm the relation

$$\sum_{k=0}^{m-1} Q_{p,k}^+ P_{p,k}\{\mathrm{Pr}\} = [1 - C(m, m\rho_p^+)]q_p$$

as the probability that an arriving customer of class p is started service immediately upon arrival and his service is preempted before completion.

We note that $G_{p-1}^*(s)$ is the LST of the DF for the length of a busy period in the M/M/1 queue with arrival rate λ_{p-1}^+ and service rate $m\mu$, which is denoted by $\mathcal{G}_{p-1}^+$. $G_{p-1}^*(s)$ is the solution to the quadratic equation

$$\lambda_{p-1}^{+}[G_{p-1}^{*}(s)]^2 - (s + \lambda_{p-1}^{+} + m\mu)G_{p-1}^{*}(s) + m\mu = 0,$$

which yields the mean and variance of $\mathcal{G}_{p-1}^{+}$:

$$E[\mathcal{G}_{p-1}^{+}] = \frac{1}{m\mu(1-\rho_{p-1}^{+})} \quad ; \quad \mathrm{Var}[\mathcal{G}_{p-1}^{+}] = \frac{1+\rho_{p-1}^{+}}{(m\mu)^2(1-\rho_{p-1}^{+})^3}.$$

Upon arrival of a tagged customer of class p, the following cases occur:

– If less than m servers are busy for serving customers of classes $1, 2, \ldots, p$, his service is started immediately. This case occurs with probability $1 - C(m, m\rho_p^{+})$.
  - If his service is not preempted, his waiting time is zero. This subcase occurs with probability $1 - q_p$.
  - If his service is preempted, he waits $\mathcal{G}_{p-1}^{+}$ time units for his service to be resumed. This subcase occurs with probability q_p. The resumed service is preempted i times with probability $(1 - r_p)(r_p)^i$ $(i = 0, 1, 2, \ldots)$ with each preemption making him wait $\mathcal{G}_{p-1}^{+}$ time units.

– If m or more servers are busy for serving customers of classes $1, 2, \ldots, p$, he waits W_p^{+} time units for his service to be started for the first time. This case occurs with probability $C(m, m\rho_p^{+})$. His service is preempted i times with probability $(1 - r_p)(r_p)^i$ $(i = 0, 1, 2, \ldots)$ with each preemption making him wait $\mathcal{G}_{p-1}^{+}$ time units. We have the LST of the DF for W_p^{+} as

$$W_p^{+}(s) = \frac{(1-\rho_p^{+})G_{p-1}^{*}(s)}{1-\rho_p^{+}G_{p-1}^{*}(s)} = \frac{m\mu(1-\rho_p^{+})[1-G_{p-1}^{*}(s)]}{s-\lambda_p+\lambda_p G_{p-1}^{*}(s)}. \tag{7}$$

Therefore, the LST of the DF for the waiting time of a tagged customer of class p is given by [5, 6]

$$\begin{aligned} W_p^{*}(s) &= [1 - C(m, m\rho_p^{+})]\left\{1 - q_p + q_p G_{p-1}^{*}(s)\sum_{i=0}^{\infty}(1-r_p)(r_p)^i[G_{p-1}^{*}(s)]^i\right\} \\ &\quad + C(m, m\rho_p^{+})W_p^{+}(s)\sum_{i=0}^{\infty}(1-r_p)(r_p)^i[G_{p-1}^{*}(s)]^i \\ &= [1 - C(m, m\rho_p^{+})]\left\{1 - q_p + \frac{q_p(1-r_p)G_{p-1}^{*}(s)}{1-r_p G_{p-1}^{*}(s)}\right\} + C(m, m\rho_p^{+})\frac{(1-r_p)W_p^{+}(s)}{1-r_p G_{p-1}^{*}(s)}. \end{aligned} \tag{8}$$

The mean waiting time for a customer of class p is given by

$$E[W_p] = \frac{[1 - C(m, m\rho_p^{+})]q_p}{m\mu(1-r_p)(1-\rho_{p-1}^{+})} + \frac{C(m, m\rho_p^{+})(1-r_p\rho_p^{+})}{m\mu(1-r_p)(1-\rho_{p-1}^{+})(1-\rho_p^{+})}.$$

We have numerically confirmed that this yields the same result as Eq. (4). The second moment of the waiting time is given by

$$E[W_p^2] = \frac{2[1 - C(m, m\rho_p^+)]q_p(1 - r_p\rho_{p-1}^+)}{(m\mu)^2(1 - r_p)^2(1 - \rho_{p-1}^+)^3}$$

$$+\frac{2C(m, m\rho_p^+)}{(m\mu)^2}\left[\frac{1 - \rho_{p-1}^+\rho_p^+}{(1 - \rho_{p-1}^+)^3(1 - \rho_p^+)^2} + \frac{r_p[2 - \rho_{p-1}^+ - \rho_p^+ - r_p(1 - \rho_{p-1}^+\rho_p)]}{(1 - r_p)^2(1 - \rho_{p-1}^+)^3(1 - \rho_p^+)}\right]. \quad (9)$$

This expression yields the same numerical values as those from the following expression derived by Zeltyn et al. [6]:

$$E[W_p^2] = \frac{2}{(m\mu)^2}\left[\frac{(1 - \rho_{p-1}^+\rho_p^+)C(m, m\rho_p^+)}{(1 - \rho_{p-1}^+)^3(1 - \rho_p^+)^2} + \frac{\rho_{p-1}^+C(m, m\rho_p^+)[1 - C(m, m\rho_{p-1}^+)]}{(1 - \rho_{p-1}^+)^3(1 - \rho_p^+)}\right.$$

$$\left. + \frac{[q_p + (r_p - q_p)C(m, m\rho_p^+)](1 - r_p\rho_{p-1}^+)}{(1 - r_p)^2(1 - \rho_{p-1}^+)^3}\right]. \quad (10)$$

The agreement of Eqs. (9) and (10) can be proved algebraically by using the relation

$$C(m, m\rho_{p-1}^+) = \frac{\rho_{p-1}^+ - r_p}{\rho_{p-1}^+(1 - r_p)} = \frac{B(m, m\rho_{p-1}^+)}{1 - \rho_{p-1}^+ + \rho_{p-1}^+B(m, m\rho_{p-1}^+)}.$$

We plot $E[W_p^2]$ in Fig. 4 for the numerical example described in Section 1.

4 Service Time

We are also interested in the total service time that each customer of class p receives before service completion in the M/M/m FCFS preemptive-resume priority queue. The total service time consists of several partial service times of two types, which we look at separately in the following.

Let $V_{p,k}^*(s)$ be the LST of the DF for the time to preemption for a customer of class p who competes for the servers with k other customers, where $0 \le k \le m - 1$. By referring to Fig. 3, we have the complete set of equations for $\{V_{p,k}^*(s); 0 \le k \le m-1\}$ as follows:

$$(s + \lambda_{p-1}^+ + \mu)V_{p,0}^*(s) = \lambda_{p-1}^+V_{p,1}^*(s),$$

$$[s + \lambda_{p-1}^+ + (k + 1)\mu]V_{p,k}^*(s) = \lambda_{p-1}^+V_{p,k+1}^*(s) + k\mu V_{p,k-1}^*(s) \quad 1 \le k \le m - 2,$$

$$(s + \lambda_{p-1}^+ + m\mu)V_{p,m-1}^*(s) = \lambda_{p-1}^+ + (m - 1)\mu V_{p,m-2}^*(s).$$

We note that $P_{p,k}\{\Pr\} = V^*_{p,k}(0)$ for $0 \le k \le m-1$. We obtain the mean $E[V_{p,k}] = -V^{*(1)}_{p,k}(0)$ as

$$E[V_{p,0}] = \sum_{j=1}^{m} \frac{(m\rho^+_{p-1})^j}{j!} \sum_{l=0}^{j-1} P_{p,l}\{\Pr\} \Bigg/ \lambda^+_{p-1} \sum_{j=0}^{m} \frac{(m\rho^+_{p-1})^j}{j!} ,$$

$$E[V_{p,m-1}] = \sum_{j=0}^{m-1} \frac{(m\rho^+_{p-1})^j}{j!} \sum_{l=j}^{m-1} P_{p,l}\{\Pr\} \Bigg/ m\mu \sum_{j=0}^{m} \frac{(m\rho^+_{p-1})^j}{j!} , \qquad (11)$$

and

$$E[V_{p,k}] = \frac{\left\{ \left[\sum_{j=k+1}^{m} \frac{(m\rho^+_{p-1})^j}{j!} \sum_{l=0}^{j-1} P_{p,l}\{\Pr\} \right] \sum_{j=0}^{k} \frac{(m\rho^+_{p-1})^j}{j!} - \left[\sum_{j=1}^{k} \frac{(m\rho^+_{p-1})^j}{j!} \sum_{l=0}^{j-1} P_{p,l}\{\Pr\} \right] \sum_{j=k+1}^{m} \frac{(m\rho^+_{p-1})^j}{j!} \right\}}{\lambda^+_{p-1} \frac{(m\rho^+_{p-1})^k}{k!} \sum_{j=0}^{m} \frac{(m\rho^+_{p-1})^j}{j!}} \qquad 1 \le k \le m-1, \qquad (12)$$

The second moment $E[V^2_{p,k}] = V^{*(2)}_{p,k}(0)$ is given by

$$E[V^2_{p,0}] = 2\sum_{j=1}^{m} \frac{(m\rho^+_{p-1})^j}{j!} \sum_{l=0}^{j-1} E[V_{p,l}] \Bigg/ \lambda^+_{p-1} \sum_{j=0}^{m} \frac{(m\rho^+_{p-1})^j}{j!} ,$$

$$E[V^2_{p,m-1}] = 2\sum_{j=0}^{m-1} \frac{(m\rho^+_{p-1})^j}{j!} \sum_{l=j}^{m-1} E[V_{p,l}] \Bigg/ m\mu \sum_{j=0}^{m} \frac{(m\rho^+_{p-1})^j}{j!} ,$$

and

$$\frac{E[V^2_{p,k}]}{2} = \frac{\left\{ \left[\sum_{j=k+1}^{m} \frac{(m\rho^+_{p-1})^j}{j!} \sum_{l=0}^{j-1} E[V_{p,l}] \right] \sum_{j=0}^{k} \frac{(m\rho^+_{p-1})^j}{j!} - \left[\sum_{j=1}^{k} \frac{(m\rho^+_{p-1})^j}{j!} \sum_{l=0}^{j-1} E[V_{p,l}] \right] \sum_{j=k+1}^{m} \frac{(m\rho^+_{p-1})^j}{j!} \right\}}{\lambda^+_{p-1} \frac{(m\rho^+_{p-1})^k}{k!} \sum_{j=0}^{m} \frac{(m\rho^+_{p-1})^j}{j!}} \qquad 1 \le k \le m-2.$$

Let $U^*_{p,k}(s)$ be the LST of the DF for the time to service completion for a customer of class p who competes for the servers with k other customers, where $0 \le k \le m-1$. By referring to Fig. 3 again, we have the complete set of equations for $\{U^*_{p,k}(s);\ 0 \le k \le m-1\}$ as follows:

$$(s+\lambda^+_{p-1}+\mu)U^*_{p,0}(s) = \mu + \lambda^+_{p-1}U^*_{p,1}(s),$$

$$[s+\lambda^+_{p-1}+(k+1)\mu]U^*_{p,k}(s) = \mu + \lambda^+_{p-1}U^*_{p,k+1}(s) + k\mu U^*_{p,k-1}(s) \qquad 1 \le k \le m-2,$$

$$(s+\lambda^+_{p-1}+m\mu)U^*_{p,m-1}(s) = \mu + (m-1)\mu U^*_{p,m-2}(s).$$

We note that $P_{p,k}\{\mathrm{Sr}\} = U^*_{p,k}(0)$ for $0 \le k \le m-1$. We obtain the mean $E[U_{p,k}] = -U^{*(1)}_{p,k}(0)$ as

$$E[U_{p,0}] = \sum_{j=1}^{m} \frac{(m\rho^+_{p-1})^j}{j!} \sum_{l=0}^{j-1} P_{p,l}\{\mathrm{Sr}\} \Bigg/ \lambda^+_{p-1} \sum_{j=0}^{m} \frac{(m\rho^+_{p-1})^j}{j!},$$

$$E[U_{p,m-1}] = \sum_{j=0}^{m-1} \frac{(m\rho^+_{p-1})^j}{j!} \sum_{l=j}^{m-1} P_{p,l}\{\mathrm{Sr}\} \Bigg/ m\mu \sum_{j=0}^{m} \frac{(m\rho^+_{p-1})^j}{j!},$$

and

$$E[U_{p,k}] = \frac{\left\{ \begin{array}{l} \left[\displaystyle\sum_{j=k+1}^{m} \frac{(m\rho^+_{p-1})^j}{j!} \sum_{l=0}^{j-1} P_{p,l}\{\mathrm{Sr}\} \right] \displaystyle\sum_{j=0}^{k} \frac{(m\rho^+_{p-1})^j}{j!} \\ - \left[\displaystyle\sum_{j=1}^{k} \frac{(m\rho^+_{p-1})^j}{j!} \sum_{l=0}^{j-1} P_{p,l}\{\mathrm{Sr}\} \right] \displaystyle\sum_{j=k+1}^{m} \frac{(m\rho^+_{p-1})^j}{j!} \end{array} \right\}}{\lambda^+_{p-1} \dfrac{(m\rho^+_{p-1})^k}{k!} \displaystyle\sum_{j=0}^{m} \frac{(m\rho^+_{p-1})^j}{j!}} \qquad 1 \le k \le m-2.$$

The second moment $E[U^2_{p,k}] = U^{*(2)}_{p,k}(0)$ is given by

$$E[U^2_{p,0}] = 2\sum_{j=1}^{m} \frac{(m\rho^+_{p-1})^j}{j!} \sum_{l=0}^{j-1} E[U_{p,l}] \Bigg/ \lambda^+_{p-1} \sum_{j=0}^{m} \frac{(m\rho^+_{p-1})^j}{j!},$$

$$E[U^2_{p,m-1}] = 2\sum_{j=0}^{m-1} \frac{(m\rho^+_{p-1})^j}{j!} \sum_{l=j}^{m-1} E[U_{p,l}] \Bigg/ m\mu \sum_{j=0}^{m} \frac{(m\rho^+_{p-1})^j}{j!},$$

and

$$\frac{E[U_{p,k}^2]}{2} = \frac{\left\{ \begin{array}{l} \left[\displaystyle\sum_{j=k+1}^{m} \frac{(m\rho_{p-1}^+)^j}{j!} \sum_{l=0}^{j-1} E[U_{p,l}] \right] \displaystyle\sum_{j=0}^{k} \frac{(m\rho_{p-1}^+)^j}{j!} \\ - \left[\displaystyle\sum_{j=1}^{k} \frac{(m\rho_{p-1}^+)^j}{j!} \sum_{l=0}^{j-1} E[U_{p,l}] \right] \displaystyle\sum_{j=k+1}^{m} \frac{(m\rho_{p-1}^+)^j}{j!} \end{array} \right\}}{\lambda_{p-1}^+ \dfrac{(m\rho_{p-1}^+)^k}{k!} \displaystyle\sum_{j=0}^{m} \frac{(m\rho_{p-1}^+)^j}{j!}} \qquad 1 \le k \le m-2.$$

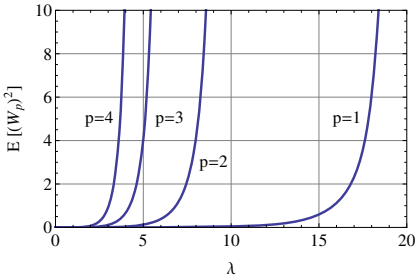


Fig. 4 Second moment of the waiting time for a customer of class p in the M/M/m FCFS preemptive-resume priority queue

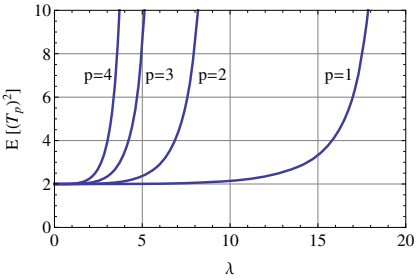


Fig. 5 Second moment of the response time for a customer of class p in the M/M/m FCFS preemptive-resume priority queue

Here we note the relations

$$E[U_{p,k}] + E[V_{p,k}] = \frac{P_{p,k}\{\mathrm{Sr}\}}{\mu} \quad ; \quad E[U_{p,k}^2] + E[V_{p,k}^2] = \frac{2E[U_{p,k}]}{\mu} \qquad 0 \le k \le m-1. \tag{13}$$

Let $S_{p,k}^*(s)$ be the LST of the DF for the total service time of a customer of class p who competes for the servers with k other customers, where $k \ge 0$. For a customer of class p who waits upon arrival or resumes service after preemption, the LST of the DF for the service time until completion is given by

$$\sum_{i=0}^{\infty} U_{p,m-1}^*(s)[V_{p,m-1}^*(s)]^i = \frac{U_{p,m-1}^*(s)}{1 - V_{p,m-1}^*(s)}.$$

Thus we have

$$S_{p,k}^*(s) = \begin{cases} U_{p,k}^*(s) + V_{p,k}^*(s) \dfrac{U_{p,m-1}^*(s)}{1 - V_{p,m-1}^*(s)} & 0 \le k \le m-1, \\ \dfrac{U_{p,m-1}^*(s)}{1 - V_{p,m-1}^*(s)} & k \ge m. \end{cases}$$

Then the LST of the DF for the total service time of a customer of class p is given by

$$S_p^*(s) = \sum_{k=0}^{\infty} Q_{p,k}^+ S_{p,k}^*(s) = \sum_{k=0}^{m-1} Q_{p,k}^+ S_{p,k}^*(s) + \sum_{k=m}^{\infty} Q_{p,k}^+ S_{p,k}^*(s)$$
$$= \sum_{k=0}^{m-1} Q_{p,k}^+ U_{p,k}^*(s) + \left\{ \sum_{k=0}^{m-1} Q_{p,k}^+ V_{p,k}^*(s) + C(m, m\rho_p^+) \right\} \frac{U_{p,m-1}^*(s)}{1 - V_{p,m-1}^*(s)}. \tag{14}$$

Using the relation in (13), we can show that

$$E[S_{p,k}] = \frac{1}{\mu} \quad ; \quad E[S_{p,k}^2] = \frac{2}{\mu^2} \qquad k \geq 0 \quad ; \quad E[S_p] = \frac{1}{\mu} \quad ; \quad E[S_p^2] = \frac{2}{\mu^2}.$$

These would be the mean and the second moment if we assumed that $S_{p,k}$ and S_p were exponentially distributed with mean $1/\mu$. However, we have not obtained $S_{p,k}^*(s)$ and $S_p^*(s)$ explicitly except for the case $p = 1$:

$$S_{1,k}^*(s) = S_1^*(s) = \frac{\mu}{s + \mu}.$$

5 Response Time

For customers of the highest priority class $p = 1$ who are never preempted, we get

$$T_1^*(s) = W_1^*(s) S_1^*(s), \tag{15}$$

where

$$W_1^*(s) = 1 - C(m, m\rho_1) + \frac{C(m, m\rho_1)(m\mu - \lambda_1)}{s + m\mu - \lambda_1}$$

is the well-known LST of the DF for the waiting time in the M/M/m FCFS queue (without priorities) with arrival rate λ_1.

For $p \geq 2$, by combining the arguments for deriving the LST of the DF for the waiting time W_p in Eq. (8) and the LST of the DF for the service time S_p in Eq. (14), the joint LST of the DF for W_p and S_p of a customer of class p is given by

$$\tilde{T}_p^*(s,s') = \sum_{k=0}^{m-1} Q_{p,k}^+ U_{p,k}^*(s') + \left\{ G_{p-1}^*(s) \sum_{k=0}^{m-1} Q_{p,k}^+ V_{p,k}^*(s') + C(m, m\rho_p^+) W_p^+(s) \right\} \frac{U_{p,m-1}^*(s')}{1 - G_{p-1}^*(s) V_{p,m-1}^*(s')}. \tag{16}$$

The marginal distributions yield

$$W_p^*(s) = \tilde{T}_p^*(s,0) \quad ; \quad S_p^*(s) = \tilde{T}_p^*(0,s),$$

which agree with Eqs. (8) and (14), respectively.

Finally, the LST of the DF for the response time T_p of a customer of class p is given by

$$T_p^*(s) = \tilde{T}_p^*(s,s) = \sum_{k=0}^{m-1} Q_{p,k}^+ U_{p,k}^*(s) + \left\{ G_{p-1}^*(s) \sum_{k=0}^{m-1} Q_{p,k}^+ V_{p,k}^*(s) + C(m, m\rho_p^+) W_p^+(s) \right\} \frac{U_{p,m-1}^*(s)}{1 - G_{p-1}^*(s) V_{p,m-1}^*(s)}. \tag{17}$$

Clearly it does *not* hold that $T_p^*(s) = W_p^*(s) S_p^*(s)$ for $p \geq 2$, which means that the total waiting time W_p and the total service time S_p are not independent for $p \geq 2$. In fact, they are positively correlated as we get from Eq. (16) the covariance of the total time and the service time:

$$\mathrm{Cov}[W_p, S_p] = \frac{\sum_{k=0}^{m-1} Q_{p,k}^+ E[V_{p,k}]}{m\mu(1-r_p)(1-\rho_{p-1}^+)} + \frac{\{[1 - C(m, m\rho_p^+)]q_p + C(m, m\rho_p^+)\} E[V_{p,m-1}]}{m\mu(1-r_p)^2(1-\rho_{p-1}^+)}, \tag{18}$$

where $E[V_{p,k}]$ and $E[V_{p,m-1}]$ are given in Eqs. (12) and (11), respectively. For customers of the highest priority class ($p = 1$), we have $T_1^*(s)$ given in Eq. (15) and $E[W_1, S_1] = 0$, which means that W_1 and S_1 are independent, because they are never preempted. We plot $\mathrm{Cov}[W_p, S_p]$ in Fig. 6 for $p \geq 2$ for the numerical example described in Section 1.

From Eq. (17), we get the mean response time $E[T_p]$ already given in Eq. (3). The second moment of the response time is given by

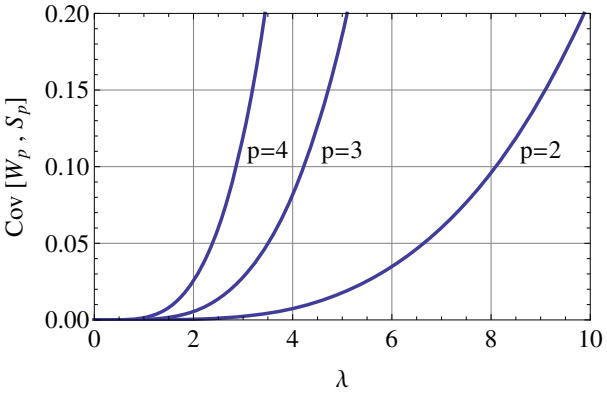


Fig. 6 Covariance of the total waiting time and the total service time for a customer of class p in the M/M/m FCFS preemptive-resume priority queue

$$E[T_p^2] = E[(W_p + S_p)^2] = E[W_p^2] + 2\mathrm{Cov}[W_p, S_p] + 2E[W_p]E[S_p] + E[S_p^2]$$

$$= \frac{2\sum_{k=0}^{m-1} Q_{p,k}^+ E[V_{p,k}]}{m\mu(1-r_p)(1-\rho_{p-1}^+)} + \frac{2\{[1-C(m,m\rho_p^+)]q_p + C(m,m\rho_p^+)\}E[V_{p,m-1}]}{m\mu(1-r_p)^2(1-\rho_{p-1}^+)}$$

$$+ \frac{2[1-C(m,m\rho_p^+)]q_p}{(m\mu)^2}\left[\frac{1-r_p\rho_{p-1}^+}{(1-r_p)^2(1-\rho_{p-1}^+)^3} + \frac{m}{(1-r_p)(1-\rho_{p-1}^+)}\right]$$

$$+ \frac{2C(m,m\rho_p^+)}{(m\mu)^2}\left[\frac{1-\rho_{p-1}^+\rho_p^+}{(1-\rho_{p-1}^+)^3(1-\rho_p^+)^2} + \frac{r_p[2-\rho_{p-1}^+-\rho_p^+-r_p(1-\rho_{p-1}^+\rho_p)]}{(1-r_p)^2(1-\rho_{p-1}^+)^3(1-\rho_p^+)}\right.$$

$$\left. + \frac{m(1-r_p\rho_p^+)}{(1-r_p)(1-\rho_{p-1}^+)(1-\rho_p^+)}\right] + \frac{2}{\mu^2}. \tag{19}$$

We plot $E[T_p^2]$ in Fig. 5 for the numerical example described in Section 1. We have confirmed that this yields the same result as obtained by the method of Segall [4].

It remains us to find the explict LST of the DF, $T_p^*(s)$, for a customer of generic class p in the M/M/m FCFS preemptive-resume priority queue.

Acknowledgment This work is supported by the Grant-in-Aid for Scientific Research (C) No. 26330354 from the Japan Society for Promotion of Science in 2015.

References

1. Brosh, I.: Preemptive priority assignment in multichannel systems. Operations Research **17**(3), 526–535 (1969)
2. Buzen, J.P., Bondi, A.B.: The response times of priority classes under preemptive resume in M/M/m queues. Operations Research **31**(3), 456–465 (1983)

3. Fujiki, M.: Fundamental theory and application on communication traffic. 5 queuein theory (part 2). Transactions of the Institute of Electronics and Communication Engineers of Japan **55**(9), 1194–1200 (1972). (in Japanese)
4. Segal, M.: A multiserver system with preemptive priorities. Operations Research **18**(2), 316–323 (1970)
5. Tatashev, A.G.: Calculation of the distribution of the waiting time in a multiple-channel queueing system with fixed priorities. Engineering Cybernetics **22**(1), 59–62 (1984). (Originally published in Tekhnicheskaya Kibernetika, No.6, pp. 163–166 (1983)
6. Zeltyn, S., Feldman, Z., Wasserkrug, S.: Waiting and sojourn times in a multi-server queue with mixed priorities. Queueing Systems **61**(4), 305–328 (2009)

Exhaustive Vacation Queue with Dependent Arrival and Service Processes

Gábor Horváth, Zsolt Saffer and Miklós Telek

Abstract This paper presents a more general class of MAP/MAP/1 exhaustive vacation queue, in which the Markov modulated arrival and service processes are dependent. This model class requires the evaluation of the busy period of quasi birth death process with arbitrary initial level, which is a new analysis element.

The model is analyzed by applying matrix analytic methods for the underlying quasi birth death process. The main result of the paper is the probability-generating function of the number of jobs in the system. Finally, a numerical example provides an insight into the behavior of the model.

Keywords Vacation queue · MAP · Dependent arrival and service process · QBD · Matrix analytic methods · stationary analysis

1 Introduction

The importance of vacation queues comes for their diverse application fields: modeling various computer systems, telecommunication protocols, manufacturing, logistics, etc. For details on analysis works on vacation models and their generalizations the reader is referred to the recent surveys [5] and [10].

Due to the versatility of the Markovian Arrival Process (MAP) [7], vacation queues with MAP input and general service times have also been investigated in several past papers [4, 8, 9]. Only a few discrete-time models have been investigated, in which

G. Horváth · Z. Saffer · M. Telek
Budapest University of Technology and Economics, Budapest, Hungary
e-mail: saffer@webspn.hit.bme.hu

G. Horváth(✉) · M. Telek
MTA-BME Information Systems Research Group, Budapest, Hungary
e-mail: hgabor@webspn.hit.bme.hu, telek@webspn.hit.bme.hu

The authors thank the support of the OTKA K101150 project.

T.V. Do et al. (eds.), *Queueing Theory and Network Applications*,
Advances in Intelligent Systems and Computing 383,
DOI: 10.1007/978-3-319-22267-7_2

both the arrival and the service processes are Markovian. MAP/PH/1 vacation models have been analyzed by A.-S. Alpha in [1, 2] and by C. Goswami and N. Selvaraju in [3].

In this paper we consider a more general class of exhaustive vacation queues with dependent Markov modulated arrival and service processes. This model class requires the introduction of a new analysis element, the evaluation of the busy period of quasi birth death (QBD) processes with arbitrary initial level. We provide the expression of the probability-generating function (PGF) of the number of jobs in the system. In the last part of the paper we provide a numerical example and investigate the effects of different vacation distributions on the mean number of jobs.

2 Model Description

We consider the dependent MAP/MAP/1 exhaustive vacation queue. The model falls in the class of single server FCFS queue with multiple vacations and exhaustive discipline [10]. According to the rule of exhaustive service discipline the server serves the jobs in the queue until it gets idle, then the server leaves for vacation for an independent and identically distributed random amount of time. If the queue is idle at the end of the vacation the server leaves for a new vacation, otherwise it starts serving the jobs in the queue. The random vacation time, its probability density function (pdf) and its Laplace transform (LT) are denoted by $\tilde{\sigma}, \sigma(t)$ and $\sigma^*(s) = E(e^{-s\tilde{\sigma}})$, respectively.

The arrivals and services are characterized by seven matrices: $\mathbf{L_v}, \mathbf{F_v}, \mathbf{B_s}$, $\mathbf{L_s}, \mathbf{F_s}, \boldsymbol{\Pi}_{vs}$ and $\boldsymbol{\Pi}_{sv}$.

- During the vacations the arrivals are given by a MAP, where the entries of $\mathbf{L_v}$ are the rates of transitions without a job arrival, and the entries of $\mathbf{F_v}$ are the rates of transitions that are accompanied by a job arrival. Matrix $\mathbf{L_v} + \mathbf{F_v}$ is therefore the generator of the continuous time Markov chain (CTMC) with N_v states which modulates the arrivals during vacation.
- When the server serves the jobs, the queue behaves as a quasi birth-death (QBD) [6] process, there the matrices $\mathbf{B_s}$, $\mathbf{L_s}$ and $\mathbf{F_s}$ contains the transition rates associated with a service completion, without service completion and job arrival, with a job arrival, respectively. In this case the generator of the modulating CTMC is $\mathbf{B_s} + \mathbf{L_s} + \mathbf{F_s}$, and it has N_s states.
- The transition between the vacation and service periods is given by $N_v \times N_s$ stochastic matrix $\boldsymbol{\Pi}_{vs}$, whose entries are the probabilities of the state transitions occurring at the end of the vacation period. The $N_s \times N_v$ matrix $\boldsymbol{\Pi}_{sv}$ has a similar role, holding the probabilities of phase transitions when the service period ends and a vacation starts.

This is a general model which covers a number of special cases, e.g., the MAP/PH/1 vacation queue and the MAP/MAP/1 vacation queue.

The stability of the model is determined by the stationary drift of the QBD during service [6]. Hence the necessary and sufficient condition of the stability of this vacation model is

$$\alpha_s \mathbf{F_s} \mathbb{1} - \alpha_s \mathbf{B_s} \mathbb{1} < 0, \tag{1}$$

where α_s is the solution of the linear system $\alpha_s \, (\mathbf{B_s} + \mathbf{L_s} + \mathbf{F_s}) = \mathbf{0}$, $\alpha_s \mathbb{1} = 1$, and $\mathbb{1}$ denotes the column vector of ones.

3 The Number of Jobs in the System

To characterize the number of jobs in the system, let us introduce the two dimensional process $\mathcal{X}(t) = \{\mathcal{N}(t), \mathcal{J}(t), t \geq 0\}$, where $\mathcal{N}(t)$ denotes the number of jobs (also referred to as *levels*) and $\mathcal{J}(t)$ denotes the state of the modulating CTMC (also referred to as *phase*) at time t. For the analysis of $\mathcal{X}(t)$ the evolution of the queue is divided to *cycles*, as shown in Figure 1. Each cycle starts with a vacation period, which is followed by a service period, and the cycle ends when the last job leaves the system. Note that a cycle can also be degenerate: if no jobs arrive during the vacation period, there is no service period (see cycle $i-1$ in Figure 1).

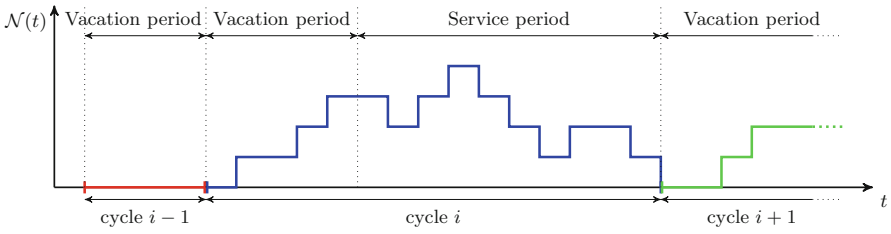


Fig. 1 Cycles in the evolution of the queue

The stationary probability that there are ℓ ($\ell \geq 1$) jobs in the system is proportional to M_ℓ, the mean time spent at level ℓ in a stationary cycle.

$$M_\ell = \underbrace{\int_{u=0}^{\infty} \sigma(u) \int_{t=0}^{u} \beta \, \mathbf{P}_\ell^{(\mathbf{v})}(t) \mathbb{1} \, dt \, du}_{M_\ell^{(v)}} + \underbrace{\int_{u=0}^{\infty} \sigma(u) \sum_{m=1}^{\infty} \beta \, \mathbf{P}_m^{(\mathbf{v})}(u) \mathbf{\Pi_{vs}} \mathbf{H}_{m,\ell} \mathbb{1} \, du}_{M_\ell^{(s)}}, \tag{2}$$

where row vector β of size N_v is the stationary phase distribution at the beginning of the vacation period, matrix $\mathbf{P}_\ell^{(\mathbf{v})}(t)$ characterizes the number of arrivals up to time t during the vacation period, defined as

$$[\mathbf{P}_\ell^{(\mathbf{v})}(t)]_{i,j} = P(\mathcal{N}(t) = \ell, \mathcal{J}(t) = j, \tilde{\sigma} > t | \mathcal{N}(0) = 0, \mathcal{J}(0) = i), \tag{3}$$

and $[\mathbf{H}_{m,\ell}]_{i,j}$ is the mean times spent in level ℓ and phase j in the service period starting from level m and phase i. The first and second term of (2), M(v) and M(s), correspond to the vacation and service period, respectively. Closed form formulas are provided for both in the next subsections. From M_ℓ the stationary distribution of $\mathcal{N}(t)$ is obtained by normalization, $q_\ell = \lim_{t\to\infty} P(\mathcal{N}(t) = \ell) = M_\ell / \sum_k M_k$.

The Evolution of the Number of Jobs During the Vacation Period. The evolution of the number of jobs during the vacation period resembles to the counting process of a MAP given by matrices $\mathbf{L_v}, \mathbf{F_v}$. Thus, for matrices $\mathbf{P}_\ell^{(\mathbf{v})}(t)$ we have

$$\frac{d}{dt}\mathbf{P}_\ell^{(\mathbf{v})}(t) = \mathbf{P}_{\ell-1}^{(\mathbf{v})}(t)\mathbf{F_v} + \mathbf{P}_\ell^{(\mathbf{v})}(t)\mathbf{L_v}, \quad \text{for } \ell > 0, \tag{4}$$

$$\frac{d}{dt}\mathbf{P}_0^{(\mathbf{v})}(t) = \mathbf{P}_0^{(\mathbf{v})}(t)\mathbf{L_v}, \tag{5}$$

with initial condition $\mathbf{P}_\ell^{(\mathbf{v})}(0) = \delta_{0,\ell}\mathbf{I}$, where δ denotes the Kronecker delta (that is $\delta_{ii} = 1$, $\delta_{ij} = 0$ for $i \neq j$). Similar to [6, Sec. 3], multiplying the ℓth equation by z^ℓ, summing up and solving the differential equation gives the generating function

$$\mathbf{P}^{(\mathbf{v})*}(z,t) = \sum_{\ell=0}^{\infty} z^\ell \mathbf{P}_\ell^{(\mathbf{v})}(t) = e^{(\mathbf{L_v}+z\mathbf{F_v})t}. \tag{6}$$

The Mean Time Spent in Different Levels During the Service Period. As a new contributions of the paper we derive matrix $\mathbf{H}_{m,\ell}$, which is the mean time spent in various phases of level ℓ starting from level m in a QBD characterized by matrices $\mathbf{B_s}, \mathbf{L_s}$ and $\mathbf{F_s}$. It is known that the mean time spent at different phases of level ℓ starting from level 0 before returning to level 0 is given by $\mathbf{R}^\ell$ [6].

But, in our vacation queue the starting level after a vacation is not 0, but the number of arrivals during the vacation, which is denoted by m. According to our best knowledge, this measure has not been investigated yet.

For $m > 0$, we define matrix $\mathbf{P}_{m,\ell}^{(\mathbf{s})}$ corresponding to the service period as

$$[\mathbf{P}_{m,\ell}^{(\mathbf{s})}(t)]_{i,j} = P(\Theta > u+t, \mathcal{N}(u+t) = \ell, \mathcal{J}(u+t) = j | \mathcal{N}(u) = m, \mathcal{J}(u) = i, \tilde{\sigma} = u),$$

where u marks the beginning and Θ marks the end of the service period, thus $\Theta = \min\{t : \mathcal{N}(u+t) = 0\}$. For $\ell > 1$ matrix $\mathbf{P}_{m,\ell}^{(\mathbf{s})}(t)$ satisfies

$$\frac{d}{dt}\mathbf{P}_{m,\ell}^{(\mathbf{s})}(t) = \mathbf{P}_{m,\ell-1}^{(\mathbf{s})}(t)\mathbf{F_s} + \mathbf{P}_{m,\ell}^{(\mathbf{s})}(t)\mathbf{L_s} + \mathbf{P}_{m,\ell+1}^{(\mathbf{s})}(t)\mathbf{B_s}, \tag{7}$$

and for $\ell = 1$ we have

$$\frac{d}{dt}\mathbf{P}^{(\mathbf{s})}_{m,1}(t) = \mathbf{P}^{(\mathbf{s})}_{m,1}(t)\mathbf{L}_\mathbf{s} + \mathbf{P}^{(\mathbf{s})}_{m,2}(t)\mathbf{B}_\mathbf{s}, \tag{8}$$

with initial values $\mathbf{P}^{(\mathbf{s})}_{m,\ell}(0) = \delta_{m,\ell}\mathbf{I}$. We are interested in the mean time spent in different states during the busy period, that is $\mathbf{H}_{m,\ell} = \int_{t=0}^{\infty} \mathbf{P}^{(\mathbf{s})}_{m,\ell}(t)\,dt$. Integrating the differential equation (7) and (8) from $t = 0$ to ∞ we get

$$\mathbf{P}^{(\mathbf{s})}_{m,\ell}(\infty) - \mathbf{P}^{(\mathbf{s})}_{m,\ell}(0) = \mathbf{H}_{m,\ell-1}\mathbf{F}_\mathbf{s} + \mathbf{H}_{m,\ell}\mathbf{L}_\mathbf{s} + \mathbf{H}_{m,\ell+1}\mathbf{B}_\mathbf{s}, \quad \text{for } \ell > 1, \tag{9}$$

$$\mathbf{P}^{(\mathbf{s})}_{m,1}(\infty) - \mathbf{P}^{(\mathbf{s})}_{m,1}(0) = \mathbf{H}_{m,1}\mathbf{L}_\mathbf{s} + \mathbf{H}_{m,2}\mathbf{B}_\mathbf{s}. \tag{10}$$

These two equations and the initial value $\mathbf{P}^{(\mathbf{s})}_{m,\ell}(0)$ lead to four different cases, in general: (1) when $\ell = 1$, (2) when $1 < \ell < m$, (3) when $1 < \ell = m$, finally, (4) when $\ell > m$. The corresponding equations are

$$-\delta_{m,1}\mathbf{I} = \mathbf{H}_{m,1}\mathbf{L}_\mathbf{s} + \mathbf{H}_{m,2}\mathbf{B}_\mathbf{s}, \tag{11}$$

$$\mathbf{0} = \mathbf{H}_{m,\ell-1}\mathbf{F}_\mathbf{s} + \mathbf{H}_{m,\ell}\mathbf{L}_\mathbf{s} + \mathbf{H}_{m,\ell+1}\mathbf{B}_\mathbf{s}, \quad \text{for } 1 < \ell < m, \tag{12}$$

$$-\mathbf{I} = \mathbf{H}_{m,m-1}\mathbf{F}_\mathbf{s} + \mathbf{H}_{m,m}\mathbf{L}_\mathbf{s} + \mathbf{H}_{m,m+1}\mathbf{B}_\mathbf{s}, \quad \text{for } m > 1, \tag{13}$$

$$\mathbf{0} = \mathbf{H}_{m,\ell-1}\mathbf{F}_\mathbf{s} + \mathbf{H}_{m,\ell}\mathbf{L}_\mathbf{s} + \mathbf{H}_{m,\ell+1}\mathbf{B}_\mathbf{s}, \quad \text{for } \ell > m, \tag{14}$$

where δ denotes the Kronecker delta again. The solution of (11)-(14) is given by a matrix-geometric combination

$$\mathbf{H}_{m,\ell} = \mathbf{\Phi}\mathbf{R}^{\ell-1} + \mathbf{\Psi}\mathbf{S}^{m-\ell}, \quad \text{for } 1 \le \ell \le m, \tag{15}$$

$$\mathbf{H}_{m,\ell} = \mathbf{H}_{m,m}\mathbf{R}^{\ell-m}, \quad \text{for } 1 \le m < \ell, \tag{16}$$

where matrices $\mathbf{R}$ and $\mathbf{S}$ are obtained such that the regular equations (12) and (14) are satisfied for any $\mathbf{\Psi}$ and $\mathbf{\Phi}$. $\mathbf{R}$ and $\mathbf{S}$ are the minimal non-negative solutions to the quadratic equations [6, Sec. 10]

$$\mathbf{0} = \mathbf{F}_\mathbf{s} + \mathbf{R}\mathbf{L}_\mathbf{s} + \mathbf{R}^2\mathbf{B}_\mathbf{s}, \quad \mathbf{0} = \mathbf{B}_\mathbf{s} + \mathbf{S}\mathbf{L}_\mathbf{s} + \mathbf{S}^2\mathbf{F}_\mathbf{s}, \tag{17}$$

Matrices $\mathbf{\Psi}$ and $\mathbf{\Phi}$ are obtained from the solution of the irregular equations (11) and (13) as

$$\mathbf{0} = \mathbf{\Phi}(\mathbf{L}_\mathbf{s} + \mathbf{R}\mathbf{B}_\mathbf{s}) + \mathbf{\Psi}(\mathbf{S}^{m-1}\mathbf{L}_\mathbf{s} + \mathbf{S}^{m-2}\mathbf{B}_\mathbf{s}), \tag{18}$$

$$-\mathbf{I} = \mathbf{\Phi}\mathbf{R}^{m-2}\underbrace{(\mathbf{F}_\mathbf{s} + \mathbf{R}\mathbf{L}_\mathbf{s} + \mathbf{R}^2\mathbf{B}_\mathbf{s})}_{\mathbf{0}} + \mathbf{\Psi}(\mathbf{S}\mathbf{F}_\mathbf{s} + \mathbf{L}_\mathbf{s} + \mathbf{R}\mathbf{B}_\mathbf{s}). \tag{19}$$

The solution of $\mathbf{\Psi}$ and $\mathbf{\Phi}$ are

$$\mathbf{\Psi} = (-\mathbf{S}\mathbf{F}_\mathbf{s} - \mathbf{L}_\mathbf{s} - \mathbf{R}\mathbf{B}_\mathbf{s})^{-1}, \tag{20}$$

$$\begin{aligned}\mathbf{\Phi} &= \mathbf{\Psi}(\mathbf{S}^{m-1}\mathbf{L}_\mathbf{s} + \mathbf{S}^{m-2}\mathbf{B}_\mathbf{s})(\mathbf{L}_\mathbf{s} + \mathbf{R}\mathbf{B}_\mathbf{s})^{-1} \\ &= -\mathbf{\Psi}\mathbf{S}^m\mathbf{F}_\mathbf{s}(\mathbf{L}_\mathbf{s} + \mathbf{R}\mathbf{B}_\mathbf{s})^{-1} = -\mathbf{\Psi}\mathbf{S}^m\mathbf{R},\end{aligned} \tag{21}$$

where we exploited various identities of the fundamental matrices of QBDs. Finally using the expressions of $\mathbf{H}_{m,\ell}$ from (16) and (15) as well as (20) and (21) we get

$$\mathbf{H}_{m,\ell} = \underbrace{-\mathbf{\Psi}\mathbf{S}^m\mathbf{R}^\ell}_{term1} + \underbrace{\mathbf{\Psi}\mathbf{R}^{\ell-m}}_{term2}, \qquad \text{for } 1 \le m \le \ell, \tag{22}$$

$$\mathbf{H}_{m,\ell} = \underbrace{-\mathbf{\Psi}\mathbf{S}^m\mathbf{R}^\ell}_{term1} + \underbrace{\mathbf{\Psi}\mathbf{S}^{m-\ell}}_{term3}, \qquad \text{for } 1 \le \ell < m. \tag{23}$$

The Mean Time Spent at Each Level in a Stationary Cycle. By applying (6) in the first term of (2), its generating function, $M_\ell^{(v)}$, can be expressed as

$$\begin{aligned}M^{(v)*}(z) &= \sum_{\ell=0}^{\infty} z^\ell M_\ell^{(v)} = \beta \int_{u=0}^{\infty} \sigma(u) \int_{t=0}^{u} e^{(\mathbf{L}_\mathbf{v} + z\mathbf{F}_\mathbf{v})t} \mathbb{1}\, dt\, du \\ &= \beta \int_{u=0}^{\infty} \sigma(u) \left(\mathbf{I} - e^{(\mathbf{L}_\mathbf{v} + z\mathbf{F}_\mathbf{v})u}\right) (-\mathbf{L}_\mathbf{v} - z\mathbf{F}_\mathbf{v})^{-1} \mathbb{1} du \\ &= \beta \left(\mathbf{I} - \sigma^*(\mathbf{L}_\mathbf{v} + z\mathbf{F}_\mathbf{v})\right) (-\mathbf{L}_\mathbf{v} - z\mathbf{F}_\mathbf{v})^{-1} \mathbb{1},\end{aligned} \tag{24}$$

where $\sigma^*(\mathbf{M})$ with square matrix $\mathbf{M}$ is defined by $\int_{u=0}^{\infty} \sigma(u)\, e^{\mathbf{M}u} du$.

Lemma 1. *For any row vector x of size N_v, matrix $\mathbf{X}$ of size $N_v \times N_s$ and matrix $\mathbf{Y}$ of size $N_s \times N_s$, if the infinite sum exists we have*

$$\sum_{m=0}^{\infty} x\, \mathbf{P}_m^{(\mathbf{v})}(t)\, \mathbf{X}\, \mathbf{Y}^m = vec^T\langle \mathbf{X}^T \rangle e^{(\mathbf{L}_\mathbf{v}{}^T \otimes \mathbf{I} + \mathbf{F}_\mathbf{v}{}^T \otimes \mathbf{Y})t} (x^T \otimes \mathbf{I}), \tag{25}$$

where $vec\langle\rangle$ is the column stacking operator, which generates a column vector from the columns of a matrix.

Proof. The proof of the lemma is omitted due to space limitations. □

$M_\ell^{(s)}$ is obtained by substituting the expressions (22) and (23) of the determined matrices $\mathbf{H}_{m,\ell}$ into the definition (2). For the generating function $M^{(s)*}(z) = \sum_{\ell=0}^{\infty} z^\ell M_\ell^{(s)}$ we get

$$M^{(s)*}(z) = \sum_{\ell=0}^{\infty} z^{\ell} M_{\ell}^{(s)} = \underbrace{-\sum_{\ell=0}^{\infty} z^{\ell} \int_{u=0}^{\infty} \sigma(u) \sum_{m=1}^{\infty} \beta \mathbf{P}_m^{(\mathbf{v})}(u) \mathbf{\Pi}_{\boldsymbol{vs}} \mathbf{\Psi} \mathbf{S}^m \mathbf{R}^{\ell} \mathbb{1}\, du}_{M_1^{(s)*}(z)}$$
$$+ \underbrace{\sum_{\ell=0}^{\infty} z^{\ell} \int_{u=0}^{\infty} \sigma(u) \sum_{m=1}^{\ell} \beta \mathbf{P}_m^{(\mathbf{v})}(u) \mathbf{\Pi}_{\boldsymbol{vs}} \mathbf{\Psi} \mathbf{R}^{\ell-m} \mathbb{1}\, du}_{M_2^{(s)*}(z)}$$
$$+ \underbrace{\sum_{\ell=0}^{\infty} z^{\ell} \int_{u=0}^{\infty} \sigma(u) \sum_{m=\ell+1}^{\infty} \beta \mathbf{P}_m^{(\mathbf{v})}(u) \mathbf{\Pi}_{\boldsymbol{vs}} \mathbf{\Psi} \mathbf{S}^{m-\ell} \mathbb{1}\, du}_{M_3^{(s)*}(z)}.$$

By applying rearrangements and making use of Lemma 1 the above three terms can be expressed in closed-form resulting a formula for $M^{(s)*}(z)$ as

$$M^{(s)*}(z) = \Big(\beta \sigma^*(\mathbf{L_v} + z\mathbf{F_v})\mathbf{\Pi}_{\boldsymbol{vs}}\mathbf{\Psi} - \mathrm{vec}^T\langle \mathbf{\Psi}^T \mathbf{\Pi}_{\boldsymbol{vs}}{}^T \rangle \sigma^*(\mathbf{L_v}^T \otimes \mathbf{I} + \mathbf{F_v}^T \otimes \mathbf{S})(\beta^T \otimes \mathbf{I})\Big)$$
$$\Big((\mathbf{I} - z\mathbf{R})^{-1} + (z\mathbf{I} - \mathbf{S})^{-1}\mathbf{S}\Big)\mathbb{1}.$$

The Generating Function of the Number of Jobs in the System. The phase $(\mathcal{J}(t))$ at the beginning of the cycles form an embedded discrete time Markov chain (DTMC). The probability matrices characterizing the number of arriving jobs and the phase transitions during the vacation period are $\int_0^{\infty} \sigma(x)\mathbf{P}_m^{(\mathbf{v})}(x)\,dx$. If m jobs are in the queue when the system enters the service period then the phase transitions are given by $\mathbf{G}^m$, where matrix $\mathbf{G}$ is the minimal non-negative solution to the matrix-quadratic equation $\mathbf{0} = \mathbf{B_s} + \mathbf{L_s}\mathbf{G} + \mathbf{F_s}\mathbf{G}^2$. Thus, the transition probability matrix of the DTMC, denoted by $\mathbf{Q}$, is expressed by

$$\mathbf{Q} = \int_0^{\infty} \sigma(x) \sum_{m=0}^{\infty} \mathbf{P}_m^{(\mathbf{v})}(x) \mathbf{\Pi}_{\boldsymbol{vs}} \mathbf{G}^m \mathbf{\Pi}_{\boldsymbol{sv}}\, dx. \tag{26}$$

The stationary distribution of $\mathbf{Q}$, denoted by β, is determined by the linear system $\beta\mathbf{Q} = \beta$, $\beta\mathbb{1} = 1$. Making use of Lemma 1, vector β is the solution to

$$\mathrm{vec}^T\langle \mathbf{\Pi}_{\boldsymbol{vs}}{}^T \rangle \sigma^*(\mathbf{L_v}^T \otimes \mathbf{I} + \mathbf{F_v}^T \otimes \mathbf{G})(\beta^T \otimes \mathbf{\Pi}_{\boldsymbol{sv}}) = \beta, \quad \beta\mathbb{1} = 1. \tag{27}$$

Theorem 1. *The generating function of the stationary number of jobs in the system, $q(z)$, is given by*

$$q(z) = \frac{1}{c}\Bigg(\beta\Big(\mathbf{I} - \sigma^*(\mathbf{L_v} + z\mathbf{F_v})\Big)(-\mathbf{L_v} - z\mathbf{F_v})^{-1}\mathbb{1} + \Big(\beta\sigma^*(\mathbf{L_v}+z\mathbf{F_v})\mathbf{\Pi}_{vs}\mathbf{\Psi} - vec^T\langle\mathbf{\Psi}^T\mathbf{\Pi}_{vs}^T\rangle\sigma^*(\mathbf{L_v}^T\otimes\mathbf{I}+\mathbf{F_v}^T\otimes\mathbf{S})(\beta^T\otimes\mathbf{I})\Big) \cdot \Big((\mathbf{I}-z\mathbf{R})^{-1} + (z\mathbf{I}-\mathbf{S})^{-1}\mathbf{S}\Big)\mathbb{1}\Bigg),$$

where β *is determined by* (27) *and the constant* c *satisfies* $\lim_{z\to 1} q(z) = 1$.

Taking the derivatives of $q(z)$ at $z \to 1$ provides the factorial moments of the number of jobs in the queue.

4 Numerical Example

This numerical example investigates the effect of the mean and the distribution of the vacation time on the mean number of jobs in the system[1]. Since during the service period the arrival and the service processes are dependent we characterize the overall effect of the Markov environment by the following matrices :

$$\mathbf{B_s} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 4 & 1 & 2 \end{bmatrix}, \mathbf{L_s} = \begin{bmatrix} -8 & 1 & 0 \\ 0 & -5 & 2 \\ 1 & 3 & -11 \end{bmatrix}, \mathbf{F_s} = \begin{bmatrix} 2 & 1 & 4 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$
$$\mathbf{F_v} = \begin{bmatrix} 3 & 1 \\ 0 & 1 \end{bmatrix}, \mathbf{L_v} = \begin{bmatrix} -5 & 1 \\ 2 & -3 \end{bmatrix}, \mathbf{\Pi}_{sv} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0.1 & 0.9 \end{bmatrix}, \mathbf{\Pi}_{vs} = \begin{bmatrix} 0.8 & 0 & 0.2 \\ 0 & 0.7 & 0.3 \end{bmatrix}. \tag{28}$$

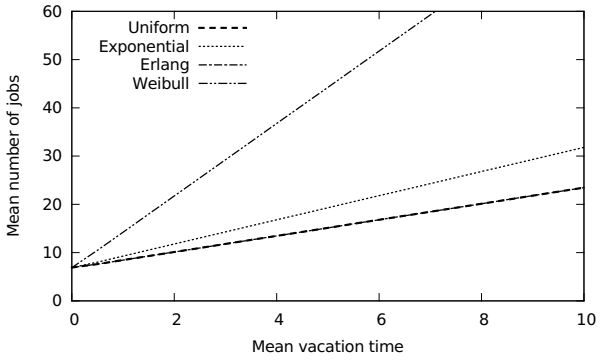


Fig. 2 The mean number of jobs in the system

[1] The Mathematica implementation can be downloaded from http://hit.bme.hu/~ghorvath/software

The computation has been performed for the following type of vacation distributions: Uniform distribution, Exponential distribution, Erlang distribution with shape parameter of 3, Weibull distribution with shape parameter of $k = 1/2$. The mean number of jobs computed from Theorem 1 is depicted in Figure 2. As expected, the number of jobs in the system is the highest when the vacation times are Weibull distributed which has the heaviest tail. The plots corresponding to the uniform and the Erlang cases match completely. These distributions have the same squared coefficient of variations (that is $1/3$), thus the results suggest same kind of insensitivity as in the M/G/1 queue.

References

1. Alfa, A.S.: A discrete MAP/PH/1 queue with vacations and exhaustive time-limited service. Oper. Res. Lett. **18**, 31–40 (1995)
2. Alfa, A.S.: Discrete time analysis of MAP/PH/1 vacation queue with gated time-limited service. Queueing Systems **29**(1), 35–54 (1998)
3. Goswami, C., Selvaraju, N.: The discrete-time MAP/PH/1 queue with multiple working vacations. Applied Mathematical Modelling **34**, 931–946 (2010)
4. Chang, S., Takine, T.: Factorization and stochastic decomposition properties in bulk queues with generalized vacations. Queueing Systems **50**(2–3), 165–183 (2005)
5. Ke, J.C., Wu, C.H., Zhang, Z.G.: Recent developments in vacation queueing models: a short survey. International Journal of Operations Research **7**(4), 3–8 (2010)
6. Latouche, G., Ramaswami, V.: Introduction to matrix analytic methods in stochastic modeling, vol. 5. SIAM (1999)
7. Neuts, M.F.: A versatile markovian point process. Journal of Applied Probability, 764–779 (1979)
8. Saffer, Z., Telek, M.: Analysis of *BMAP*/*G*/1 vacation model of non-*M*/*G*/1-type. In: Thomas, N., Juiz, C. (eds.) EPEW 2008. LNCS, vol. 5261, pp. 212–226. Springer, Heidelberg (2008)
9. Saffer, Z., Telek, M.: Closed form results for BMAP/G/1 vacation model with binomial type disciplines. Publ. Math. Debrecen **76**(3), 359–378 (2010)
10. Tian, N., Zhang, Z.G.: Vacation queueing models: theory and applications, vol. 93. Springer Science & Business Media (2006)

Delay Analysis of a Queue with General Service Demands and Phase-Type Service Capacities

Michiel De Muynck, Herwig Bruneel and Sabine Wittevrongel

Abstract We present the analysis of a non-classical discrete-time queueing model where customers demand variable amounts of work from a server that is able to perform this work at a varying rate. The service demands of the customers are integer numbers of *work units*. They are assumed to be independent and identically distributed (i.i.d.). The service capacities, i.e., the numbers of work units that the server can process in the consecutive slots, are also assumed to be i.i.d. and have a rational probability generating function (pgf). Finally, the numbers of customer arrivals in each slot are i.i.d. as well. We analyze this model analytically using contour integration. Our main result is an expression for the pgf of the customer delay in steady state, from which expressions for the moments of the delay can be derived.

Keywords Discrete-time queueing theory · Service demands · Service capacities · Complex contour integration

1 Introduction

In many naturally occurring single-server queueing phenomena, the customers require different amounts of service from the server, and the rate at which the server is able to provide that service also varies over time. Examples are packet-switched routers where the packet sizes are variable and the available bandwidth fluctuates over time, web services where the available processing power fluctuates due to background processes or shared hosting, etc.

Most classical queueing models that try to take these two effects into account do this by using the notion of "service time", which is the amount of time that the server needs to fully serve one customer. This single notion, however, is not always adequate

M. De Muynck(✉) · H. Bruneel · S. Wittevrongel
Department of Telecommunications and Information Processing,
Stochastic Modeling and Analysis of Communication Systems Research Group,
Ghent University, Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium
e-mail: {michiel.demuynck,hb,sw}@telin.ugent.be

T.V. Do et al. (eds.), *Queueing Theory and Network Applications*,
Advances in Intelligent Systems and Computing 383,
DOI: 10.1007/978-3-319-22267-7_3

to model both effects. This is because if the amount of service that each customer requires (which we refer to as the "service demand" of the customer) varies from customer to customer, and the amount of work that the server is able to perform per time unit (which we refer to as the "service capacity" of the server) varies over time, then there may be a non-trivial correlation between the consecutive service times. In discrete-time queueing systems the notion of "service time" may even become ill-defined, as the amount of time that each customer spends in service may depend on the state of the queue and the service demands of the other customers.

In this paper, we analyze a discrete-time queue with variable service demands and variable service capacities. Time is divided into fixed-length intervals, called time slots, and both the service demands of the customers and the service capacities in each time slot are assumed to be integer numbers of "work units". The numbers of customer arrivals per slot are assumed to be independent and identically distributed (i.i.d.), as are the service demands per customer and the service capacities per slot.

This type of queueing model has been analyzed before, under various restrictions for the distribution of the service capacities. In [1] and [2], it was assumed that the service capacities follow a geometric distribution. In [3], the model was analyzed under the restriction that service capacities are deterministically equal to a given constant. Finally, in [4], this model was again analyzed, now with the restriction that the distribution of the service capacities has finite support. This was a generalization of [3], as the deterministic distribution has a support of 1. It was, however, not a generalization of [1] and [2], as the geometric distribution does not have finite support.

In the present paper, we analyze the queueing model under the assumption that the service capacities have a rational probability generating function (pgf). Note that all the restrictions on the service capacities in the previous papers [1]-[4] imply that the pgf of the service capacities must be rational. Additionally, all discrete phase-type distributions have a rational pgf, so the model is very widely applicable. The analysis in this paper is focused on the customer delay, and our main result is an expression for the pgf of the customer delay in steady state.

2 Queueing Model

The queueing model that we study in this paper is a discrete-time queueing model, where time is divided into contiguous fixed-length intervals, referred to as (time) slots. The numbers of arriving customers during the consecutive slots are i.i.d. from slot to slot. We denote their common pgf as $A(z)$ and their mean as $\lambda \triangleq A'(1)$.

Each customer has a *service demand*, expressed as a positive integer number of work units. This is exactly the amount of work that the server will have to perform, possibly over the course of multiple time slots, to completely serve the customer. The service demands of the customers are assumed to be i.i.d. from customer to customer. The common pgf of the service demands of the customers is denoted as $S(z)$. The mean service demand is denoted as $\tau \triangleq S'(1)$.

The number of work units that the server can execute in a time slot is referred to as the *service capacity* of the server during that time slot. These service capacities

are assumed to be non-negative integers that are i.i.d. from slot to slot, and we denote their common pgf as $R(z)$, which is assumed to be a rational function. The mean service capacity is denoted as $\mu \triangleq R'(1)$.

The server cannot initiate the service of a customer during the arrival slot of that customer. Stated otherwise, the service of a customer can start at the earliest during the slot following his arrival slot, even if the customer arrives in an empty system.

Customers from the queue are served sequentially by the server in first-come first-served (FCFS) order. In each slot, no more work units are performed than the available service capacity for that slot. If during a slot the available service capacity is less than the (remaining) service demand of the customer currently in service, then that customer's service simply continues in the next slot, with a reduced remaining service demand. Conversely, if the service capacity is greater than the remaining service demand of the customer currently in service, then that customer leaves the system and the server will use its remaining service capacity to immediately start the service of the next customer in the queue (if any). This is repeated until either the whole service capacity of the server during that slot has been used or there are no customers left in the queue that still require service.

The final assumption that we make is that the numbers of arrivals in each slot, the service demands of the customers, and the service capacities during each slot, are mutually independent.

3 Unfinished Work

Before deriving the expression for the pgf of the customer delay in this system, we first derive an expression for the pgf $U(z)$ of the unfinished work at the beginning of an arbitrary slot in steady state, i.e., the sum of (remaining) service demands of all the customers in the system at the beginning of the arbitrary slot.

We begin by introducing a notation for some of the random variables pertaining to the state of the system in slot k. We denote the unfinished work at the beginning of slot k as u_k, the number of customer arrivals during slot k as a_k, the service demand of the ith ($i = 1, 2, ..., a_k$) customer entering the system during slot k as $s_{k,i}$ and the service capacity during slot k as r_k. In every slot k, the following system equation holds between these random variables:

$$u_{k+1} = (u_k - r_k)^+ + \sum_{i=1}^{a_k} s_{k,i} \ , \tag{1}$$

where $(...)^+ = \max(..., 0)$. Taking the z-transform of both sides of this equation, taking the limit for $k \to \infty$ and using the fact that all the above random variables pertaining to the same slot k (i.e., u_k, a_k, r_k and the $s_{k,i}$'s) are all independent of each other, we obtain

$$U(z) = A(S(z)) \lim_{k\to\infty} E\left[z^{(u_k - r_k)^+}\right] \ . \tag{2}$$

Since u_k and r_k are independent, we can use a method based on complex contour integration, similar to the one presented in [5] (for the analysis of the classical discrete-time $G^{(G)}/Geo/1$ queue), to further work out the above equation. Under the assumption of a stable system, i.e., under the equilibrium condition $\lambda\tau < \mu$, the following equation is then obtained for the steady-state pgf $U(z)$ of the unfinished work at the beginning of a slot (see [5]):

$$U(z) = A(S(z)) \left[U(z)R(1/z) + (z-1) \sum_{\zeta \in \mathcal{S}_R^{-1}} F_\zeta(z) \right] . \tag{3}$$

Equation (3) is valid for all $z \notin \mathcal{S}_R^{-1}$, where $\mathcal{S}_R$ denotes the set of singularities (including, if applicable, ∞) of $R(z)$, $\mathcal{S}_R^{-1}$ denotes the inverse of this set, i.e., $\{z : 1/z \in \mathcal{S}_R\}$, and $F_\zeta(z)$ is defined as

$$F_\zeta(z) = \frac{1}{2\pi\imath} \oint_{C_\zeta} \frac{U(\xi)R(1/\xi)}{(\xi - z)(\xi - 1)} \, \mathrm{d}\xi \ , \tag{4}$$

with $\imath^2 = -1$ and C_ζ a small contour around ζ but not around any other singularity of $R(1/\xi)$, nor any singularity of $U(\xi)$, nor around 1 or z.

Let us now assume that the service-capacity pgf $R(z)$ is a rational function. Then all singularities ζ of $R(1/z)$ are poles and we can write

$$R(1/z) = \frac{P_R(z)}{Q_R(z)} = \frac{P_R(z)}{\prod_{\zeta \in \mathcal{S}_R^{-1}} (z-\zeta)^{\mu_\zeta}} \ , \tag{5}$$

wherein $P_R(z)$ and $Q_R(z)$ are two mutually prime polynomials and μ_ζ denotes the multiplicity of a singularity ζ. Note that the degree of $P_R(z)$ cannot be higher than the degree $m = \sum_{\zeta \in \mathcal{S}_R^{-1}} \mu_\zeta$ of $Q_R(z)$, since $\lim_{z\to\infty} R(1/z) = R(0) \in [0, 1]$. Therefore, using the expression for the residue of a complex function in a pole ζ with multiplicity μ_ζ, we easily find that the contour integral $F_\zeta(z)$ takes the form

$$F_\zeta(z) = \sum_{k=1}^{\mu_\zeta} \frac{c_k}{(z-\zeta)^k} \ , \tag{6}$$

for yet unknown constants c_k, or hence,

$$\sum_{\zeta \in \mathcal{S}_R^{-1}} F_\zeta(z) = \frac{N(z)}{Q_R(z)} \ , \tag{7}$$

with $N(z)$ an unknown polynomial of degree $m - 1$. Hence, we get

$$U(z) = \frac{(z-1)A(S(z))N(z)}{Q_R(z) - A(S(z))P_R(z)} \ . \tag{8}$$

Using Rouché's theorem it can be shown (see e.g. [6]) that the denominator $T(z) = Q_R(z) - A(S(z))P_R(z)$ has exactly m zeros inside or on the unit circle, one of which is equal to 1. Since $U(z)$ must remain bounded in these zeros, the numerator of $U(z)$ has to vanish as well, which completely determines the polynomial $N(z)$ and the pgf $U(z)$ except for a constant factor. With the normalization condition $U(1) = 1$, we finally get the following expression for $U(z)$:

$$U(z) = (\mu - \lambda\tau)\frac{(z-1)A(S(z))}{1 - R(1/z)A(S(z))} \prod_{\zeta \in \mathcal{S}_R^{-1}} \left(\frac{1-\zeta}{z-\zeta}\right)^{\mu_\zeta} \prod_{\xi \in \mathcal{N}_T^-} \left(\frac{z-\xi}{1-\xi}\right)^{n_\xi} \ , \tag{9}$$

where $\mathcal{N}_T^-$ denotes the set of zeros of $T(z)$ inside or on the unit circle, excluding the zero at $z = 1$, and n_ξ denotes the multiplicity of a zero ξ in this set.

4 Customer Delay

In this section, we derive an expression for the pgf $D(z)$ of the delay d_C that an arbitrary customer C experiences in the system in steady state, under a FCFS scheduling discipline. This delay is measured as the number of slots between the end of the arrival slot of the customer and the end of the slot during which the customer leaves. Note that the customer delay cannot be 0, since a customer that arrives during a slot cannot receive any service during that same slot.

We start the delay analysis with the derivation of the pgf of a related quantity, v_C, the unfinished work observed by the customer C upon arrival. It is defined as the total number of work units present in the system just after the arrival slot of customer C, but to be executed before or during the service of customer C. Mathematically, v_C is therefore defined as

$$v_C = (u_J - r_J)^+ + \sum_{i=1}^{f_C+1} s_{J,i} \ , \tag{10}$$

where J denotes the arrival slot of customer C, f_C is the number of customers that arrive in slot J but are to be served before C, and $s_{J,i}$ is the service demand of the ith customer in slot J. It is well-known (see e.g. [7]) that for any queue with independent, ordered arrivals, the pgf of f_C is given by $(A(z)-1)/(\lambda(z-1))$. Using this property and equation (2), the pgf $V(z)$ of v_C then follows immediately as

$$V(z) = \frac{U(z)}{A(S(z))} \cdot \frac{A(S(z)) - 1}{\lambda\,(S(z) - 1)} \cdot S(z) \ . \tag{11}$$

The delay d_C of customer C is related to the quantity v_C as follows. Customer C will still be in the system at the start of a slot if and only if fewer than v_C work units have been executed since the end of the arrival slot J of C. In other words,

$$d_C > k \quad \Leftrightarrow \quad v_C > r_{J+1} + r_{J+2} + \ldots + r_{J+k} \ . \tag{12}$$

We denote the above sum of k independent service capacities as $r_J^{(k)}$. Its pgf is given by $R(z)^k$. Using the fact that v_C and $r_J^{(k)}$ are independent, we then find

$$\frac{D(z)-1}{z-1} = \sum_{k=0}^{\infty} \text{Prob}[d_C > k] z^k = \sum_{k=0}^{\infty} \sum_{i=0}^{\infty} \sum_{j=0}^{i-1} \text{Prob}[v_C = i] \, \text{Prob}[r_J^{(k)} = j] z^k \ . \tag{13}$$

The inversion formula for probability generating functions states that

$$\text{Prob}[r_J^{(k)} = j] = \frac{1}{2\pi\imath} \oint_L \frac{R(\zeta)^k}{\zeta^{j+1}} \, \mathrm{d}\zeta \ , \tag{14}$$

where L is a contour around the origin such that $\forall \zeta \in L : |\zeta| < \mathcal{R}_R$, where $\mathcal{R}_X$ denotes the radius of convergence of a pgf $X(z)$. Note that the radius of convergence of $R(z)$ and that of $R(z)^k$ are equal. Equation (13) now reduces to

$$\begin{aligned} \frac{D(z)-1}{z-1} &= \frac{1}{2\pi\imath} \sum_{k=0}^{\infty} \sum_{i=0}^{\infty} \sum_{j=0}^{i-1} \oint_L \text{Prob}[v_C = i] \frac{\big(zR(\zeta)\big)^k}{\zeta^{j+1}} \, \mathrm{d}\zeta \\ &= \frac{1}{2\pi\imath} \sum_{k=0}^{\infty} \sum_{i=0}^{\infty} \oint_L \text{Prob}[v_C = i] \big(zR(\zeta)\big)^k \frac{\zeta^{-i}-1}{1-\zeta} \, \mathrm{d}\zeta \ . \end{aligned} \tag{15}$$

The above infinite summation of contour integrals is equal to the contour integral of the infinite series (i.e., we may "swap" the summation and integration symbols) if the contour L is chosen such that the resulting infinite series is uniformly convergent. It is important to question when such a contour can be constructed. This is the case if $\forall \zeta \in L : |1/\zeta| < \mathcal{R}_V$ and $|zR(\zeta)| < 1$. The former condition imposes a lower bound on $|\zeta|$, whereas the latter imposes an upper bound on $|R(\zeta)|$ that depends on z. Since this upper bound is most severe when ζ is real and positive, and since $R(\zeta)$ is an increasing function on the part of the real axis where $0 \le \zeta < \mathcal{R}_R$, the bounds can be rewritten as $R(1/\mathcal{R}_V) < R(|\zeta|) < |1/z|$. We conclude that a contour can be constructed if and only if $|z| < 1/R(1/\mathcal{R}_V)$. It follows that the radius of convergence $\mathcal{R}_D$ of $D(z)$ is given by $\mathcal{R}_D = 1/R(1/\mathcal{R}_V)$.

If $|z| < \mathcal{R}_D$, we may construct L as described above and bring the summations in (15) inside the integral. We obtain

$$\frac{D(z)-1}{z-1} = \frac{1}{2\pi\imath} \oint_L \frac{V(1/\zeta)-1}{(1-\zeta)\big(1-zR(\zeta)\big)} \, \mathrm{d}\zeta \ . \tag{16}$$

Substituting $z = 0$ in (16), in view of $D(0) = 0$, we get

$$1 = \frac{1}{2\pi\imath} \oint_L \frac{V(1/\zeta) - 1}{1 - \zeta} \, \mathrm{d}\zeta \ .$$

Using this result in (16) again, we find

$$D(z) = \frac{z}{2\pi\imath} \oint_L \frac{V(1/\zeta) - 1}{1 - \zeta} \cdot \frac{1 - R(\zeta)}{1 - zR(\zeta)} \, \mathrm{d}\zeta \ . \tag{17}$$

We can split the integrand into two terms, as follows:

$$\frac{V(1/\zeta)}{1 - \zeta} \frac{1 - R(\zeta)}{1 - zR(\zeta)} - \frac{1}{1 - \zeta} \frac{1 - R(\zeta)}{1 - zR(\zeta)} \ . \tag{18}$$

The latter term has no poles inside L, since L was chosen such that $\forall \zeta \in L : |zR(\zeta)| < 1$, which implies (due to Rouché's theorem) that $1 - zR(\zeta)$ has no zeros inside L, and since the simple zero of the denominator at $\zeta = 1$ (if that is inside L) is canceled by the zero of the numerator at $\zeta = 1$. We conclude that the contribution of the latter term to the value of the contour integral in (17) is zero. Therefore we can rewrite (17) as

$$D(z) = \frac{z}{2\pi\imath} \oint_L \frac{V(1/\zeta)}{1 - \zeta} \cdot \frac{1 - R(\zeta)}{1 - zR(\zeta)} \, \mathrm{d}\zeta \ . \tag{19}$$

Now we change the integration variable in (19) to $\xi = 1/\zeta$ (which yields a factor $-1/\xi^2$ in the integrand), and we invert the integration path L into L' but still integrate in counter-clockwise sense (which yields an extra factor of -1, since the inversion of L is a clockwise path). This leads to the expression

$$D(z) = \frac{z}{2\pi\imath} \oint_{L'} \frac{V(\xi)}{\xi \, (\xi - 1)} \cdot \frac{1 - R(1/\xi)}{1 - zR(1/\xi)} \, \mathrm{d}\xi \ , \tag{20}$$

where L' is a contour where $\forall \xi \in L' : \mathcal{R}_R^{-1} < |\xi| < \mathcal{R}_V$ and $|zR(1/\xi)| < 1$.

If the pgf of the service capacities is a rational function, i.e., if $R(z)$ is given by (5), then (20) can be rewritten as

$$D(z) = \frac{z}{2\pi\imath} \oint_{L'} \frac{V(\xi)}{\xi \, (\xi - 1)} \cdot \frac{Q_R(\xi) - P_R(\xi)}{Q_R(\xi) - zP_R(\xi)} \, \mathrm{d}\xi \ . \tag{21}$$

We now focus on the poles of the integrand in (21). Since the service demand of each customer is at least 1, $S(0)$ must equal 0, and by (11) $V(0)$ must equal 0 as well. The zero of the factor $V(\xi)$ in the numerator of the integrand at $\xi = 0$ then ensures that the factor ξ in the denominator does not cause a pole of the integrand at $\xi = 0$. Furthermore, the factor $Q_R(\xi) - P_R(\xi)$ ensures that the factor $(\xi - 1)$ in

the denominator does not cause a pole of the integrand at $\xi = 1$. Finally, since the contour L' was chosen such that $\forall \xi \in L' : |\xi| < \mathcal{R}_V$, $V(\xi)$ has no poles inside L' either. Therefore the only poles of the integrand are the zeros for ξ of

$$Q_R(\xi) - zP_R(\xi) \ , \tag{22}$$

or equivalently, of

$$1 - zR(1/\xi) \ . \tag{23}$$

We denote these zeros as $\alpha_k(z), k \in [0, N]$ for some N, and their multiplicities as $m_k, k \in [0, N]$. It can easily be seen that all these zeros lie inside L'. Indeed, the contour L' was chosen such that $\forall \xi \in L' : |zR(1/\xi)| < 1$. This implies that $|zP_R(\xi)| < |Q_R(\xi)|$, so using Rouché's theorem we can say that (22) has as many zeros inside L' as $Q_R(\xi)$. But all m zeros (counting with multiplicities) of $Q_R(\xi)$ must lie inside L', because L' was chosen such that $\forall \xi \in L' : |\xi| > 1/\mathcal{R}_R$. This means that (22) must have exactly m zeros for ξ inside L'. Moreover, these must be all the zeros of (22), since (22) is a polynomial in ξ of degree m (see Section 3).

Now we can apply Cauchy's Residue Theorem to calculate the value of the contour integral in (21). This gives

$$D(z) = z \sum_{k=0}^{N} \operatorname*{Res}_{\xi=\alpha_k(z)} \left[\frac{V(\xi)}{\xi\,(\xi - 1)} \cdot \frac{Q_R(\xi) - P_R(\xi)}{Q_R(\xi) - zP_R(\xi)} \right] \ , \tag{24}$$

where the residue of $\xi = \alpha_k(z)$ is given by

$$\frac{1}{(m_k - 1)!} \lim_{\xi \to \alpha_k(z)} \frac{\mathrm{d}^{m_k - 1}}{\mathrm{d}\xi^{m_k - 1}} \left[(\xi - \alpha_k(z))^{m_k} \frac{V(\xi)}{\xi\,(\xi - 1)} \cdot \frac{Q_R(\xi) - P_R(\xi)}{Q_R(\xi) - zP_R(\xi)} \right] \ . \tag{25}$$

Since all quantities in expression (24) are known or can be calculated numerically (when z is known), this expression may be used to evaluate $D(z)$ for any z. However, due to the $(m_k - 1)$st derivative with respect to ξ in (25), the evaluation of $D(z)$ may be difficult in practice if $m_k > 1$.

Note that if a zero ξ of (23) has a multiplicity $m_k > 1$, then $R'(1/\xi)/\xi = 0$. Since $R'(1/\xi)/\xi$ is the derivative of $R(1/\xi)$, a rational function with degree of the numerator and denominator at most m, there are at most $2m - 1$ values of ξ for which $R'(1/\xi)/\xi = 0$, with at most $2m - 1$ corresponding values of z (see (23)). Therefore, for all but at most $2m - 1$ values of z, the zeros $\alpha_k(z)$ are distinct, so that $m_k = 1$ for all k. For those z, a substantially simpler expression for $D(z)$ is available, because we can simplify (24) to

$$D(z) = \sum_{k=0}^{m-1} V\big(\alpha_k(z)\big) \frac{\alpha_k(z)}{\alpha_k(z) - 1} \frac{1 - R\big(1/\alpha_k(z)\big)}{R'\big(1/\alpha_k(z)\big)} \ . \tag{26}$$

Due to (23), we have that $R\big(1/\alpha_k(z)\big) = 1/z$. This allows to simplify (26) further to

$$D(z) = \frac{z-1}{z} \sum_{k=0}^{m-1} \frac{V\big(\alpha_k(z)\big)}{R'\big(1/\alpha_k(z)\big)} \cdot \frac{\alpha_k(z)}{\alpha_k(z)-1} \ . \tag{27}$$

Substituting the expressions (11) and (9) we previously found for $V(z)$ and $U(z)$, and again using (23) to simplify the result, we finally obtain

$$\begin{aligned} D(z) = \frac{\mu - \lambda\tau}{\lambda} \sum_{k=0}^{m-1} \frac{1-z}{R'\big(1/\alpha_k(z)\big)} \cdot \frac{S\big(\alpha_k(z)\big)}{S\big(\alpha_k(z)\big)-1} \cdot \frac{1 - A\big(S(\alpha_k(z))\big)}{z - A\big(S(\alpha_k(z))\big)} \cdot \alpha_k(z) \\ \cdot \prod_{\zeta \in \mathcal{S}_R^{-1}} \left(\frac{1-\zeta}{\alpha_k(z)-\zeta} \right)^{\mu_\zeta} \cdot \prod_{\xi \in \mathcal{N}_T^{-}} \left(\frac{\alpha_k(z)-\xi}{1-\xi} \right)^{n_\xi} \ . \end{aligned} \tag{28}$$

This expression still contains the functions $\alpha_k(z)$. These are zeros of an m-degree polynomial with coefficients that depend on z, and for these zeros a closed-form solution is generally not available. Therefore, inverting the pgf $D(z)$ analytically is very difficult. However, inverting this pgf numerically, using methods such as those described in [8], is very straight-forward.

Additionally, expression (28) can be used to calculate the expected value and other moments of the delay, in view of the moment-generating property of pgfs.

5 Numerical Examples

In this section, we briefly give some illustrative numerical examples. In the first example, shown in Fig. 1, we study the impact of the service-capacity distribution on the mean customer delay under varying loads $\rho = \lambda\tau/\mu$.

We consider 4 different service-capacity distributions, all with mean $\mu = 10$: deterministic ($R_1(z) = z^{10}$), negative binomial with parameter $r = 5$ ($R_2(z) = 1/(3-2z)^5$), geometric ($R_3(z) = 1/(11-10z)$) and a weighted mixture of 2 geometric distributions with means 5 and 30 such that the overall mean $\mu = 10$,

$$R_4(z) = \frac{26 - 25z}{(6-5z)(31-30z)}.$$

The variances of these 4 distributions are respectively 0, 30, 110, and 310.

In Fig. 1 it can be seen that, generally, a higher variance of the service capacity leads to a higher mean delay. This is to be expected, since more variability on the service capacities should in general lead to a burstier service process, which should in turn cause longer queues. However, in Fig. 1 there is also one case where the opposite is true: under low load, the system with negative binomially distributed service capacities has a lower mean delay than the system with deterministic capacities, despite the fact that the deterministic distribution has the lowest variance. Under high

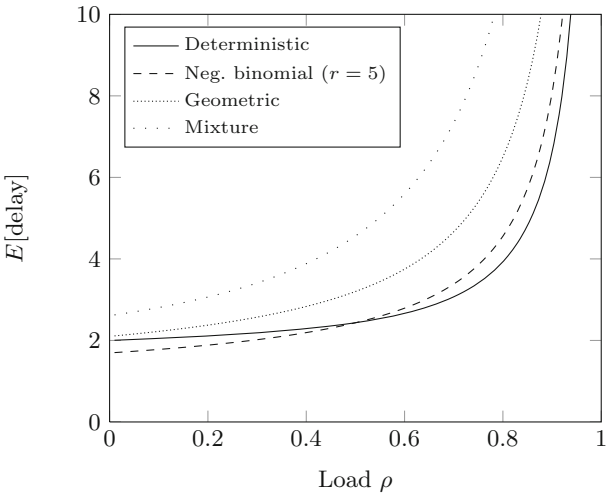


Fig. 1 Mean customer delay for a queue with Poisson arrivals with varying λ, deterministic service demands of 11 work units and various distributions for the service capacities (as indicated), all with mean $\mu = 10$.

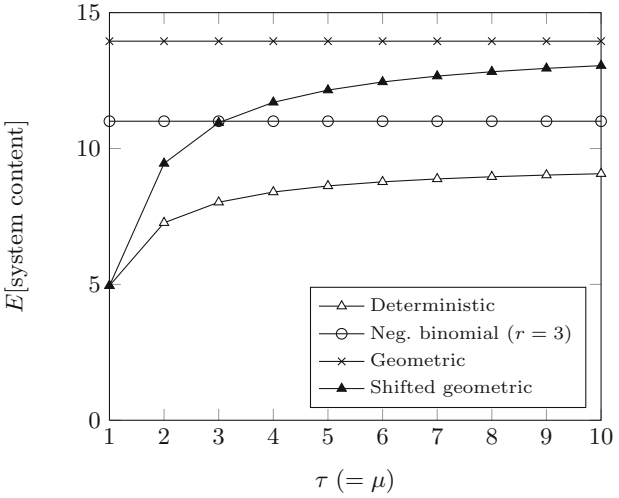


Fig. 2 Mean system content for a queue with Poisson arrivals with $\lambda = 0.9$, shifted geometric (i.e., with minimum 1) service demands with varying mean τ and various distributions for the service capacities (as indicated), with mean $\mu = \tau$.

load it is the other way around again. This is because for the system with deterministic service capacities, the lack of variance in the service process is helpful under high load, but it makes the minimum delay experienced by each customer at least 2 slots, whereas a system with variable service capacities may under low load achieve

a mean customer delay below 2. The exact influence of the coefficient of variation on the mean delay is very non-linear and depends on the zeros $\mathcal{S}_R^{-1}$, $\mathcal{N}_T^-$ and $\alpha_k(1)$.

In a second example, shown in Fig. 2, we keep the ratio τ/μ and all other system parameters fixed and scale τ and μ together to observe the impact of their actual values on the mean system content (obtained from the mean delay through Little's law $E[\text{system content}] = \lambda E[\text{delay}]$).

Note that for geometric service capacities (and shifted geometric demands), the mean system content does not depend on the actual values of τ and μ but merely on their ratio. This effect was already observed in [1] and was called the "geometric invariance property". However, note that while it holds for geometric service capacities, it does not hold for shifted geometric service capacities, as the mean system content clearly depends on μ. From Fig. 2 it can also be seen that the invariance property also holds for negative binomial service capacities.

Acknowledgments This research has been partly funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office.

References

1. Bruneel, H., Wittevrongel, S., Claeys, D., Walraevens, J., Discrete-time queues with variable service capacity: a basic model and its analysis, Annals of operations research (2013) (accepted for publication). doi:10.1007/s10479-013-1428-y
2. Walraevens, J., Bruneel, H., Claeys, D., Wittevrongel, S.: The discrete-time queue with geometrically distributed service capacities revisited. In: Dudin, A., De Turck, K. (eds.) ASMTA 2013. LNCS, vol. 7984, pp. 443–456. Springer, Heidelberg (2013)
3. Bruneel, H., Rogiest, W., Walraevens, J., Wittevrongel, S.: On queues with general service demands and constant service capacity. In: Norman, G., Sanders, W. (eds.) QEST 2014. LNCS, vol. 8657, pp. 210–225. Springer, Heidelberg (2014)
4. De Muynck, M., Wittevrongel, S., Bruneel, H., A discrete-time queue with finite-support service capacities. In: Book of Abstracts of ECQT 2014, Ghent, p. 34, August 20–22, 2014
5. Vinck, B., Bruneel, H.: Analyzing the discrete-time $G^{(G)}/Geo/1$ queue using complex contour integration. Queueing Systems **18**(1–2), 47–67 (1994)
6. Adan, I.J.B.F., van Leeuwaarden, J.S.H., Winands, E.M.M.: On the applicationof Rouchè's theorem in queueing theory. Oper. Res. Letters, 355–360 (20050
7. Mitrani, I.: Modelling of Computer and Communication Systems. Cambridge University Press, Cambridge (1987)
8. Abate, J., Whitt, W.: Numerical inversion of probability generating functions. Oper. Res. Letters **12**, 245–251 (1992)

A Queueing Approximation of MMPP/PH/1

Azam Asanjarani and Yoni Nazarathy

Abstract We consider the well-studied MMPP/PH/1 queue and illustrate a method to find an almost equivalent model, the MTCP/PH/1. MTCP stands for Markovian Transition Counting Process. It is a counting process that has similar characteristics to MMPP (Markov Modulated Poisson Process). We prove that for a class of MMPPs there is an equivalent class of MTCPs. We then use this property to suggest an approximation for MMPP/PH/1 in terms of the first two moments. We numerically show that the steady state characteristics of MMPP/PH/1 are well approximated by the associated MTCP/PH/1 queue. Our numerical analysis leaves some open problems on bounds of the approximations. Of independent interest, this paper also contains a lemma on the workload expression of MAP/PH/1 queues which to the best of our knowledge has not appeared elsewhere.

Keywords Markov modulated poisson process · MAP/PH/1 · Queueing

1 Introduction

Queuing theory finds a variety of applications such as telecommunication networks, healthcare and manufacturing, see for instance [6]. One of the most useful queueing models is the MAP/PH/1 queue, see for example [14]. The Markovian Arrival Process (MAP) is a counting process based on a background finite-state Continuous-Time Markov Chain (CTMC). MAP can be considered as a generalisation of the Poisson process where the inter-arrival times of a MAP are not necessarily independent of each other, nor exponentially distributed. The Phase type (PH) distribution is a generalization of the exponential distribution and is based on the distribution of time until absorption in a finite-state CTMC. These two matrix-analytic objects make up the MAP/PH/1 queue: the arrival process is MAP, and the service times are assumed i.i.d. from a PH distribution.

A. Asanjarani(✉) · Y. Nazarathy
The University of Queensland, Brisbane, Queensland, Australia
e-mail: a.asanjarani@uq.edu.au

T.V. Do et al. (eds.), *Queueing Theory and Network Applications*,
Advances in Intelligent Systems and Computing 383,
DOI: 10.1007/978-3-319-22267-7_4

Comparison of different stochastic processes to find a versatile model for describing observed data in an accurate manner is a fundamental objective in stochastic modelling. In modelling a variety of phenomena such as queueing processes, the Markov Modulated Poisson Process (MMPP), a special case of MAP, can be applied. The MMPP has a variety of applications in modelling bursty traffic. The motivation behind the vast applications of MMPP is that MMPP keeps the tractability of the Poisson process while enabling non-zero correlation between inter-arrival times. See for example [5], [10] and [13].

In this paper we introduce an alternative model to MMPP which we refer to as the Markovian Transition Counting Process (MTCP). MTCP is a MAP which counts every transition of the background CTMC. We believe it is more tractable and more computationally convenient than the MMPP. We find relations between MTCP and MMPP, focusing on the case of a two state background CTMC for the MMPP. We prove that in some cases, the first two moments of MMPP and MTCP can be matched. We refer to these cases as slow MMPPs. This implies that the intensity of arrivals is greater than the total intensity of state changes per state. From a modelling perspective, slow MMPPs are perhaps the most useful MMPPs because non-slow MMPPs have characteristics quite similar to the Poisson process.

In using MTCP for queues, we investigate the behaviour of the MTCP/PH/1 queue as an alternative to the MMPP/PH/1 queue. Here, we address this question empirically through extensive numerical experiments. We show that the basic steady state characteristics (mean and variance of the queue) of a given MMPP/PH/1 queue can be emulated by an MTCP/PH/1 queue almost without relative error in most cases, and with relative errors that are bounded at the worst case by 9%. These preliminary results are significant for the emerging body of research dealing with finding alternative (but similar) queueing models.

As a stochastic modeller chooses a suitable queueing model for a given situation, there is typically more than one choice. Knowing that MTCP/PH/1 is similar to MMPP/PH/1 allows the modeller to have more freedom in model choice. In future research we shall integrate this within a statistical model-selection framework, fitting queueing models to data. Towards that end, a key advantage of using MTCP/PH/1 instead of MMPP/PH/1 is that the MTCP is more informative than the MMPP. In fact we believe that our MTCP is better suited for parameter estimation since for this model, each observed event corresponds to exactly one transition in the background (unobserved) CTMC.

The remainder of this paper is structured as follows: In Section 2 we overview the MMPP/PH/1 queue and treat it as a Quasi-Birth-Death (QBD) process. We also present a lemma on the workload expression of MAP/PH/1 queues which to the best of our knowledge has not appeared elsewhere. In Section 3 we introduce the new model, MTCP, as a special MAP. In Section 4 we show that for a slow MMPP_2, a useful substitute MTCP_4 exists. In fact, we prove that the first and second moments of these two model classes (slow MMPP and MTCP) can be matched. In Section 5 numerical results for approximating a given $\text{MMPP}_2/\text{PH}_2/1$ with an $\text{MTCP}_4/\text{PH}_2/1$ are presented. We conclude in Section 6.

2 The MMPP/PH/1 Queue

The MMPP/PH/1 queue is a special case of the general single-server queue MAP/G/1, where the stream of arrivals and service mechanism are modelled by MMPP and PH distribution respectively. Figure 1 illustrates an example of an $\text{MMPP}_2/\text{PH}_2/1$ queue. Methods of analysing the MMPP/PH/1 queueing models can be found in [7] and [10]. In this paper we use the uniform framework of QBD processes which is an efficient way to analyse more general models using matrix-analytic methods, see [9].

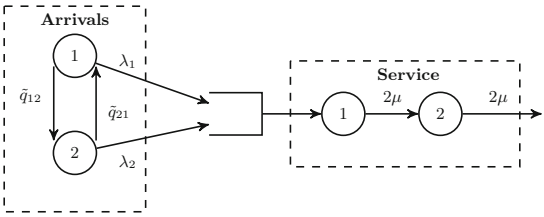


Fig. 1 A schematic illustration of the $\text{MMPP}_2/\text{E}_2/1$ queue (E_2 is a special case of PH_2 and stands for Erlang, where in this case it has a mean of μ^{-1}). The circles illustrate phases of the arrival and/or service mechanism.

MMPP. An MMPP is simply an arrival process which consists of a finite number of Poisson processes, modulated by a CTMC. In other words, MMPP is a special case of a doubly stochastic Poisson process whose arrival rate is modulated by the states of an irreducible finite-state CTMC, which is referred to as the phase process. The parameters of an MMPP of order p are the vector of Poisson arrival rates associated with each phase, $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_p)'$ as well as the parameters of the p-state background CTMC: the transition rate matrix Q and the initial distribution of the background CTMC, taken as a row vector $\boldsymbol{\alpha}$.

PH Distribution. The time until absorption into state 0 (absorbing state) of a finite-state CTMC with q transient state and one absorbing state is said to have a phase type (PH) distribution of order q. A PH distribution of order q is parametrised by $\boldsymbol{\eta}$ and T, where $\boldsymbol{\eta}$ is the initial distribution over the transient states (taken as a row vector) and the matrix $T = \{t_{ij}\}_{i,j=1,\ldots,q}$ specifies the transition rates between the transient states of the CTMC. PH distributions are very versatile and are dense in the class of distributions defined on the non-negative real numbers [3]. Moreover, PH distributions are used in a wide range of applications, see for instance [1] and [8].

QBD and MMPP/PH/1. A continuous-time homogeneous QBD_r is a Markov process characterised by a two dimensional state space $\{(n, i) : 0 \leq n, 1 \leq i \leq r\}$, which are called the level and the phase of the state, respectively. A transition from (n, i) to (n', j) is possible only when $|n' - n| < 2$ and the transition rate from (n, i) to (n', j) may depend on i, j and $|n' - n|$, but not on the specific values of n and n'. When ordering the states in lexicographic order, the transition rate matrix of a

QBD_r has the following form:

$$A = \begin{pmatrix} B_0 & B_1 & & & 0 \\ B_{-1} & A_0 & A_1 & & \\ & A_{-1} & A_0 & A_1 & \\ & & A_{-1} & A_0 & A_1 \\ 0 & & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{1}$$

In representing the $\mathrm{MMPP}_p/\mathrm{PH}_q/1$ queue as a QBD_r, where $r = p \times q$, the phase records (in lexicographic order) both the background state of the MMPP (arrival) and the current phase of the service (see Figure 1 for illustration of the phases in the special case of $\mathrm{MMPP}_2/\mathrm{E}_2/1$). The level, represents the number of items in the system.

Modelled as a QBD, we have:

$$B_{-1} = I_p \otimes \mathbf{t}, \quad B_0 = C, \quad B_1 = \mathrm{diag}(\boldsymbol{\lambda}) \otimes \boldsymbol{\eta},$$

where $C = Q - \mathrm{diag}(\boldsymbol{\lambda})$ and where $\otimes$ is the Kronecker product. Here $\mathbf{t} = -T\mathbf{1}$, where $\mathbf{1}$ is a column vector of 1's with appropriate dimension.

Further,

$$A_{-1} = I_p \otimes \mathbf{t}\boldsymbol{\eta}, \quad A_0 = I_p \otimes T + C \otimes I_q, \quad A_1 = \mathrm{diag}(\boldsymbol{\lambda}) \otimes I_q.$$

As is well known in the theory of QBDs, the stationary distribution of a positive-recurrent QBD, $\boldsymbol{\pi}$, admits a matrix-geometric form $\pi_n = \pi_{n-1}R$, where R is the solution of a quadratic fixed-point matrix equation $R = A_1 + RA_0 + R^2A_{-1}$ and π_n are row vectors of dimension r, see [9]. We use the-state-of-the-art SMC solver to find the matrix R and the stationary distribution of a given QBD, see [4]. It is easy to show that A is irreducible due to the properties of the building blocks and irreducibility of Q. Moreover, characterizing the positive-recurrence can be done as follows[1].

Lemma 1. *The QBD representing a $MAP_p/PH_q/1$ queue is positive-recurrent if and only if,*

$$\rho := \frac{\boldsymbol{\beta} A_1 \mathbf{1}}{\boldsymbol{\beta} A_{-1} \mathbf{1}} = \frac{\Lambda}{\frac{1}{-\boldsymbol{\eta} T^{-1} \mathbf{1}}} < 1,$$

where $-\boldsymbol{\eta} T^{-1}\mathbf{1}$ is the first moment of PH_q with parameters $(\boldsymbol{\eta}, T)$, $\Lambda = \boldsymbol{\pi} D\mathbf{1}$ is the first moment of a time-stationary MAP_p with parameters $(\boldsymbol{\pi}, C, D)$[2], and $\boldsymbol{\beta}$ is the stationary distribution of $A_{-1} + A_0 + A_1$.

[1] To the best of our knowledge, the algebra behind this intuitive lemma has not appeared elsewhere.

[2] The QBD representation of $\mathrm{MAP}_p/\mathrm{PH}_q/1$ generalises the $\mathrm{MMPP}_p/\mathrm{PH}_q/1$ representation, with $\mathrm{diag}(\boldsymbol{\lambda})$ being replaced by D (see next Section for MAPs).

Proof. The fact that the left hand side of ρ is a necessary and sufficient condition for positive recurrence follows from the theory of QBDs (see [9], Theorem 7.2.4). It remains to show that both representations of ρ agree.

First we show that $\boldsymbol{\beta} = \boldsymbol{\pi} \otimes \boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ is the unique solution of $\boldsymbol{\gamma}(T + \mathbf{t}\boldsymbol{\eta}) = \mathbf{0}'$ and $\boldsymbol{\gamma}\mathbf{1} = 1$[3]. It is immediate that $(\boldsymbol{\pi} \otimes \boldsymbol{\gamma})\mathbf{1} = 1$. Further, we have

$$\begin{aligned}(\boldsymbol{\pi} \otimes \boldsymbol{\gamma})(A_{-1} + A_0 + A_1) &= (\boldsymbol{\pi} \otimes \boldsymbol{\gamma})\big(I_p \otimes \mathbf{t}\boldsymbol{\eta} + (I_p \otimes T + C \otimes I_q) + D \otimes I_q\big) \\ &= (\boldsymbol{\pi} \otimes \boldsymbol{\gamma})\big(I_p \otimes (\mathbf{t}\boldsymbol{\eta} + T) + (C + D) \otimes I_q\big) \\ &= (\boldsymbol{\pi} \otimes \boldsymbol{\gamma})\big((C + D) \otimes (\mathbf{t}\boldsymbol{\eta} + T)\big) \\ &= \mathbf{0}',\end{aligned}$$

where the last two steps follow since $(A \otimes B)(A' \otimes B') = AA' \otimes BB'$ when matrix dimensions agree for the multiplication.

Now we need to show that $\frac{\boldsymbol{\beta} A_1 \mathbf{1}}{\boldsymbol{\beta} A_{-1} \mathbf{1}} = \frac{\Lambda}{\frac{1}{-\boldsymbol{\eta} T^{-1} \mathbf{1}}}$ or equivalently:

$$\boldsymbol{\beta} A_1 \mathbf{1} = (\boldsymbol{\beta} A_{-1} \mathbf{1})\, \Lambda\, (-\boldsymbol{\eta} T^{-1} \mathbf{1})$$

which for the $\text{MAP}_p/\text{PH}_q/1$ queue is written as:

$$(\boldsymbol{\pi} \otimes \boldsymbol{\gamma})(D \otimes I_q)\mathbf{1} = (\boldsymbol{\pi} \otimes \boldsymbol{\gamma})(I_p \otimes \mathbf{t}\boldsymbol{\eta})\mathbf{1}(\boldsymbol{\pi} D \mathbf{1})(-\boldsymbol{\eta} T^{-1}\mathbf{1}). \tag{2}$$

For the left hand side, we have

$$(\boldsymbol{\pi} \otimes \boldsymbol{\gamma})(D \otimes I_q)\mathbf{1} = (\boldsymbol{\pi} D \otimes \boldsymbol{\gamma})\mathbf{1} = \boldsymbol{\pi} D \mathbf{1}.$$

Therefore we need to show that the right hand side of (2) is equal to $\boldsymbol{\pi} D \mathbf{1}$, or equivalently:

$$(\boldsymbol{\pi} \otimes \boldsymbol{\gamma})(I_p \otimes \mathbf{t}\boldsymbol{\eta})\mathbf{1}(-\boldsymbol{\eta} T^{-1}\mathbf{1}) = 1.$$

Since $\boldsymbol{\pi}\mathbf{1} = \boldsymbol{\eta}\mathbf{1} = 1$, we have $(\boldsymbol{\pi} \otimes \boldsymbol{\gamma})(I_p \otimes \mathbf{t}\boldsymbol{\eta})\mathbf{1} = (\boldsymbol{\pi} \otimes \boldsymbol{\gamma}\mathbf{t}\boldsymbol{\eta})\mathbf{1} = \boldsymbol{\gamma}\mathbf{t}$. Moreover, from $\boldsymbol{\gamma}(T + \mathbf{t}\boldsymbol{\eta}) = \mathbf{0}'$ we have $\boldsymbol{\gamma}\mathbf{t}\boldsymbol{\eta} = -\boldsymbol{\gamma} T$ which results in $\boldsymbol{\gamma}\mathbf{t}(-\boldsymbol{\eta} T^{-1}\mathbf{1}) = 1$. □

3 MAPs and the Markovian Transition Counting Process

MAP. A MAP is a pure birth process which can be considered as a special case of the QBD: a MAP is a two-dimensional Markov process with parameters $(\boldsymbol{\alpha}, C, D)$ where $\boldsymbol{\alpha}$ is the initial distribution of the finite-state CTMC and the matrix $C = B_0 = A_0$ records the transitions of the background CTMC with no arrival. The event intensity matrix $D = B_1 = A_1$ has non-negative elements and describes the transitions of the background CTMC with an arrival. The matrices A_{-1} and B_{-1} are zero matrices.

[3] Note that $\boldsymbol{\gamma}$ is the limiting distribution of the phase in a PH_q-renewal process.

Moreover, we have $C + D = Q$, where Q is the transition rate matrix of the CTMC. A MAP with parameters $(\boldsymbol{\alpha}, C, D)$ is time-stationary if $\boldsymbol{\alpha} = \boldsymbol{\pi}$, where $\boldsymbol{\pi}$ is the stationary distribution of the phase process, i.e. $\boldsymbol{\pi} Q = \mathbf{0}'$ and $\boldsymbol{\pi}\mathbf{1} = 1$.

For a time-stationary MAP, the mean and variance of the number of counts at any time t are given by the following formulas, see Chapter XI of [3]:

$$\mathbb{E}[N(t)] = \Lambda\, t = \boldsymbol{\pi} D \mathbf{1}\, t, \tag{3}$$

$$\mathrm{Var}\Big(N(t)\Big) = \{\Lambda - 2\Lambda^2 + 2\,\boldsymbol{\pi} D Q^{-} D\mathbf{1}\}\, t + 2\,\boldsymbol{\pi} D Q^{-}(e^{Qt} - I) Q^{-} D\mathbf{1}, \tag{4}$$

where $Q^{-} = (\mathbf{1}\boldsymbol{\pi} - Q)^{-1}$.

The class of MAPs contains most of the commonly used point processes such as the Poisson process ($D = \lambda$, where λ is the Poisson rate and $C = -\lambda$) and MMPP ($D = \mathrm{diag}(\boldsymbol{\lambda})$ where $\boldsymbol{\lambda}$ is the vector of Poisson rates and $C = Q - D$). In this research, we introduce and investigate a class of MAPs as follows:

Definition 1. *A **Markovian Transition Counting Process (MTCP)** is a two-dimensional Markov process $\{(\bar{N}(t), X(t)); t \geq 0\}$ where $\bar{N}(t)$ counts every transition of an irreducible CTMC $X(\cdot)$ on $[0, t]$.*

Therefore MTCP is a special type of MAP where we have $\bar{D} = \bar{Q} - \mathrm{diag}(\bar{Q})$ and the parameters of MTCP are just the parameters of the background CTMC. MTCPs and MMPPs are in a sense the extreme cases of MAPs. In an MMPP, the events do not coincide with state transitions (with probability 1). In contrast, in an MTCP the events are precisely all the transitions of the CTMC. This fact motivates the idea of finding relations between MTCP and MMPP. An early reference that analyses both MTCPs and MMPPs (although not using these names) is [12]. We now show some further relations.

4 Relations between MTCP and MMPP

In Proposition 3.2 of [11], the authors showed that every MTCP has an associated MMPP with the same two first moments. For completeness, we present this proposition of [11] in an alternative form here, including the proof.

Proposition 1. *Let $\bar{N}(t)$ be the counting processes of a time-stationary MTCP$_p$. Then there is an MMPP$_p$, with the counting processes $\tilde{N}(t)$, such that their first and second moments are matched. That is, for $\forall t \geq 0$,*

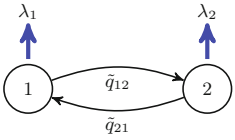
$$\mathbb{E}[\tilde{N}^k(t)] = \mathbb{E}[\bar{N}^k(t)]\,, \qquad \textit{for} \quad k = 1, 2\,.$$

Proof. Assume that the event matrix of the MTCP$_p$ is given by $\bar{D} = \bar{Q} - \mathrm{diag}(\bar{Q})$. Consider an MMPP$_p$ with the same background Markov chain and set $\tilde{D} = -\mathrm{diag}(\bar{Q})$. Now from (3) and (4), we just need to show that $\bar{D}\mathbf{1} = \tilde{D}\mathbf{1}$ and $\boldsymbol{\pi}\bar{D} = \boldsymbol{\pi}\tilde{D}$. Since $\bar{Q}\mathbf{1} = 0$ and $\boldsymbol{\pi}\bar{Q} = \mathbf{0}'$ the result follows. □

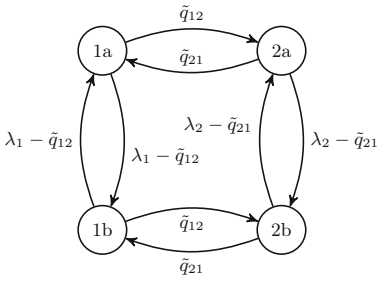
The proof shows that in order to construct an MMPP matching an MTCP with parameter $\bar{Q}$: set $\boldsymbol{\lambda} = -\text{diag}(\bar{Q})$ and $\tilde{Q} = \bar{Q}$. The question is now how to construct MTCPs matching MMPPs. Based on the above proposition, the answer is given for the special case of MMPPs where $\boldsymbol{\lambda} = -\text{diag}(\tilde{Q})$, i.e. $\lambda_i = \sum_{j \neq i} \tilde{q}_{ij}$. But this is a very restricted case since it does not leave any freedom with λ_i.

We now show that for each instance in a class of MMPPs (of order 2), where $\lambda_i > \sum_{j \neq i} \tilde{q}_{ij}$ which we call "slow MMPPs", there is an associated MTCP (of order 4) that exhibits the same first and second moments for the counting process. We believe a similar construction holds for arbitrary $p > 2$ (relating MTCP_{2p} to MMPP_p), this remains the subject of future work.

Definition 2. *A **slow Markov Modulated Poisson Process (slow MMPP)** is an MMPP where for any phase i in the phase process, the arrival rate is greater than the rate of leaving that phase, i.e. $\lambda_i > \sum_{j \neq i} \tilde{q}_{ij}$.*



(a) Transition diagram of the phase process of an MMPP_2.



(b) Transition diagram of the phase process of related MTCP_4.

Fig. 2 An MMPP_2 and its associated MTCP_4

Given MMPP parameters, $\boldsymbol{\lambda}$ and $\tilde{Q}$, we can associate an MTCP_4 to any slow MMPP_2 as illustrated in Figure 2. The transition rate matrix $\bar{Q}$ and the event intensity matrix $\bar{D}$ of the associated MTCP_4 are given as follows:

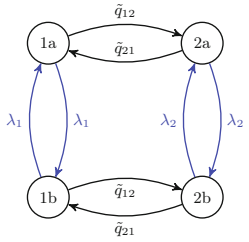
$$\bar{Q} = \left(\begin{array}{cc|cc} -\lambda_1 & \lambda_1 - \tilde{q}_{12} & \tilde{q}_{12} & 0 \\ \lambda_1 - \tilde{q}_{12} & -\lambda_1 & 0 & \tilde{q}_{12} \\ \hline \tilde{q}_{21} & 0 & -\lambda_2 & \lambda_2 - \tilde{q}_{21} \\ 0 & \tilde{q}_{21} & \lambda_2 - \tilde{q}_{21} & -\lambda_2 \end{array}\right), \quad \bar{D} = \bar{Q} - \text{diag}(\bar{Q}). \tag{5}$$

We now have the following:

Proposition 2. *Let $\tilde{N}(t)$ and $\bar{N}(t)$ be the counting processes of a time-stationary slow MMPP_2 and its associated MTCP_4, respectively. Then, these processes have the same first and second moment. That is, for $\forall t \geq 0$,*

$$\mathbb{E}[\tilde{N}^k(t)] = \mathbb{E}[\bar{N}^k(t)]\,, \qquad \textit{for} \quad k = 1, 2\,.$$

Proof. We first construct a MAP_4 with the same counting process as the MMPP_2 by coupling the events of the phase process of MMPP_2. When the process is in phase k, Figure 3 shows the structure of a coupled MAP that results in transition from phase k_a to k_b or vice versa. $\tilde{Q}$ is the phase transition matrix and $\tilde{D}$ is the event intensity matrix of the resulting MAP_4.



$$\tilde{Q} = \left(\begin{array}{cc|cc} -(\lambda_1+\tilde{q}_{12}) & \lambda_1 & \tilde{q}_{12} & 0 \\ \lambda_1 & -(\lambda_1+\tilde{q}_{12}) & 0 & \tilde{q}_{12} \\ \hline \tilde{q}_{21} & 0 & -(\lambda_2+\tilde{q}_{21}) & \lambda_2 \\ 0 & \tilde{q}_{21} & \lambda_2 & -(\lambda_2+\tilde{q}_{21}) \end{array}\right)$$

$$\tilde{D} = \begin{pmatrix} 0 & \lambda_1 & 0 & 0 \\ \lambda_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_2 \\ 0 & 0 & \lambda_2 & 0 \end{pmatrix}$$

Fig. 3 Transition diagram of the phase process of the coupled MAP_4 and its matrices

To find the stationary distribution of the associated MAP_4, $\tilde{\boldsymbol{\pi}}$, we need to solve $\tilde{\boldsymbol{\pi}}\tilde{Q} = \mathbf{0}'$, $\tilde{\boldsymbol{\pi}}\mathbf{1} = 1$. In the same way, we can find the stationary distribution of MTCP_4, $\bar{\boldsymbol{\pi}}$, i.e. we have the following systems of equations:

$$\begin{cases} -(\lambda_1+\tilde{q}_{12})\tilde{\pi}_1 + \lambda_1\tilde{\pi}_2 + \tilde{q}_{21}\tilde{\pi}_3 = 0 \\ \lambda_1\tilde{\pi}_1 - (\lambda_1+\tilde{q}_{12})\tilde{\pi}_2 + \tilde{q}_{21}\tilde{\pi}_4 = 0 \\ \tilde{q}_{12}\tilde{\pi}_1 - (\lambda_2+\tilde{q}_{21})\tilde{\pi}_3 + \lambda_2\tilde{\pi}_4 = 0 \\ \tilde{\pi}_1 + \tilde{\pi}_2 + \tilde{\pi}_3 + \tilde{\pi}_4 = 1 \end{cases} \qquad \begin{cases} -\lambda_1\bar{\pi}_1 + (\lambda_1-\tilde{q}_{12})\bar{\pi}_2 + \tilde{q}_{21}\bar{\pi}_3 = 0 \\ (\lambda_1-\tilde{q}_{12})\bar{\pi}_1 - \lambda_1\bar{\pi}_2 + \tilde{q}_{21}\bar{\pi}_4 = 0 \\ \tilde{q}_{12}\bar{\pi}_1 - \lambda_2\bar{\pi}_3 + (\lambda_2-\tilde{q}_{21})\bar{\pi}_4 = 0 \\ \bar{\pi}_1 + \bar{\pi}_2 + \bar{\pi}_3 + \bar{\pi}_4 = 1 \end{cases}$$

Both of the above are uniquely solved by

$$\pi_1 = \pi_2 = \frac{\tilde{q}_{21}}{2(\tilde{q}_{12}+\tilde{q}_{21})}, \qquad \text{and} \qquad \pi_3 = \pi_4 = \frac{\tilde{q}_{12}}{2(\tilde{q}_{12}+\tilde{q}_{21})}.$$

Therefore, these two processes have the same stationary distribution $\boldsymbol{\pi}$.

Now since $\tilde{D}\mathbf{1} = \bar{D}\mathbf{1} = (\lambda_1, \lambda_1, \lambda_2, \lambda_2)'$ one can find from (3):

$$\mathbb{E}[\tilde{N}(t)] = \mathbb{E}[\bar{N}(t)].$$

To compute the variance, first we verify that:

$$\boldsymbol{\pi}\tilde{D} = \boldsymbol{\pi}\bar{D} = \left(\frac{\tilde{q}_{21}\lambda_1}{2(\tilde{q}_{12}+\tilde{q}_{21})} \quad \frac{\tilde{q}_{21}\lambda_1}{2(\tilde{q}_{12}+\tilde{q}_{21})} \quad \frac{\tilde{q}_{12}\lambda_2}{2(\tilde{q}_{12}+\tilde{q}_{21})} \quad \frac{\tilde{q}_{12}\lambda_2}{2(\tilde{q}_{12}+\tilde{q}_{21})}\right).$$

Explicit calculation of the fundamental matrices $\tilde{Q}^-$ and $\bar{Q}^-$ shows that even though these matrices are not the same, it holds that $\boldsymbol{\pi}\tilde{D}\tilde{Q}^-\tilde{D}\mathbf{1} = \boldsymbol{\pi}\bar{D}\bar{Q}^-\bar{D}\mathbf{1}$. In addition, by explicitly calculating the matrix exponential, we have:

$$\boldsymbol{\pi}\tilde{D}\tilde{Q}^-(e^{\tilde{Q}t} - I)\tilde{Q}^-\tilde{D}\mathbf{1} = \frac{(e^{-(\tilde{q}_{12}+\tilde{q}_{21})t} - 1)\tilde{q}_{12}\tilde{q}_{21}(\lambda_1 - \lambda_2)^2}{(\tilde{q}_{12}+\tilde{q}_{21})^4} = \boldsymbol{\pi}\bar{D}\bar{Q}^-(e^{t\bar{Q}} - I)\bar{Q}^-\tilde{D}\mathbf{1}.$$

Therefore, from (4), $\mathrm{Var}\left(\tilde{N}(t)\right) = \mathrm{Var}\left(\bar{N}(t)\right)$ and the proof is complete. □

Remark 1. Note that Proposition 2 only holds for slow MMPPs. Otherwise the construction of a MAP_4 from a given MMPP_2 does not hold due to some non-positive off-diagonal elements $\lambda_i - \tilde{q}_{ij}$ in the matrices $\bar{Q}$ and $\bar{D}$.

5 The Steady-State Queue Approximation

In this section we use the results of the previous section to approximate a given (slow) $\mathrm{MMPP}_2/\mathrm{PH}_2/1$ with an $\mathrm{MTCP}_4/\mathrm{PH}_2/1$. In general, our computations are for MAP/PH/1 queues where the service time distributions are parametrized by their workloads and their Squared Coefficient of Variations (SCVs) which we denote by c^2. We have $c^2 = \frac{1}{2}$ in the case of Erlang-2 (E_2) distribution: the sum of two i.i.d. exponential random variables with rate $\frac{2\Lambda}{\rho}$, where Λ is the arrival rate as in (3) and ρ is the workload. In the case of $c^2 = 1$, we use exponentially distributed random variables with rate $\mu = \frac{\Lambda}{\rho}$. For the case of $c^2 > 1$, we use the Hyperexponential-2 (H_2) distribution which is a mixture of two independent exponential random variables. With probability $p = \frac{1}{2c^2-1}$ we take an exponential distribution with rate $\frac{\Lambda}{c^2\rho}$ and with probability $1 - p$ we take an exponential distribution with rate $\frac{2\Lambda}{\rho}$. It is easy to verify that this H_2 random variable has mean 1 and the desired c^2.

We compute the matrix R and the stationary distribution of $\mathrm{MMPP}_2/\mathrm{PH}_2/1$ and $\mathrm{MTCP}_4/\mathrm{PH}_2/1$ as QBDs by using the SMC solver. The numerical computation for finding the relative errors, $\frac{\text{true value-approximate value}}{\text{true value}}$, shows the same properties for the curves of the relative error of mean and SCV of steady state queue for all of the above mentioned processes.

Figure 4 (left) shows different relative errors of the steady state mean for various service time SCVs. The bigger the SCV of service time, the less relative error of the mean. Figure 4 (right) shows different relative errors of the steady state SCV for various service time SCVs. The minimum absolute value of the relative error is again for the case that the service distribution is hyperexponential, i.e. the bigger the SCV of service time, the less absolute value of the relative error of SCV of steady state queue.

Both of these families of curves are bell-shaped. The only difference is that in contrast to the relative error of means which has positive values, the relative error of SCV of the steady state queue has negative values. This shows that the true value for mean is always greater than the approximate one and the opposite holds for SCV.

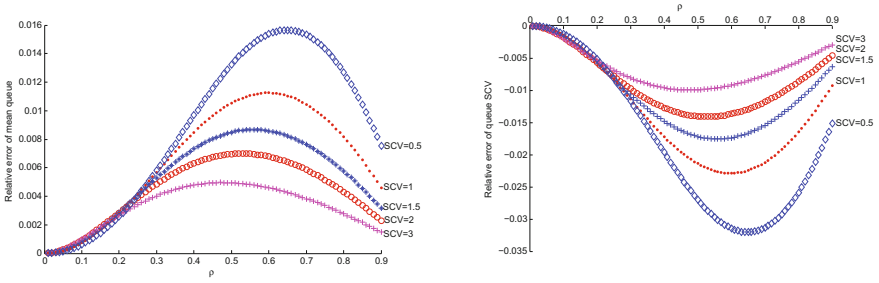


Fig. 4 The relative mean error (left) and relative SCV error (right) of a steady state queue. The MMPP_2 model used has $\tilde{q}_{12} = \tilde{q}_{21} = 5$, $\lambda_1 = 10$, $\lambda_2 = 20$. Then the mean service time is varied to accommodate for the desired ρ.

From further investigation of the variance (not appearing in the figures) it also holds that the true variance is less than or equal to the approximated variance.

As is evident from the figures, in any case, the relative error is negligible. Note though, that for more bursty arrival processes we may have bigger relative errors than those in the figure, yet we carried out an extensive computational study to find an empirical boundary for relative error. Assuming that λ_1 is constant (=10) and varying the values of λ_2, $\tilde{q}_{12}$ and $\tilde{q}_{21}$ gives the results in Table 1 for the maximum relative error. These empirical results indeed suggest that the MTCP/PH/1 is a very sensible alternative model to MMPP/PH/1.

Table 1 Maximum relative error of mean queue in approximation of $\text{MMPP}_2/\text{PH}_2/1$ queue by $\text{MTCP}_4/\text{PH}_2/1$ queue where $\lambda_1 = 10$. Note that the H_2 case corresponds to $c^2 = 1.1$.

Model	λ_2	$\tilde{q}_{12}$	$\tilde{q}_{21}$	Max Relative Error of Mean Queue
$\text{MMPP}_2/\text{E}_2/1$	500	8	70	0.0893
$\text{MMPP}_2/\text{M}/1$	300	9	70	0.0725
$\text{MMPP}_2/\text{H}_2/1$	400	5	70	0.0715

6 Conclusions and Future Work

As illustrated in this paper, MMPPs can perhaps be replaced by MTCPs for modelling purposes. We have shown a theoretical relationship between the two processes and an empirical relationship between their associated queueing models. Our focus in this conference paper is on being expository, hence we focused on the case of $p = 2$. A question that arises is: "Can we construct an MTCP to match a non-slow MMPP with the same mean and variance?".

In further work we plan to handle the general case, for $p > 2$, where we believe similar results may hold. Proving the empirical bounds that we found for the queueing approximations remains a challenge.

Of further interest is the issue of parameter estimation of MTCPs. Our belief is that since data traces generated by MTCPs are more informative than those generated by MMPPs, there is a promise in devising a good parameter estimation method for MTCPs.

Acknowledgments We thank Sophie Hautphenne for useful discussions and advice. Yoni Nazarathy is supported by Australian Research Council (ARC) grants DP130100156 and DE130100291.

Azam Asanjarani is supported by DE130100291.

References

1. Asmussen, S.: Phase-type representations in random walk and queueing problems. Ann. of Prob., 772–789 (1992)
2. Asmussen, S.: Matrix-analytic models and their analysis. Scand. Jour. of Stat. **27**(2), 193–226 (2000)
3. Asmussen, S.: Applied probability and queues. Stochastic Modelling and Applied Probability, vol. 51. Springer (2003)
4. Bini, D.A., Meini, B., Steffé, S., Van Houdt, B.: Structured Markov chains solver: software tools. In: Proc. 2006 Workshop on Tools for Solving Structured Markov Chains, Article No. 14. ACM (2006)
5. Fischer, W., Meier-Hellstern, K.: The Markov-modulated Poisson process (MMPP) cookbook. Perf. Eval. **18**(2), 149–171 (1993)
6. Fomundam, S., Herrmann, J.W.: A survey of queuing theory applications in healthcare (2007)
7. Gun, L.: An Algorithmic Analysis of the MMPP/G/1 Queue (No. ISR-TR-88-40). Maryland Univ. College Park Inst. For Systems Research (1988)
8. Horváth, A., Telek, M.: Markovian modeling of real data traffic: heuristic phase type and MAP fitting of heavy tailed and fractal like samples. In: Calzarossa, M.C., Tucci, S. (eds.) Performance 2002. LNCS, vol. 2459, pp. 405–434. Springer, Heidelberg (2002)
9. Latouche G., Ramaswami, V.: Introduction to matrix analytic methods in stochastic modeling, vol. 5. SIAM (1999)
10. Lucantoni, D.M.: The BMAP/G/1 queue: a tutorial. In: Donatiello, L., Nelson, R. (eds.) SIGMETRICS 1993 and Performance 1993. LNCS, vol. 729, pp. 330–358. Springer, Heidelberg (1993)
11. Nazarathy, Y., Weiss, G.: The asymptotic variance rate of the output process of finite capacity birth-death queues. Queueing Sys. **59**(2), 135–156 (2008)
12. Rudemo, M.: Point processes generated by transitions of Markov chains. Adv. Appl. Prob., 262–286 (1973)
13. Ramesh, N.I.: Statistical analysis on Markov-modulated Poisson processes. Environmetrics **6**(2), 165–179 (1995)
14. Riska, A., Squillante, M., Yu, S.Z., Liu, Z., Zhang, L.: Matrix-analytic analysis of a MAP/PH/1 queue fitted to web server data. Matrix-Analytic Methods; Theory and Applications, 333–356 (2002)

Part II
Queueing Applications

Throughput Analysis for the Opportunistic Channel Access Mechanism in CRNs with Imperfect Sensing Results

Shiying Ge, Shunfu Jin and Wuyi Yue

Abstract In order to reduce the average delay of secondary user (SU) packets and adapt to various tolerance for transmission interruption, in this paper, we propose a novel opportunistic channel access mechanism in cognitive radio networks (CRNs). Considering the preemptive priority of primary user (PU) packets, as well as the sensing errors caused by SUs, we can model the network system as a priority queue with two classes of packets, however, these two classes of packets may interfere with each other. We first analyze the stationary distribution of the queueing model, then we derive the expression for the throughput of SU packets. Finally, we provide numerical experiments to optimize the energy sensing threshold with a maximum throughput of SU packets.

Keywords Cognitive radio networks · Opportunistic channel access · Sensing threshold · Throughput

1 Introduction

In future wireless application, such as the 5th generation (5G) networks, the demand for wireless spectrum resources will have a huge increase [1]. Now, a large portion of the assigned spectrum remains under-utilized. This is the key reason leading to

S. Ge · S. Jin (✉)
School of Information Science and Engineering, Yanshan University,
Qinhuangdao 066004, China
e-mail: gesyemail@163.com, jsf@ysu.edu.cn

S. Ge · S. Jin
Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province,
Qinhuangdao 066004, China

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

T.V. Do et al. (eds.), *Queueing Theory and Network Applications*,
Advances in Intelligent Systems and Computing 383,
DOI: 10.1007/978-3-319-22267-7_5

the shortage of spectrum resources [2]. How to improve the utilization of spectrum resources is a hot topic of research. Cognitive radio (CR) is predicted to be one of the most popular wireless technologies due to its efficient spectrum utilization. In such a situation, cognitive radio networks (CRNs) emerge as required [3]. Recently, the opportunistic channel access mechanism in CRNs has been paid more attention aiming to improve the spectrum effectively.

In [4], a strategy for improving the priority of the interrupted SU was proposed to reduce the delay of interrupted SU packets. However, the delay of all the SU packets was not analyzed mathematically. In [5], a dynamic channel selection approach was proposed to reduce the overhead caused by interrupted SU packets. According to channel idle probability and average waiting time in the channel queue, a value was calculated for each channel to estimate how much a channel was suitable for being selected when the interruption occurred. However, the perfect sensing results of SUs was assumed.

In practice, there exist two kinds of errors associated with channel sensing, namely, missed detection and false alarm [6]. In [7], a cooperative channel sensing strategy was analyzed. The performance of the network in terms of maximum throughput with optimal number of SU packets and sensing time was investigated. Unfortunately, the preemptive priority of PU packet to SU packet was neglected.

Inspired by the above observations, aiming to reduce the average delay of all the SU packets and adapt to various tolerance for transmission interruption, in this paper, we propose a new opportunistic channel access mechanism. Considering the preemptive priority of primary user PU packets and the sensing errors caused by SUs, we establish a priority queueing model with two classes of packets to capture the stochastic behavior of the network system. The two classes of packets may interfere with each other. By using the method of Markov chain, we analyze the model to investigate the influence of admission threshold and feedback probability on the system performance.

2 Opportunistic Channel Access Mechanism and System Model

In this section, we first present the opportunistic channel access mechanism that we proposed in CRNs.

We consider a CRN with a single licensed channel. The channel is used by PU packets preemptively and shared by SU packets opportunistically. To minimum the delay of PU packets, no buffer is prepared for PU packets. That is to say, the number of PU packets in the system is no more than 1. To improve the throughput of SU packets, we set a buffer for SU packets. Moreover, we set an admission threshold H ($H > 0$) and an admission probability r ($0 \leq r \leq 1$).

When an SU packet arrives at the system, the central controller will compare the number L of SU packets in the system with the admission threshold H. If $L \geq H$, the arriving SU packet will be admitted to join the system with probability r and

refused by the system with probability $\bar{r}$. Otherwise, the arriving SU packet will be admitted with probability 1. After admitted, the arriving SU packet will queue at the buffer.

Under the schedule of the central controller, the SU will sense the channel periodically. According to the sensing results, the SU packet will decide wether to occupy the channel for transmission or not. Moreover, if an SU packet occupying the channel is interrupted, it will leave the system with probability q $(0 \leq q \leq 1)$ and return the buffer with probability $\bar{q} = 1 - q$. We define E as the value of detection energy and τ as the energy threshold.

With the stochastic behavior of the network, we consider the system model in discrete-time field. We suppose that the arrival processes for both PU packets and SU packets follow Bernoulli distributions with arrival rates λ_{pu} and λ_{su}, respectively. In addition, we suppose that the service times of a PU packet and an SU packet follow geometrical distributions with parameters μ_{pu} and μ_{su}, respectively.

We define the total number $X_n = i$ $(i = 0, 1, 2, \ldots)$ of SU packets in the system as the system level and the channel state $Y_n = j$ $(j = 0, 1, 2, 3)$ as the system stage at the instant n^+. Then the process $\{(X_n, Y_n),\ n \geq 1\}$ constitutes a two-dimensional Markov process. Let $\pi_{i,j}$ be the steady-state probability that the system level is i and the system stage is j. Therefore, $\pi_{i,j}$ can be given by

$$\pi_{i,j} = \lim_{n \to \infty} P\{X_n = i, Y_n = j\}. \tag{1}$$

3 Stationary Probability Distribution

3.1 Missed Detection and False Alarm

When an SU senses the channel via energy detection, two kinds of sensing errors in terms of missed detection and false alarm are unavoidable.

Let p_{md} be the missed detection ratio and p_{fa} be the false alarm ratio. Referencing [6], p_{md} and p_{fa} can be given as follows:

$$\begin{cases} p_{md} = 1 - Q\left(\left(\dfrac{\tau}{\sigma^2} - \gamma - 1\right)\sqrt{\dfrac{t_s f_s}{2\gamma + 1}}\right) \\ p_{fa} = Q\left(\left(\dfrac{\tau}{\sigma^2} - 1\right)\sqrt{t_s f_s}\right) \end{cases} \tag{2}$$

where τ is the energy threshold defined in Section 2, t_s is the sensing time, f_s is the sensing frequency, γ is the signal-to-noise ratio, σ is the variance of noise and $Q(v)$ is the tail probability of the standard normal distribution as follows:

$$Q(v) = \frac{1}{\sqrt{2\pi}} \int_v^{\infty} \exp\left(-\frac{t^2}{2}\right) dt.$$

3.2 Transition Probability Matrix

We define $\boldsymbol{P}$ as the one step transition probability matrix of the Markov process (X_n, Y_n), $n \geq 1$. Let $\boldsymbol{P}_{ik}$ be the transition probability sub-matrices for the number of SU packets in the system changing from i ($i = 0, 1, 2, \ldots$) to k ($k = 0, 1, 2, \ldots$). $\boldsymbol{P}_{ik}$ is discussed as follows.
(1) If $i = 0$ and $k = 0$, it means that there is no SU packet arrival at the system during the one step transition from the system level 0. $\boldsymbol{P}_{00}$ can be given by

$$\boldsymbol{P}_{00} = \bar{\lambda}_{su} \begin{bmatrix} \bar{\lambda}_{pu} & \lambda_{pu} & 0 & 0 \\ \bar{\lambda}_{pu}\mu_{pu} & V_{pu} & 0 & 0 \end{bmatrix}$$

where $V_{pu} = \bar{\mu}_{pu} + \lambda_{pu}\mu_{pu}$.
(2) If $i = 0$ and $k = 1$, it means that there is an SU packet arrival at the system during the one step transition from the system level 0. $\boldsymbol{P}_{01}$ can be given by

$$\boldsymbol{P}_{01} = \lambda_{su} \begin{bmatrix} \bar{\lambda}_{pu} & \lambda_{pu} & \bar{\lambda}_{pu} & \lambda_{pu} \\ \bar{\lambda}_{pu}\mu_{pu} & V_{pu} & \bar{\lambda}_{pu}\mu_{pu} & V_{pu} \end{bmatrix} \times \boldsymbol{K}$$

where

$$\boldsymbol{K} = \begin{bmatrix} p_{fa} & 0 & 0 & 0 \\ 0 & \bar{p}_{md} & 0 & 0 \\ 0 & 0 & \bar{p}_{fa} & 0 \\ 0 & 0 & 0 & p_{md} \end{bmatrix}.$$

(3) If $i = 1$ and $k = 0$, it means that there is an SU packet departure and no arrival at the system during the one step transition from the system level 1. $\boldsymbol{P}_{10}$ can be given as follows:

$$\boldsymbol{P}_{10} = \bar{\lambda}_{su} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \mu_{su}\bar{\lambda}_{pu} & \mu_{su}\lambda_{pu} & 0 & 0 \\ \bar{\lambda}_{pu} & \lambda_{pu} & 0 & 0 \end{bmatrix} + \bar{\lambda}_{su} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \bar{q}\bar{\mu}_{su}\bar{\lambda}_{pu} & \bar{q}\bar{\mu}_{su}\lambda_{pu} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \times \boldsymbol{K}.$$

(4) Let $\boldsymbol{B}_0$ be the transition probability matrix representing that there is an SU packet arrival at the system and no SU packet departure during the one step transition from system levels higher than 0. $\boldsymbol{B}_0$ is given as follows:

$$\boldsymbol{B}_0 = \lambda_{su} \begin{bmatrix} \bar{\lambda}_{pu} & \lambda_{pu} & \bar{\lambda}_{pu} & \lambda_{pu} \\ \bar{\lambda}_{pu}\mu_{pu} & V_{pu} & \bar{\lambda}_{pu}\mu_{pu} & V_{pu} \\ q\bar{\mu}_{su}\bar{\lambda}_{pu} & q\bar{\mu}_{su}\lambda_{pu} & \bar{\mu}_{su}\bar{\lambda}_{pu} & \bar{\mu}_{su}\lambda_{pu} \\ 0 & 0 & 0 & 0 \end{bmatrix} \times \boldsymbol{K}.$$

(i) For the case of $1 \leq i < H$ and $k = i + 1$, $\boldsymbol{P}_{i,i+1}$ is given by

$$\boldsymbol{P}_{i,i+1} = \boldsymbol{B}_0.$$

(ii) For the case of $i \geq H$ and $k = i + 1$, $\boldsymbol{P}_{i,i+1}$ is denoted as $\boldsymbol{A}_0$. $\boldsymbol{A}_0$ is given by

$$\boldsymbol{A}_0 = r\boldsymbol{B}_0.$$

(5) Let $\boldsymbol{B}_1$ be the transition probability matrix representing the number of SU packets to be fixed at a value lager than 0 during one step transition. $\boldsymbol{B}_1$ is given as follows:

$$\boldsymbol{B}_1 = \begin{bmatrix} \bar{\lambda}_{su} & 0 & 0 & 0 \\ 0 & \bar{\lambda}_{su} & 0 & 0 \\ 0 & 0 & \lambda_{su}(\mu_{su} + \bar{q}\bar{\mu}_{su}) + \bar{\lambda}_{su}q\bar{\mu}_{su} & 0 \\ 0 & 0 & 0 & \lambda_{su} \end{bmatrix} \times \begin{bmatrix} \bar{\lambda}_{pu} & \lambda_{pu} & \bar{\lambda}_{pu} & \lambda_{pu} \\ \bar{\lambda}_{pu}\mu_{pu} & V_{pu} & \bar{\lambda}_{pu}\mu_{pu} & V_{pu} \\ \bar{\lambda}_{pu} & \lambda_{pu} & \bar{\lambda}_{pu} & \lambda_{pu} \\ \bar{\lambda}_{pu} & \lambda_{pu} & \bar{\lambda}_{pu} & \lambda_{pu} \end{bmatrix} \times \boldsymbol{K}.$$

(i) For the case of $1 \leq i < H$ and $k = i$, $\boldsymbol{P}_{ii}$ is given by

$$\boldsymbol{P}_{ii} = \boldsymbol{B}_1.$$

(ii) For the case of $i = H$ and $k = i$, $\boldsymbol{P}_{HH}$ is given by

$$\boldsymbol{P}_{HH} = \boldsymbol{B}_1 + \bar{r}\lambda_{su} \begin{bmatrix} \bar{\lambda}_{pu} & \lambda_{pu} & \bar{\lambda}_{pu} & \lambda_{pu} \\ \bar{\lambda}_{pu} & \lambda_{pu} & \bar{\lambda}_{pu} & \lambda_{pu} \\ (2q-1)\bar{\mu}_{su}\bar{\lambda}_{pu} & (2q-1)\bar{\mu}_{su}\lambda_{pu} & \bar{\mu}_{su}\bar{\lambda}_{pu} & \bar{\mu}_{su}\lambda_{pu} \\ 0 & 0 & 0 & 0 \end{bmatrix} \times \boldsymbol{K}.$$

(iii) For the case of $i > H$ and $k = i$, $\boldsymbol{P}_{ii}$ is denoted as $\boldsymbol{A}_1$. $\boldsymbol{A}_1$ is given by

$$\boldsymbol{A}_1 = \boldsymbol{B}_1 + \bar{r}\lambda_{su} \times \begin{bmatrix} \bar{\lambda}_{pu} & \lambda_{pu} & \bar{\lambda}_{pu} & \lambda_{pu} \\ \bar{\lambda}_{pu} & \lambda_{pu} & \bar{\lambda}_{pu} & \lambda_{pu} \\ (2q\bar{\mu}_{su} - 1)\bar{\lambda}_{pu} & (2q\bar{\mu}_{su} - 1)\lambda_{pu} & (2\bar{\mu}_{su} - 1)\bar{\lambda}_{pu} & (2\bar{\mu}_{su} - 1)\lambda_{pu} \\ -\bar{\lambda}_{pu} & -\lambda_{pu} & -\bar{\lambda}_{pu} & -\lambda_{pu} \end{bmatrix} \times \boldsymbol{K}.$$

(6) Let $\boldsymbol{B}_2$ be the transition probability matrix representing that there is no SU packet arrival at the system and an SU packet departure during the one step transition from system levels higher than 0. $\boldsymbol{B}_2$ is given as follows:

$$\boldsymbol{B}_2 = \bar{\lambda}_{su} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ (\mu_{su} + \bar{q}\bar{\mu}_{su})\bar{\lambda}_{pu} & (\mu_{su} + \bar{q}\bar{\mu}_{su})\lambda_{pu} & \mu_{su}\bar{\lambda}_{pu} & \mu_{su}\lambda_{pu} \\ \bar{\lambda}_{pu} & \lambda_{pu} & \bar{\lambda}_{pu} & \lambda_{pu} \end{bmatrix} \times \boldsymbol{K}.$$

(i) For the case of $1 \leq i < H$ and $k = i - 1$, $\boldsymbol{P}_{i,i-1}$ is given by

$$\boldsymbol{P}_{i,i-1} = \boldsymbol{B}_2.$$

(ii) For the case of $i = H$ and $k = H - 1$, $\boldsymbol{P}_{H,H-1}$ is given by

$$\boldsymbol{P}_{H,H-1} = \boldsymbol{B}_2 + \bar{q}\bar{\mu}_{su}\bar{r}\lambda_{su} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \bar{\lambda}_{pu} & \lambda_{pu} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \times \boldsymbol{K}.$$

(iii) For the case of $i > H$ and $k = i - 1$, $\boldsymbol{P}_{i,i-1}$ is denoted as $\boldsymbol{A}_2$. $\boldsymbol{A}_2$ is given by

$$\boldsymbol{A}_2 = \boldsymbol{B}_2 + \bar{r}\lambda_{su} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ (\bar{q}\bar{\mu}_{su} + \mu_{su})\bar{\lambda}_{pu} & (\bar{q}\bar{\mu}_{su} + \mu_{su})\lambda_{pu} & \mu_{su}\bar{\lambda}_{pu} & \mu_{su}\lambda_{pu} \\ \bar{\lambda}_{pu} & \lambda_{pu} & \bar{\lambda}_{pu} & \lambda_{pu} \end{bmatrix} \times \boldsymbol{K}.$$

From above, we obtained all the nonzero sub-matrices in the transition probability matrix $\boldsymbol{P}$. Then $\boldsymbol{P}$ can be given as follows:

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{P}_{00} & \boldsymbol{P}_{01} & & & & & & & \\ \boldsymbol{P}_{10} & \boldsymbol{P}_{11} & \boldsymbol{P}_{12} & & & & & & \\ & \boldsymbol{P}_{21} & \boldsymbol{P}_{22} & \boldsymbol{P}_{23} & & & & & \\ & & \ddots & \ddots & \ddots & & & & \\ & & & \boldsymbol{P}_{H-1,H-2} & \boldsymbol{P}_{H-1,H-1} & \boldsymbol{P}_{H-1,H} & & & \\ & & & & \boldsymbol{P}_{H,H-1} & \boldsymbol{P}_{HH} & \boldsymbol{A}_0 & & \\ & & & & & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & \\ & & & & & & \ddots & \ddots & \ddots \end{bmatrix}. \tag{3}$$

With $\boldsymbol{P}$, we can get the steady-state probability distribution π_{ij} defined in Eq. (1) by employing the matrix-geometric solution method.

4 Throughput of SU Packets and Numerical Experiments

In order to evaluate the system performance of the mechanism proposed in Section 2, we present the performance measure for the throughput of SU packets. The throughput ϕ of SU packets is defined as the number of SU packets transmitted successfully per slot. ϕ is given as follows:

$$\phi = \sum_{i=0}^{\infty} \pi_{i,2}\mu_{su}. \tag{4}$$

By using the same settings for the system parameters as in [6], we set $1 \leq \tau \leq 7$, $\lambda_{pu} = 0.05$, $\mu_{su} = 0.8$, $\mu_{su} = 0.7$ as an example, and provide numerical experiments to show the change trend of the throughput packets. We conclude that the analysis results match well with the simulation results.

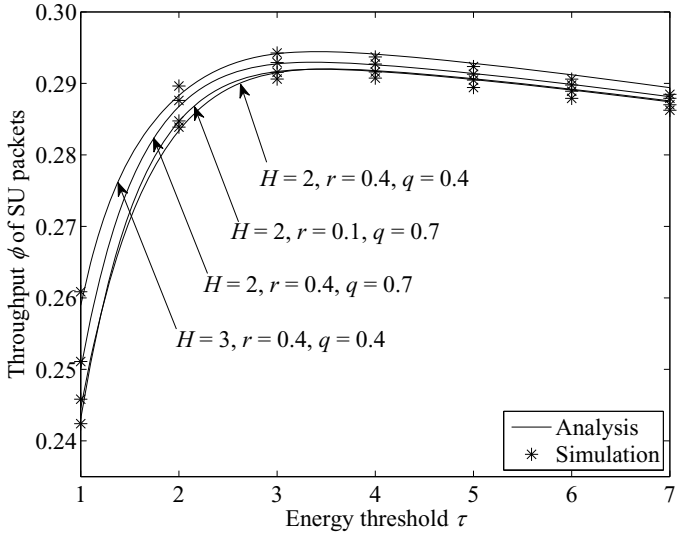


Fig. 1 Throughput ϕ of SU packets versus the energy threshold τ

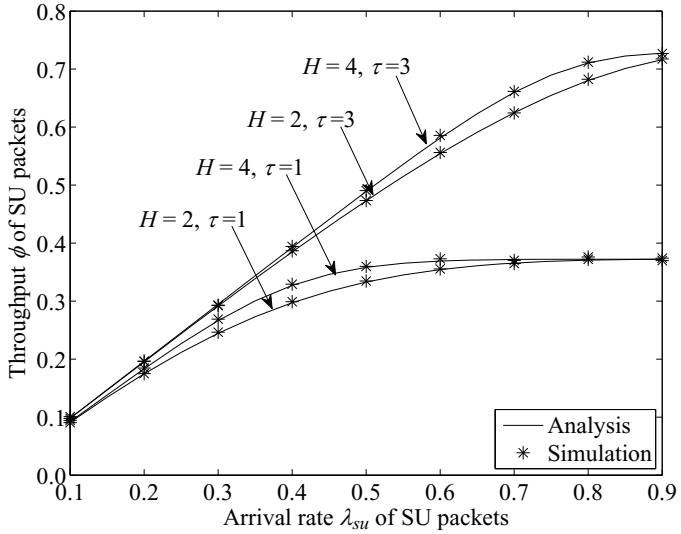


Fig. 2 Throughput ϕ of SU packets versus the arrival rate λ_{su} of SU packets

Figure 1 shows how the throughput ϕ of SU packets changes with the energy threshold τ for the arrival rate $\lambda_{su} = 0.3$ of SU packets.

In Fig. 1, we find that the throughput ϕ of SU packets will increase firstly and then decreases slowly as the energy threshold τ increases. When the energy threshold is smaller, the false alarm ratio is greater, so the false alarm is the main reason to impact

the throughput of the SU packets. As the energy threshold increases, the false alarm radio becomes smaller, the probability that an SU packet leaves the system because of the false alarm is less, so the throughput will increase. On the other hand, with the energy threshold increase, the missed detection becomes greater and it will lead more SU packets collide with PU packets. So, the throughput of SU packets will decrease.

Figure 2 demonstrates how the throughput ϕ of SU packets changes with the arrival rate λ_{su} of SU packets for admission probability $r = 0.4$ and feedback probability $q = 0.4$.

In Fig. 2, we notice that the throughput ϕ of SU packets will increase with the arrival rate λ_{su} of SU packets. The larger the arrival rate of SU packets is, the more the probability is that an SU packet arrives at the system. So, the throughput of SU packets will increase.

5 Conclusion

In this paper, we proposed a novel opportunistic channel access mechanism with admission threshold and probabilistic feedback in CRNs to reduce the average delay of SU packets. Based on the imperfect sensing results of SUs, we established a priority queueing model with two classes of packets to capture the stochastic behavior. We analyzed system model in steady state by using matrix-geometric solution method. Finally, with numerical experiments, we evaluated the throughput of SU packets.

Acknowledgments This work was supported in part by National Natural Science Foundation (No. 61472342), China and was supported in part by MEXT, Japan.

References

1. Hu, R.Q., Qian, Y.: An Energy Efficient and Spectrum Efficient Wireless Heterogeneous Network Framework for 5G Systems. IEEE Communications Magazine **52**, 94–101 (2014)
2. Ghosh, G., Das, P., Chatterjee, S.: Cognitive Radio and Dynamic Spectrum Access-A Study. International Journal of Next-Generation Networks **6**, 43–60 (2014)
3. Altrad, O., Muhaidat, S., Al-Dweik, A., Shami, A., Yoo, P.D.: Opportunistic Spectrum Access in Cognitive Radio Networks under Imperfect Spectrum Sensing. IEEE Transactions on Vehicular Technology **63**, 920–925 (2014)
4. Zhang, Y., Jiang, T., Zhang, L., Wei, P.: Analysis on the transmission delay of priority-based secondary users in cognitive radio networks. In: Proc. 2013 International Conference on Wireless Communications and Signal Processing, CD-ROM, 6 pages. IEEE Press, Hangzhou (2013)
5. Kahvand, M., Soleimani, M.T., Dabiranzohouri, M.: Channel selection in cognitive radio networks: a new dynamic approach. In: Proc. 11th IEEE Malaysia International Conference on Communications, pp. 407–411. IEEE Press, Kuala Lumpur (2013)
6. Liang, Y.C., Zeng, Y., Peh, E.C.Y., Hoang, A.T.: Sensing-Throughput Tradeoff for Cognitive Radio Networks. IEEE Transactions on Wireless Communications **7**, 1326–1337 (2008)
7. Bhowmick, A., Das, M.K., Biswas, J., Roy, S.D., Kundu, S.: Throughput optimization with cooperative spectrum sensing in cognitive radio network. In: Proc. Souvenir of the 2014 IEEE International Advance Computing Conference, pp. 329–332. IEEE Press, Gurgaon (2014)

Throughput Analysis of Multichannel Cognitive Radio Networks Based on Stochastic Geometry

Seunghee Lee and Ganguk Hwang

Abstract In this paper, we consider an underlay type cognitive radio network with multiple secondary users who contend to access multiple heterogeneous primary channels. With the help of stochastic geometry we develop a new analytical model to analyze the throughput of a random channel access protocol where each secondary user determines whether to access a primary channel based on a given access probability. Due to the interference-free region that we newly introduce we can easily analyze the throughput of a random channel access protocol. Numerical examples are provided to validate our analysis.

1 Introduction

There are rapidly increasing new wireless services that need high transmission rate and network throughput. Such an explosion of new wireless services causes the scarcity of the radio spectrum. In order to solve the scarcity of the radio spectrum, a concept of cognitive radio (CR) has been introduced. A CR network consists of licensed primary users (PUs) who have a priority to occupy their designated radio spectrum and unlicensed secondary users (SUs) who are allowed to access the radio spectrum as far as they do not interfere the PUs at all (an overlay type) or affect the PUs very limitedly (an underlay type) [1].

In order to capture the random feature of user locations in the network, stochastic geometry [2] is widely used. Interference depends upon the path loss and the fading characteristics of wireless interface. Both of them can be interpreted as functions of the distance between users in the network, where stochastic geometry is involved. For a comprehensive understanding, we refer the readers to a survey paper [3] and the references therein.

S. Lee · G. Hwang(✉)
Department of Mathematical Sciences, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
e-mail: oz0386@kaist.ac.kr, guhwang@kaist.edu

T.V. Do et al. (eds.), *Queueing Theory and Network Applications*,
Advances in Intelligent Systems and Computing 383,
DOI: 10.1007/978-3-319-22267-7_6

Pioneering studies on CR networks with stochastic geometry assume that the spectrum access of each SU does not depend on the spatial distribution of PUs in the network [4]-[6]. This assumption fails to characterize the networks in practice because there is a strong correlation between the spatial distribution of PUs and the spectrum access of SUs. The modeling of the dependency brings us many difficulties in performance analysis. One of key difficulties in analysis is that the distribution of the interference created by active SUs cannot be obtained in an analytical form. Here, active SUs denotes the SUs who transmit their packets. To overcome such difficulties various models and approximations have been proposed, e.g., [7, 8, 9, 10].

In this paper, we consider a random access protocol in an underlay CR network with multiple heterogeneous primary channels and multiple SUs. Each SU is able to transmit its packet in channel k if it selects channel k, decides to access channel k with a given access probability, and finally decides that channel k is idle after sensing. We assume that the access probability depends on the channel that the SU selects, which is important because we consider heterogeneous primary channels. We aim to analyze the throughput performance of an arbitrary SU. To overcome the difficulties mentioned above, we propose to consider the so-called *interference-free region* of a channel for an active SU. With the help of the interference-free region we can approximate the throughput in a simple way. More details will be given in Section 3. Numerical and simulation results are provided to validate our analysis.

The organization of this paper is as follows. In Section 2, we describe the system model. In Section 3, we analyze the active probability, the coverage probability (COP), and throughput. To validate our analysis numerical and simulation results are provided in Section 4. We give our conclusions in Section 5.

2 System Model

We consider an underlay type CR network with N primary channels. The PUs for each primary channel in the network are assumed to follow a marked Poisson Point Process (PPP). The SUs in the network are assumed to follow another marked PPP. We assume that the intensities of PUs of channels are different, so that SUs have heterogeneous primary channels. The time axis is divided into slots of equal size and one slot time is needed to transmit a packet for all users.

2.1 Network Model

Let $\tilde{\Phi}_{P,k} = \{(X_{i,k}, a_{i,k}, \mathbf{F}_{i,k})\}$ be a marked PPP with intensity $\lambda_{p,k}$ on $\mathbb{R}^2$ which represents the set of PUs of channel k, and $\tilde{\Phi}_S = \{(Y_i, b_i, c_i, \mathbf{F}_i)\}$ be a marked PPP with intensity λ_s on $\mathbb{R}^2$ which represents the set of SUs. Here

(1) $\Phi_{P,k} = \{X_{i,k}\}$ denotes the locations of PUs of channel k, and $\Phi_S = \{Y_i\}$ denotes the locations of SUs. All PUs and SUs are potential transmitters. The receivers are assumed to be located at a distance R from their corresponding transmitters in a random direction as in the so-called bipolar model.

(2) $\{A_{i,k}\}$ are independent and identically distributed (i.i.d.) random variables (r.v.s). For PU i, the r.v. $A_{i,k}$ denotes the activity of PU i in channel k: $A_{i,k} = 1$ with probability $\pi_{k,0}$ (i.e., PU i is active), and $A_{i,k} = 0$ with probability $\pi_{k,1}$ (i.e., PU i is inactive).
$\{d_i\}$ are i.i.d. r.v.s. For SU i, the r.v. d_i denotes the accessibility of SU i: $d_i = 1$ with probability a_k (i.e., SU i can access channel k), and $d_i = 0$ with probability $1 - a_k$ (i.e., SU i does not access any channel) when SU i selects channel k. The probabilities a_k, $1 \leq k \leq N$, are called the common access probabilities for SUs.
(3) c_i is the r.v. which represents the channel that SU i selects. The probability that any SU selects channel k is denoted by s_k with $\sum_{k=1}^{N} s_k = 1$.
(4) Each SU is allowed to transmit its packet in channel k if it selects channel k with probability s_k, decides to access channel k with access probability a_k, and determines channel k to be idle after sensing channel k. In this case, the SU is called an active SU.
(5) $\mathbf{F}_{i,k} = \{F_k(i, j)\}$ denotes a collection of the fading power $F_k(i, j)$ from transmitter i to receiver j at the transmitting time in channel k. If the transmitter i is an active PU in channel k, we assume that $F_k(i, j)$ is an exponential r.v. with mean $1/\mu_p$. If the transmitter i is an active SU in channel k, we assume that $F_k(i, j)$ is an exponential r.v. with mean $1/\mu_s$. We further assume that all fading powers are independent.

2.2 Sensing Model

We tag an arbitrary SU, say SU i, and call it the tagged SU. Let the location of the tagged SU be denoted by Y_i. When the tagged SU selects a channel for a packet transmission, it first senses the selected channel and determines the channel to be idle if the aggregated received signal from all active PUs of the selected channel is smaller than *a priori* given sensing threshold T_1.

To be more specific, suppose that the tagged SU selects and senses channel k. Let $\Phi^a_{P,k}$ be the set of all active PUs of channel k, that is,

$$\Phi^a_{P,k} = \{X_{j,k} \in \Phi_{P,k} : A_{j,k} = 1\}.$$

Then the received signal at the tagged SU is expressed as

$$I_{\Phi^a_{P,k}}(Y_i) = \sum_{X_{j,k} \in \Phi^a_{P,k}} F_k(j, i)/|X_{j,k} - Y_i|^{\alpha}$$

where α is the path loss exponent. This refers to the aggregated signal at the tagged SU from active PUs of channel k in $\Phi^a_{P,k}$. Consequently, channel k is identified as idle if the following condition is satisfied: $I_{\Phi^a_{P,k}}(Y_i) < T_1$. For simplicity, the noise-free model is assumed in this paper.

SU j located at Y_j can transmit its packet in channel k if it selects channel k, decides to access channel k, and determines that channel k is idle. These conditions can be encoded in the following indicator:

$$e_{j,k} = \mathbb{1}(c_j = k, d_j = 1, I_{\Phi^a_{P,k}}(Y_j) < T_1).$$

2.3 Transmission Model

Consider the tagged SU who selects channel k. Let the location of the corresponding receiver of the tagged SU, called the tagged SU receiver, be denoted by y_i. The tagged SU receiver can successfully receive the packet from the tagged SU if the Signal to Interference Ratio (SIR) of the signal from Y_i to y_i is larger than *a priori* given success threshold T_2. Here, the interference is the signal transmitted from the other transmitters including active PUs and SUs in the network. We denote the aggregate interference from active PUs at the tagged SU receiver y_i by $I_{\Phi^a_{P,k}}(y_i)$. $I_{\Phi^a_{P,k}}(y_i)$ is given by $I_{\Phi^a_{P,k}}(y_i) = \sum_{X_{j,k} \in \Phi^a_{P,k}} F_k(j,i)/|X_{j,k} - y_i|^\alpha$. Let $\Phi_{S,k}$ be the set of active SUs who transmit their packets in channel k, that is, $\Phi_{S,k} = \{Y_j \in \Phi_S : e_{j,k} = 1\}$. Then $I_{\Phi_{S,k} \setminus Y_i}(y_i) = \sum_{Y_j \in \Phi_{S,k} \setminus Y_i} F_k(j,i)/|Y_j - y_i|^\alpha$ which is the aggregate interference from active SUs in $\Phi_{S,k} \setminus Y_i$. Therefore, the transmission is successful if the following condition holds:

$$SIR_{i,k} := \frac{F_k(i,i)/R^\alpha}{I_{\Phi^a_{P,k}}(y_i) + I_{\Phi_{S,k} \setminus Y_i}(y_i)} > T_2. \tag{1}$$

3 Performance Analysis

In this section, we analyze the active probability and the coverage probability of the tagged SU in the CR network. We assume that all SUs are saturated, that is, all SUs always have packets to transmit.

3.1 Analysis of the Active Probability

We derive the active probability which is defined by the probability that the tagged SU can transmit its packet.

Proposition 1. *The probability that the tagged SU can transmit its packet in channel k of a CR network is* $\mathbf{P}\{e_{i,k} = 1\} = s_k a_k \mathbf{P}\{I_{\Phi^a_{P,k}}(Y_i) < T_1\}$ *where*

$$\mathbf{P}\{I_{\Phi^a_{P,k}}(Y_i) < T_1\} \approx \int_0^\infty \int_0^\infty \int_0^{\frac{T_1}{x+y}} \frac{2\left(\pi_{k,0}\lambda_{p,k}\pi\right)^2}{\alpha} e^{-\pi_{k,0}\lambda_{p,k}\pi u^{-2/\alpha}} u^{-\frac{2}{\alpha}-1} \times \left\{ u^{-\frac{2}{\alpha}} - \left(\frac{T_1 - yu}{x}\right)^{-\frac{2}{\alpha}} \right\} \mu_p e^{-\mu_p x} \mu_p e^{-\mu_p y} du dx dy.$$

Proof. Since the analytical computation of $\mathbf{P}\{I_{\Phi^a_{P,k}}(Y_i) < T_1\}$ turns out to be very difficult, we approximate it by considering the interference from the nearest and the second nearest active PUs from the tagged SU. Due to space limitation, we do not provide the detailed proof.

3.2 Analysis of the Coverage Probability

In this subsection, we derive the coverage probability (COP) that is defined by the conditional probability for the tagged SU to transmit successfully, given that the sensing result is idle. Noting that it is not easy to compute the COP because $I_{\Phi^a_{P,k}}$ and $I_{\Phi_{S,k}}$ are not independent and their joint distribution is unknown, we approximate the COP by introducing the concept of the interference-free region of the tagged SU. The motivation is explained as follows. First note that we compute the COP under the condition that the tagged SU determines channel k to be idle. With the condition, there are no active PUs in channel k or even if there are some active PUs in channel k, they are all very far away from the tagged SU, so that the tagged SU experiences almost no interference from active PUs in channel k.

With the motivation we define the interference-free region of the tagged SU in channel k, given that the tagged SU determines channel k to be idle, by the region around the tagged SU (for simplicity, we use a circle centered at the tagged SU) such that any other SUs in the region also experiences almost no interference from active PUs in channel k.

To determine the interference-free region, we first focus on $I_{\Phi^a_{P,k}}(Y_i)$, the interference from all active PUs in channel k, and consider a new PPP with a new intensity $\lambda'_{p,k}$ where the interference I'_1 from the nearest PU in the new PPP at the tagged SU satisfies

$$\mathbf{P}\{I_{\Phi^a_{P,k}}(Y_i) < T_1\} = \mathbf{P}\{I'_1 < T_1\} = \int_0^\infty e^{-\lambda'_{p,k}\pi(x/T_1)^{2/\alpha}} f_{F_P}(x)dx.$$

where $f_{F_P}(x)$ is the pdf of the fading power of a PU (an exponential r.v. with mean $1/\mu_p$), from which we compute the new intensity $\lambda'_{p,k}$ with the help of Proposition 1.

Let R_P be the radius of the circle centered at the nearest PU in the new PPP such that the signal of the nearest PU in the new PPP at the boundary of the circle, is equal to the sensing threshold T_1, that is,

$$F_P \cdot R_P^{-\alpha} = T_1 \tag{2}$$

where F_P is the fading power of the nearest PU in the new PPP.

Definition 1. The radius of the interference-free region $R_{if,k}$ is defined by

$$R_{if,k} = \mathbf{E}[D'_1 | D'_1 > \mathbf{E}[R_P]] - \mathbf{E}[R_P]$$

where D'_1 is the distance between the nearest PU in the new PPP and the tagged SU.

By using (2) and the pdf $f_{D_1'}(r_1)$ of D_1' [11] we can determine $R_{if,k}$ as given in the following lemma whose proof is omitted due to space limitation.

Lemma 1. *The expectation of R_P and the conditional expectation of D_1' are given by*

$$\mathbf{E}[R_P] = \left(\frac{1}{\mu_p T_1}\right)^{1/\alpha} \cdot \Gamma\left(\frac{1}{\alpha} + 1\right)$$

and

$$\mathbf{E}[D_1' | D_1' > \mathbf{E}[R_P]] = \frac{\int_{\mathbf{E}[R_P]}^{\infty} r_1 f_{D_1'}(r_1) dr_1}{\int_{\mathbf{E}[R_P]}^{\infty} f_{D_1'}(r_1) dr_1},$$

respectively, where $f_{D_1'}(r_1) = 2\pi\lambda'_{p,k} r_1 e^{-\pi\lambda'_{p,k} r_1^2}$.

Based on Lemma 1, we are ready to approximate the COP.

Proposition 2. *Given that channel k is sensed as idle, the COP that the tagged SU can successfully transmit its packet is given by*

$$\mathbf{P}\{SIR_{i,k} > T_2 | I_{\Phi^a_{P,k}}(Y_i) < T_1\} \approx \tilde{\mathcal{L}}_{I_{\Phi^a_{P,k}}(y_i)}(\mu_s T_2 R^\alpha) \tilde{\mathcal{L}}_{I_{\Phi_{S,k}\setminus Y_i}(y_i)}(\mu_s T_2 R^\alpha)$$

where

$$\tilde{\mathcal{L}}_{I_{\Phi^a_{P,k}}(y_i)}(\mu_s T_2 R^\alpha) \approx \exp\left\{-2\pi\pi_{k,0}\lambda_{p,k} \int_{R_{if,k}}^{\infty} \frac{r}{1 + \mu_p r^\alpha / \mu_s T_2 R^\alpha} dr\right\}$$

and

$$\tilde{\mathcal{L}}_{I_{\Phi_{S,k}\setminus Y_i}(y_i)}(\mu_s T_2 R^\alpha) \approx \exp\left\{-2\pi s_k a_k \lambda_s \int_0^{R_{if,k}} \frac{r}{1 + r^\alpha / T_2 R^\alpha} dr\right\}.$$

Proof. We start with (1).

$$\begin{aligned}
&\mathbf{P}\{SIR_{i,k} > T_2 | I_{\Phi^a_{P,k}}(Y_i) < T_1\} \\
&\quad = \mathbf{P}\{F_k(i,i) > T_2 R^\alpha (I_{\Phi^a_{P,k}}(y_i) + I_{\Phi_{S,k}\setminus Y_i}(y_i)) | I_{\Phi^a_{P,k}}(Y_i) < T_1\} \\
&\quad = \mathbf{E}\left[e^{-\mu_s T_2 R^\alpha (I_{\Phi^a_{P,k}}(y_i) + I_{\Phi_{S,k}\setminus Y_i}(y_i))} | I_{\Phi^a_{P,k}}(Y_i) < T_1\right].
\end{aligned}$$

With the help of the interference-free region, as explained below we can approximate $\mathbf{E}\left[e^{-\mu_s T_2 R^\alpha (I_{\Phi^a_{P,k}}(y_i) + I_{\Phi_{S,k}\setminus Y_i}(y_i))} | I_{\Phi^a_{P,k}}(Y_i) < T_1\right]$ by only considering the interference from active PUs outside the interference-free region and interference

from active SUs inside the interference-free region, which are considered to be independent. Moreover, since the tagged SU and its receiver are assumed to be located closely, the interference-free region centered at the tagged SU can be considered as the circle centered at the tagged SU receiver.

With our approximation assumption and the Laplace functional of the marked PPP, We approximate $\mathbf{E}\left[e^{-\mu_s T_2 R^\alpha I_{\Phi_{S,k}\setminus Y_i}(y_i)} | I_{\Phi^a_{P,k}}(Y_i) < T_1\right]$ by (we omit the details here due to space limitation)

$$\begin{aligned}\tilde{\mathcal{L}}_{I_{\Phi_{S,k}\setminus Y_i}(y_i)}(\mu_s T_2 R^\alpha) &:= \mathbf{E}\left[e^{-\mu_s T_2 R^\alpha \sum_{Y_j\in\Phi_{S,k}\setminus Y_i} F_k(j,i)/|Y_j - y_i|^\alpha} | I_{\Phi^a_{P,k}}(Y_i) < T_1\right] \\ &\approx \exp\left\{-2\pi s_k a_k \lambda_s \int_0^{R_{if,k}} \frac{r}{1 + r^\alpha / T_2 R^\alpha} dr\right\}.\end{aligned}$$

Similarly, we obtain

$$\begin{aligned}\tilde{\mathcal{L}}_{I_{\Phi^a_{P,k}}(y_i)}(\mu_s T_2 R^\alpha) &:= \mathbf{E}\left[e^{-\mu_s T_2 R^\alpha I_{\Phi^a_{P,k}}(y_i)} | I_{\Phi^a_{P,k}}(Y_i) < T_1\right] \\ &\approx \exp\left\{-2\pi \pi_{k,0} \lambda_{p,k} \int_{R_{if,k}}^{\infty} \frac{r}{1 + \mu_p r^\alpha / \mu_s T_2 R^\alpha} dr\right\}.\end{aligned}$$

3.3 *Analysis of Throughput*

We define the throughput T of the tagged SU as the probability that the tagged SU receiver successfully receives a packet at an arbitrary slot time. Combining the active probability and the COP derived in Proposition 1 and Proposition 2, the throughput T of the tagged SU can be approximated by

$$T(\mathbf{s}, \mathbf{a}) \approx \sum_{k=1}^{N} s_k a_k \mathbf{P}\{I_{\Phi^a_{P,k}}(Y_i) < T_1\} \mathcal{L}_{I_{\Phi^a_{P,k}}(y_i)}(\mu_s T_2 R^\alpha) \mathcal{L}_{I_{\Phi_{S,k}\setminus Y_i}(y_i)}(\mu_s T_2 R^\alpha)$$

where $\mathbf{s} = (s_1, \cdots, s_N)$ and $\mathbf{a} = (a_1, \cdots, a_N)$.

4 Numerical Results

In this section, we provide some numerical results to validate our analysis. We use Matlab to simulate the cognitive radio network. The following network parameters are used in numerical and simulation analysis. The number of wireless channels is $N = 2$. The averages of transmission powers are $\frac{1}{\mu_p} = 40\ mW$ and $\frac{1}{\mu_s} = 8\ mW$. The path loss exponent is $\alpha = 4$. The distance between a transmitter and receiver pair is $R = 10\ m$. The success threshold is $T_2 = 1\ dB$. The observation space window is $200 \times 200\ m^2$.

We investigate the behavior of throughput for different parameters values. Fig. 1 shows the throughput as we change b_1 from 0 to $1-b_2$ where $b_i = s_i a_i$, $1 \leq i \leq 2$. In the figure, b_2 is chosen to maximize the throughput of channel 2, i.e., $b_2 = 0.4143$. Other parameters are $\pi_{1,0}\lambda_{p,1} = 0.00005$ and $\pi_{2,0}\lambda_{p,2} = 0.00015$. From Fig. 1 where the distribution of SUs is dense, we see that throughput is concave and the optimal throughput is achieved at some point. Moreover, we see that our analytic results are well matched with the simulation results.

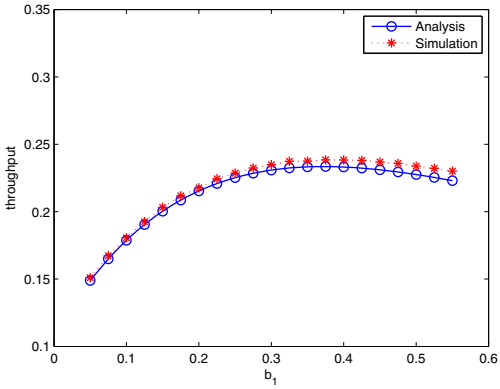


Fig. 1 Throughput ($\lambda_s = 0.005$, $T_1 = 0.0001$)

5 Conclusions

In this paper, we considered an underlay type cognitive radio network with multiple secondary users. We analyzed the throughput performance of the tagged secondary user with the help of stochastic geometry. We introduced the interference-free region to analyze the throughput of an arbitrary secondary user. Our numerical results validated our analysis.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2014R1A1A2055410).

References

1. Goldsmith, A., Jafar, S.A., Maric, I., Srinivasa, S.: Breaking spectrum gridlock with cognitive radios: an information theoretic perspective. Proc. IEEE **97**(5), 894–914 (2009)
2. Stoyan, D., Kendall, W., Mecke, J.: Stochastic Geometry and Its Applications, 2nd edn. John Wiley and Sons (1996)
3. Haenggi, M., Andrews, J.G., Baccelli, F., Dousse, O., Franceschetti, M.: Stochastic Geometry and Random Graphs for the Analysis and Design of Wireless Networks. IEEE J. Select. Areas Commun. **27**(7), September 2009

4. Yin, C., Chen, C., Liu, T., Cui, S.: Generalized results of transmission capacities for overlaid wireless networks. In: Proc. IEEE Int. Symp. Inf. Theory, Seoul, Korea, pp. 1774–1778, June 2009
5. Vaze, R.: Transmission capacity of spectrum sharing ad hoc networks with multiple antennas. IEEE Trans. Wireless Commun. **10**(7), 2334–2340 (2011)
6. Lee, J., Andrews, J.G., Hong, D.: Spectrum-sharing transmission capacity. IEEE Trans. Wireless Commun. **10**(9), 3053–3063 (2011)
7. Nguyen, T.V., Baccelli, F.: A probabilistic model of carrier sensing based cognitive radio. In: 2010 IEEE Symposium on New Frontiers in Dynamic Spectrum, pp. 1–12, April 2010
8. Lee, C., Haenggi, M.: Interference and Outage in Poisson Cognitive Networks. IEEE Trans. Wireless Commun. **11**(4), 1392–1401 (2012)
9. Song, X., Yin, C., Liu, D., Zhang, R.: Spatial opportunity in cognitive radio networks with threshold-based opportunistic spectrum access. In: 2013 IEEE International Conference on Communications (ICC), pp. 2695–2700, June 2013
10. Busson, A., Jabbari, B., Babaei, A., Vèque, V.: Interference and Throughput in Spectrum Sensing Cognitive Radio Networks using Point Processes. Journal of Communications and Networks **16**(1), 67–80 (2014)
11. Moltchanov, D.: Distance distributions in random networks. Ad Hoc Networks **10**(6), 1146–1166 (2012)

Performance Comparison Between Two Kinds of Priority Schemes in Cognitive Radio Networks

Yuan Zhao and Wuyi Yue

Abstract In this paper, we consider a cognitive radio network with multiple Secondary Users (SUs). The SU packets are divided into SU1 packets and SU2 packets, and the SU1 packets have higher priority than the SU2 packets. Different from the conventional preemptive priority scheme (called Scheme I), we propose a non-preemptive priority scheme for the SU1 packets (called Scheme II) to guarantee the transmission continuity of the SU2 packets. By constructing a three-dimensional Markov chain, we give the transition probability matrix of the Markov chain, and obtain the steady-state distribution of the system model. Accordingly, we derive some performance measures, such as the interrupted rate of the SU1 packets and the interrupted rate of the SU2 packets. Lastly, we provide numerical experiments to compare the system performance between the two priority schemes.

Keywords Cognitive radio networks · Preemptive priority · Non-preemptive priority · Markov chain

1 Introduction

In conventional cognitive radio networks there are two kinds of users, namely, Primary Users (PUs) and Secondary Users (SUs). The PUs have priority, with the SUs making use of the licensed spectrum only when the spectrum is not occupied by the PUs [1]. Most of studies in cognitive radio networks have been focused on the interaction between PUs and SUs. Hamza et al. assumed that the SU in the

Y. Zhao
School of Computer and Communication Engineering, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China
e-mail: yuanzh85@163.com

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

T.V. Do et al. (eds.), *Queueing Theory and Network Applications*,
Advances in Intelligent Systems and Computing 383,
DOI: 10.1007/978-3-319-22267-7_7

system acted as a relay for the PU in the event of transmission failure [2]. Using queueing analysis, they obtained the throughputs of the PUs and the SUs in the system. Considering that the transmission of the SUs may be interrupted by the PUs, Chu et al. proposed a spectrum handoff strategy in the overlay cognitive radio networks [3]. Their numerical results showed that their proposed spectrum handoff strategy could effectively reduce the waiting time during the spectrum handoff. However, the studies mentioned above did not attribute different priority levels to SUs.

We note that there are various types of data in communication networks, for example, real-time data and non-real-time data. The real-time data requires higher priority. So, it is also necessary to grade the SUs in cognitive radio networks. Lee et al. divided the SUs into SU1 and SU2 in a multi-channel cognitive radio network [4]. They assumed the SU1 calls have higher priority than the SU2 calls, and the SU1 calls can interrupt the transmission of the SU2 calls. By building a continuous-time Markov chain, they analyzed the performance of the SU1 and the SU2, respectively. Zhang et al. equipped the SUs with different priority levels to guarantee the quality of service for the SUs with higher priority [5]. By applying a preemptive resume priority M/M/1 queueing model, they evaluated the transmission delay for the interrupted SUs.

Moreover, we note that in most of the research considering the prioritization of the SUs in cognitive radio networks, the SUs with higher priority would interrupt the transmission of the SUs with lower priority immediately (preemptive priority scheme). However, in practical networks, in order to guarantee the transmission continuity of the SUs, the SUs with higher priority may wait until the transmission of the SUs with lower priority are completed.

In this paper, we propose a non-preemptive priority scheme for the SU packets with higher priority. Considering the digital nature of modern networks, we construct a three-dimensional discrete-time Markov chain and then derive some performance measures. In addition, we provide numerical experiments to compare the system performance between the two priority schemes.

2 System Model

In this paper, we consider a cognitive radio network with a single channel. The PU packets are supposed to have preemptive priority over the SU packets. The packets generated from the SUs are classified into SU1 packets and SU2 packets, and the SU1 packets have higher priority than the SU2 packets. Considering the lowest priority of the SU2 packets, a buffer is prepared for the SU2 packets to reduce possible loss of those packets. We call this buffer the “SU2 buffer".

When a PU packet arrives at the system, if the channel is occupied by another PU packet, this newly arriving packet will leave the system to find another available channel. If the channel is occupied by an SU1 packet, the transmission of that SU1 packet will be interrupted by the PU packet and the interrupted SU1 packet will leave the system to find another available channel. If the channel is occupied by an SU2

packet, the transmission of the SU2 packet will also be interrupted by the PU packet and the interrupted SU2 packet will return back to the SU2 buffer.

The SU1 packets have a higher priority access to the channel than the SU2 packets. For example, when an SU1 packet and an SU2 packet arrive at the system simultaneously (there is no PU packet arrival), if the channel is idle, the newly arriving SU1 packet will occupy the channel, while the newly arriving SU2 packet has to queue in the SU2 buffer.

In the case of an SU1 packet arriving at the system (namely, there is no PU packet arrival) during the transmission of an SU2 packet, we propose a non-preemptive priority scheme. We assume the newly arriving SU1 packet will be blocked and leave the system to find another available channel in order to guarantee the transmission continuity of the SU2 packets. In the following, in order to clarify the presentation, we call the preemptive priority scheme where the newly arriving SU1 packet will interrupt the transmission of the SU2 packet and occupy the channel immediately "Scheme I" and the proposed non-preemptive priority scheme we call "Scheme II".

We assume an early arriving system with a slotted timing structure, and the time axis is ordered by $t = 1, 2, \ldots$. We suppose that the arrival intervals of the PU packets, the SU1 packets and the SU2 packets follow geometrical distributions with parameters λ_1, λ_{21} and λ_{22}, respectively. Moreover, we assume that the transmission time of a PU packet, an SU1 packet and an SU2 packet follow geometrical distributions with parameters μ_1, μ_{21} and μ_{22}, respectively.

We denote $L_n^{(1)}$, $L_n^{(21)}$ and $L_n^{(22)}$ as the number of PU packets, SU1 packets and SU2 packets in the system at the instant $t = n^+$, respectively, where n represents the time epoch of the slot boundary. Then, $\left\{L_n^{(22)}, L_n^{(21)}, L_n^{(1)}\right\}$ constitutes a three-dimensional discrete-time Markov chain with the state space $\boldsymbol{M}$ as follows:

$$\boldsymbol{M} = \{(i, 0, 0) \cup (i, 0, 1) \cup (i, 1, 0) : 0 \leq i \leq \infty\}. \tag{1}$$

3 Performance Analysis

Let $\boldsymbol{P}$ be the state transition probability matrix of the three-dimensional discrete-time Markov chain. $\boldsymbol{P}$ can be given in a block-structure form as follows:

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{C}_0 & \boldsymbol{B}_0 & & \\ \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & \\ & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 \\ & & \ddots & \ddots & \ddots \end{pmatrix} \tag{2}$$

where each non-zero block in $\boldsymbol{P}$ is a 3×3 matrix and can be discussed as follows. Hereafter, we use the overbar notation to denote the probability of a complement event, for instance, $\bar{\lambda}_1 = 1 - \lambda_1$. Moreover, we introduce $\zeta = \lambda_{21}\mu_{21} + \bar{\mu}_{21}$ and $\vartheta = \lambda_{22}\mu_{22} + \bar{\lambda}_{22}\bar{\mu}_{22}$ in following equations for compactness of presentation.

(1)$\boldsymbol{C}_0$ is the one-step transition probability matrix for the number of SU2 packets in the system being fixed at 0. $\boldsymbol{C}_0$ can be given by

$$\boldsymbol{C}_0 = \bar{\lambda}_{22}\boldsymbol{U} \tag{3}$$

where $\boldsymbol{U}$ can be given as follows:

$$\boldsymbol{U} = \begin{pmatrix} \bar{\lambda}_1\bar{\lambda}_{21} & \lambda_1 & \bar{\lambda}_1\lambda_{21} \\ \bar{\lambda}_1\bar{\lambda}_{21}\mu_1 & \lambda_1\mu_1 + \bar{\mu}_1 & \bar{\lambda}_1\lambda_{21}\mu_1 \\ \bar{\lambda}_1\bar{\lambda}_{21}\mu_{21} & \lambda_1 & \bar{\lambda}_1\zeta \end{pmatrix}. \tag{4}$$

(2)$\boldsymbol{B}_0$ is the one-step transition probability matrix for the number of SU2 packets in the system increasing from 0 to 1. $\boldsymbol{B}_0$ can be given by

$$\boldsymbol{B}_0 = \lambda_{22}\boldsymbol{U}. \tag{5}$$

(3)$\boldsymbol{A}_2$ is the one-step transition probability matrix for the number of SU2 packets in the system decreasing from i to $i-1$ ($1 \le i \le \infty$). $\boldsymbol{A}_2$ can be given by

$$\boldsymbol{A}_2 = \mu_{22}\bar{\lambda}_{22}\boldsymbol{V} \tag{6}$$

where $\boldsymbol{V}$ can be given as follows:

$$\boldsymbol{V} = \begin{pmatrix} \bar{\lambda}_1\bar{\lambda}_{21} & \lambda_1 & \bar{\lambda}_1\lambda_{21} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{7}$$

(4)$\boldsymbol{A}_1$ is the one-step transition probability matrix for the number of SU2 packets in the system being fixed at i ($1 \le i \le \infty$). $\boldsymbol{A}_1$ can be given as follows:
For the case of Scheme I, $\boldsymbol{A}_1$ can be given by

$$\boldsymbol{A}_1 = \begin{pmatrix} \bar{\lambda}_1\bar{\lambda}_{21}\vartheta & \lambda_1\vartheta & \bar{\lambda}_1\lambda_{21}\vartheta \\ \bar{\lambda}_{22}\bar{\lambda}_1\bar{\lambda}_{21}\mu_1 & \bar{\lambda}_{22}(\lambda_1\mu_1 + \bar{\mu}_1) & \bar{\lambda}_{22}\bar{\lambda}_1\lambda_{21}\mu_1 \\ \bar{\lambda}_{22}\bar{\lambda}_1\bar{\lambda}_{21}\mu_{21} & \bar{\lambda}_{22}\lambda_1 & \bar{\lambda}_{22}\bar{\lambda}_1\zeta \end{pmatrix}. \tag{8}$$

For the case of Scheme II, $\boldsymbol{A}_1$ can be given by

$$\boldsymbol{A}_1 = \begin{pmatrix} \bar{\lambda}_1(\bar{\lambda}_{22}\bar{\mu}_{22} + \lambda_{22}\mu_{22}\bar{\lambda}_{21}) & \lambda_1\vartheta & \bar{\lambda}_1\lambda_{21}\lambda_{22}\mu_{22} \\ \bar{\lambda}_{22}\bar{\lambda}_1\bar{\lambda}_{21}\mu_1 & \bar{\lambda}_{22}(\lambda_1\mu_1 + \bar{\mu}_1) & \bar{\lambda}_{22}\bar{\lambda}_1\lambda_{21}\mu_1 \\ \bar{\lambda}_{22}\bar{\lambda}_1\bar{\lambda}_{21}\mu_{21} & \bar{\lambda}_{22}\lambda_1 & \bar{\lambda}_{22}\bar{\lambda}_1\zeta \end{pmatrix}. \tag{9}$$

(5)$\boldsymbol{A}_0$ is the one-step transition probability matrix for the number of SU2 packets in the system increasing from i to $i+1$ ($1 \le i \le \infty$). $\boldsymbol{A}_0$ can be given by

$$\boldsymbol{A}_0 = \lambda_{22}\boldsymbol{W}. \tag{10}$$

For the case of Scheme I, $\boldsymbol{W}$ can be given as follows:

$$\boldsymbol{W} = \begin{pmatrix} \bar{\lambda}_1\bar{\lambda}_{21}\bar{\mu}_{22} & \lambda_1\bar{\mu}_{22} & \bar{\lambda}_1\lambda_{21}\bar{\mu}_{22} \\ \bar{\lambda}_1\bar{\lambda}_{21}\mu_1 & \lambda_1\mu_1 + \bar{\mu}_1 & \bar{\lambda}_1\lambda_{21}\mu_1 \\ \bar{\lambda}_1\bar{\lambda}_{21}\mu_{21} & \lambda_1 & \bar{\lambda}_1\zeta \end{pmatrix}. \tag{11}$$

For the case of Scheme II, $\boldsymbol{W}$ can be given follows:

$$\boldsymbol{W} = \begin{pmatrix} \bar{\mu}_{22}\bar{\lambda}_1 & \bar{\mu}_{22}\lambda_1 & 0 \\ \bar{\lambda}_1\bar{\lambda}_{21}\mu_1 & \lambda_1\mu_1 + \bar{\mu}_1 & \bar{\lambda}_1\lambda_{21}\mu_1 \\ \bar{\lambda}_1\bar{\lambda}_{21}\mu_{21} & \lambda_1 & \bar{\lambda}_1\zeta \end{pmatrix}. \tag{12}$$

The steady-state distribution $\pi_{i,j,k}$ of $\left\{L_n^{(22)}, L_n^{(21)}, L_n^{(1)}\right\}$ is then defined as

$$\pi_{i,j,k} = \lim_{n\to\infty} P\left\{L_n^{(22)} = i, L_n^{(21)} = j, L_n^{(1)} = k\right\} \tag{13}$$

where $0 \le i \le \infty$, $j = 0, 1$, $k = 0, 1$. Moreover, we note that j and k can not be equate to 1 at the same time.

The structure of the transition probability matrix $\boldsymbol{P}$ indicates that the three-dimensional Markov chain follows a Quasi Birth and Death (QBD) process. By using the matrix-geometric solution method [6], we can obtain the numerical results for the steady-state distribution $\pi_{i,j,k}$ defined in Eq. (13).

Next, by using the steady-state distribution $\pi_{i,j,k}$, we present various performance measures of this system model.

We define the interrupted rate γ_{21} of the SU1 packets as the number of SU1 packets that are interrupted by the PU packets per slot. The expression of the interrupted rate γ_{21} of the SU1 packets can be given as follows:

$$\gamma_{21} = \sum_{i=0}^{\infty} \pi_{i,1,0}\bar{\mu}_{21}\lambda_1. \tag{14}$$

We define the interrupted rate γ_{22} of the SU2 packets as the number of SU2 packets that are interrupted by the SU1 packets or the PU packets per slot. The expression of the interrupted rate γ_{22} of the SU2 packets can be given for two cases.

For the case of Scheme I:

$$\gamma_{22} = \sum_{i=1}^{\infty} \pi_{i,0,0}\bar{\mu}_{22}(1 - \bar{\lambda}_1\bar{\lambda}_{21}). \tag{15}$$

For the case of Scheme II:

$$\gamma_{22} = \sum_{i=1}^{\infty} \pi_{i,0,0} \bar{\mu}_{22} \lambda_1. \tag{16}$$

4 Numerical Experiments

In this section, we compare the interrupted rate of the SU1 packets and the interrupted rate of the SU2 packets between Scheme I and Scheme II. The time length of one slot is assumed to be 1 ms. By referencing [7] and following the IEEE 802.11 b/g standard, the data rate in Physical Layer is assumed to be 11 Mbps. The average packet size is assumed to be 2, 750 Bytes. Moreover, the arrival rate λ_{22} of the SU2 packets is assumed to be $\lambda_{22} = 0.1$.

Figures 1 compares the interrupted rate γ_{21} of the SU1 packets between Scheme I and Scheme II.

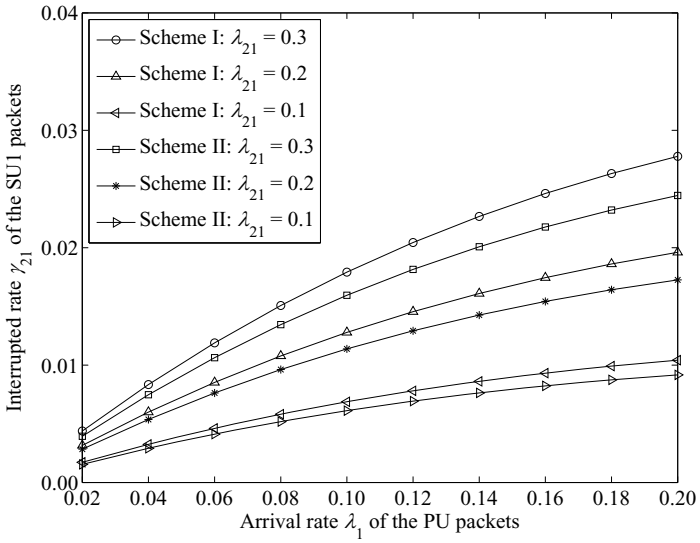


Fig. 1 Change trend for the interrupted rate γ_{21} of the SU1 packets.

From Fig. 1, we observe that the interrupted rate γ_{21} of the SU1 packets increases as the arrival rate λ_1 of the PU packets increases. This is because as the arrival rate of the PU packets increases, the possibility for the transmission of the SU1 packets to be interrupted by the PU packets will be higher, and this will increase the interrupted rate of the SU1 packets.

On the other hand, as shown in Fig. 1, the interrupted rate γ_{21} of the SU1 packets increases as the arrival rate λ_{21} of the SU1 packets increases. The reason is that the

larger the arrival rate of the SU1 packets is, the more the SU1 packets will occupy the channel, and this will result in a greater interrupted rate of the SU1 packets.

Furthermore, the interrupted rate γ_{21} of the SU1 packets in Scheme I is greater than that in Scheme II for the same parameter settings. The reason is that in Scheme I, a newly arriving SU1 packet can interrupt the transmission of the SU2 packet on the channel. In other words, in the case of Scheme I, the possibility of the SU1 packets occupying the channel is higher, and the possibility for the transmission of the SU1 packets to be interrupted by the PU packets will also be higher. As a result, the interrupted rate of the SU1 packets will be greater in Scheme I.

Figure 2 compares the interrupted rate γ_{22} of the SU2 packets between Scheme I and Scheme II.

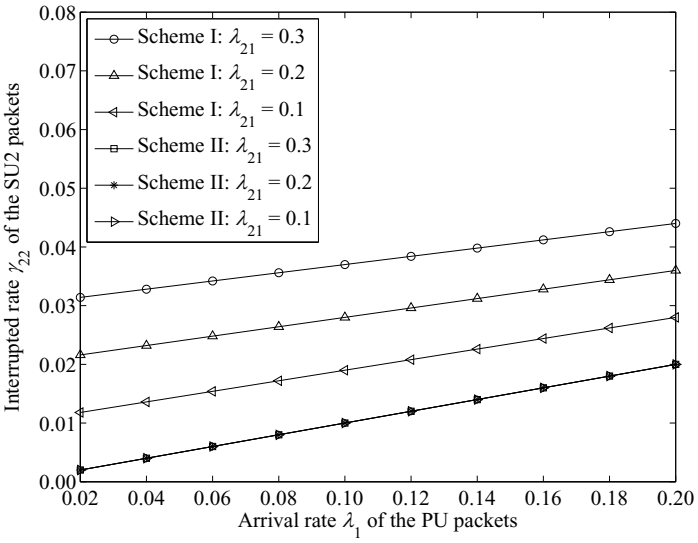


Fig. 2 Change trend for the interrupted rate γ_{22} of the SU2 packets.

From Fig. 2, we find that as the arrival rate λ_1 of the PU packets increases, the interrupted rate γ_{22} of the SU2 packets shows an increasing tendency. This is obviously because the greater the arrival rate of the PU packets is, the higher the possibility that the transmission of the SU2 packets to be interrupted, so the larger the interrupted rate of the SU2 packets will be.

On the other hand, Fig. 2 shows an increasing arrival rate λ_{21} of the SU1 packets causes an increase in the interrupted rate γ_{22} of the SU2 packets in Scheme I. This is because in Scheme I, a newly arriving SU1 packet can interrupt the transmission of the SU2 packet, and this will result in a higher interrupted rate of the SU2 packets.

Furthermore, as the arrival rate λ_{21} of the SU1 packets increases, the interrupted rate γ_{22} of the SU2 packets will not be changed in Scheme II. This is because in Scheme II, a newly arriving SU1 packet will not interrupt the transmission of the SU2 packets, so the arrival rate of the SU1 packets will not influence the interrupted

rate of the SU2 packets. Moreover, because of the preemptive priority mechanism, Scheme I experiences a higher interrupted rate γ_{22} of the SU2 packets than Scheme II.

5 Conclusions

In this paper, we investigated the system performance of cognitive radio networks, in which the SU packets in the system were divided into SU1 packets with higher priority and SU2 packets with lower priority. For the purpose of guaranteeing the transmission continuity of the SU2 packets, a non-preemptive priority scheme was proposed for the SU1 packets. A three-dimensional discrete-time Markov chain was constructed and the transition probability matrix was given. With the steady-state distribution, some performance measures for the SU1 packets and SU2 packets were derived. Finally, with numerical experiments, we showed that compared with the preemptive priority scheme, the proposed non-preemptive priority scheme for the SU1 packets could reduce the interrupted rate of the SU2 packets effectively.

Acknowledgments This work was supported in part by National Natural Science Foundation (No. 61472342), China and was supported in part by MEXT, Japan.

References

1. Zhang, Z., Long, K., Wang, J.: Self-organization paradigms and optimization approaches for cognitive radio technologies: A survey. IEEE Wireless Communications **20**, 36–42 (2013)
2. Hamza, D., Aïssa, S.: Enhanced primary and secondary performance through cognitive relaying and leveraging primary feedback. IEEE Transactions on Vehicular Technology **63**, 2236–2247 (2014)
3. Chu, J., Ma, R., Feng, K.: Stochastic spectrum handoff protocols for partially observable cognitive radio networks. Wireless Networks **20**, 1003–1022 (2014)
4. Lee, Y., Park, C.G., Sim, D.B.: Cognitive radio spectrum access with prioritized secondary users. Applied Mathematics & Information Sciences **6**, 595S–601S (2012)
5. Zhang, Y., Jiang, T., Zhang, L., Qu, D., Peng, W.: Analysis on the transmission delay of priority-based secondary users in cognitive radio networks. In: Proceedings of the International Conference on Wireless Communications & Signal Processing (WCSP), CD-ROM, 6 pages. IEEE Press, Hangzhou (2013)
6. Alfa, A.S.: Queueing Theory for Telecommunications: Discrete Time Modelling of a Single Node System. Springer, New York (2010)
7. Kim, K.J., Kwak, K.S., Choi, B.D.: Performance analysis of opportunistic spectrum access protocol for multi-channel cognitive radio networks. Journal of Communications and Networks **15**, 77–86 (2013)

Performance Analysis of Binary Exponential Backoff MAC Protocol for Cognitive Radio in the IEEE 802.16e/m Network

Shengzhu Jin, Bong Dae Choi and Doo Seop Eom

Abstract We propose a distributed MAC protocol for cognitive radio when primary network is IEEE 802.16e/m WiMAX. Our proposed MAC protocol is Truncated Binary Exponential Backoff Algorithm where backoff stage of algorithm is doubled at each collision, and backoff counter is operated by frame basis and is freezed at a frame with no idle slots. We model our proposed MAC protocol as a 3-dimensional discrete-time Markov chain and obtain steady state probability of the Markov chain by using a censored Markov chain method. Based on this steady state probability, we obtain the throughput, packet loss probability and packet delay distribution of secondary users. Our numerical examples show that initial contention window size can be determined according to the number of secondary users in order to obtain higher throughput for secondary users, and the maximum backoff stage has a large impact on the secondary user's packet loss probability.

Keywords Cognitive radio · Exponential backoff MAC protocol · Censored Markov chain · Throughput

S. Jin
ROBOTIS Co., Ltd., Seoul, Korea
e-mail: irarykim@gmail.com

B.D. Choi (✉)
Research Institute for ICT, Korea University, Seoul, Korea
e-mail: queue@korea.ac.kr

D.S. Eom
The School of Electrical Engineering, Korea University, Seoul, Korea
e-mail: eomds@korea.ac.kr

B.D. Choi—This work was supported by the National Research Foundation of Korea grants funded by Korea government(MEST)(No.2012-008099).

T.V. Do et al. (eds.), *Queueing Theory and Network Applications*,
Advances in Intelligent Systems and Computing 383,
DOI: 10.1007/978-3-319-22267-7_8

1 Introduction

Cognitive radio technologies have been receiving a lot of attention from industry and academia since it is known as a promising technique to enhance the utilization of the existing radio spectrum[1]. The main idea behind a cognitive radio is that the unlicensed users (also called secondary users (SUs)) opportunistically exploit the licensed spectrum unused by the primary users (PUs).

For related works applicable to IEEE 802.16e/m WiMAX [2],[3], we focus on the papers [4]and [5] where MAC protocols adopt 802.11 binary backoff scheme. Fantacci et al.[4] proposed scheduling based and contention based cognitive radio MAC protocols for SUs with IEEE 802.16 networks as the PU network, and simulation are presented. Chong et al.[5] proposed a slot-based MAC protocol for SUs where primary network has multiple channels and operates on frame-by-frame basis. To the best of our knowledge, there has been no work on comprehensive QoS analysis of cognitive radio network whose primary network is IEEE 802.16e/m.

In this paper, we propose a distributed contention-based MAC protocol for SUs in a cognitive radio network where the primary network is IEEE 802.16e/m. As a MAC protocol for SUs, a truncated binary exponential backoff (TBEB) scheme is adopted and the backoff unit of the algorithm is one frame in IEEE 802.16m WiMAX. We construct a 3-dimensional Markov chain incorporating the variation of the number of idle slots in a frame and (backoff stage, backoff counter) of the tagged SU. By applying censored Markov chain method, we obtain the steady state probability of Markov chain. Based on the steady state probability of the Markov chain, we obtain throughput, packet loss probability.

2 System Model

In this paper, the primary network of cognitive radio network is IEEE 802.16e/m network which is operated in Time Division Duplexing (TDD) mode, because it is the most commonly used scheme in practical implementation of WiMAX systems.

In TDD mode, a frame is divided into two subframes: a downlink (from base station(BS) to mobile station(MS)) subframe followed by an uplink (from MS to BS) subframe. The BS for primary users allocates resources in data regions of uplink subframes and downlink subframes to PUs for uplink and downlink transmissions, respectively, and this scheduling information is broadcasted to PUs through DL-MAP and UL-MAP which are the beginning parts of every downlink subframe. The DL-MAP and UL-MAP in each downlink subframe contains the allocation message for resources of current downlink subframe and following uplink subframe. In this work, we only consider the uplink transmission, thus we only focus on resources of uplink subframes. The data region of an uplink subframe consists of *slots*. A slot is a resource unit assigned to PUs. Since the uplingk traffic size is affected by the PU's activities (i.e., time-varying), thus there might be some slots unused by PUs in each uplink frame and the number of such slots (*idle slots*) is time-varying.

By allowing SUs to use the idle slots, the utilization of resources can be significantly improved. Therefore, it is an important issue to develop an efficient opportunistic spectrum access scheme for SUs.

In practice, the size of slots are different according to the frequency band it locates and also according to whether the operation is Time Division Multiplexing (TDM) or Frequency Division Multiplexing (FDM). However, in our work, for simplicity, we assume that the slot size is a constant and thus the number of slots in an uplink subframe is fixed.

3 Cognitive Radio Protocol

In this section, we introduce the operation of PUs and SUs in our proposed protocol.

3.1 Primary Users' Operations

We assume that the PUs' traffic is voice data generated by N_p independent active primary voice sources (PV sources). Each active PV source alternates between talkspurt and silent periods. Voice data is generated during talkspurt periods, while no data is generated during silent periods which is due either to listening periods or gaps between words. Let each PV source occupies R_t slots per uplink subframe during talkspurt periods. We model the operation of each PV source as a discrete time ON-OFF process, with one frame as the time unit, and the one-step transition probability matrix is given as follows.

$$\mathbf{P}_{\text{voice}} = \begin{array}{c} \\ \text{talkspurt} \\ \text{silent} \end{array} \begin{array}{c} \begin{array}{cc} \text{talkspurt} & \text{silent} \end{array} \\ \begin{pmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{pmatrix} \end{array}, \tag{1}$$

Let N_{slot} be the number of slots in an uplink subframe and X_n be the the number of PV sources in talkspurt period during the nth uplink subframe. If $X_n > N_{slot}$, we assume that only N_{slot} of them are allowed to transmit voice data while others are blocked. The mechanism of admission control of PV sources is the outside the scope of this paper.

3.2 Secondary Users' Operations

As mentioned above, the BS broadcasts resource allocation messages for PUs through DL-MAP and UL-MAP in a control message at the beginning of each frame. We assume that the SUs can get information on the number and the positions of idle slots in the frame. This can be done either by overhearing the UL-MAP or cognitive radio BS (CR-BS) broadcasting this information to SUs. Another assumption is that each SU uses only one slot at every frame.

For distributed opportunistic spectrum access for SUs, we adopt truncated binary exponential backoff (TBEB) scheme whose backoff unit is a frame unit in IEEE 802.16 e/m WiMaX. The detailed operation for TBEB scheme is as follows:

– Whenever a SU has a packet to transmit, the tagged SU chooses a backoff counter value uniformly in $[0, W_0 - 1]$ where W_0 is called the initial contention window and SU is called at its backoff stage 0.
– If at least one of N_{slot} slots is idle at a frame, the SU decreases its backoff counter by one regardless of other SUs' operations in next frame, otherwise, the SU's backoff counter is frozen.
– If the backoff counter of the SU reaches zero at a frame and in that frame there is at least one idle slot unused by PUs, the SU chooses a slot randomly among the idle slot(s) and transmit its packet in that slot.
  - If no other SUs choose the same slot as the SU chose, the packet's transmission of the SU is successful.
  - Otherwise, i.e., if another SU chooses the same slot as the tagged SU chose, both packets collide.
– Suppose a SU with backoff stage i and contention window W_i collides,
  - if i is less than pre-defined maximum backoff stage m, this SU increases its backoff stage by one and doubles its contention window ($W_{i+1} = 2 \times W_i$);
  - If $i = m$, this SU gives up transmitting the packet (i.e., this packet is lost) and sets its backoff stage to 0, i.e., each SU only has m retrial chances.

4 Markov Chains for Describing PUs and a Tagged SU's Operations

In this section, we first model the operation of a tagged PU as a discrete time ON-OFF process, then we construct a 3-dimensional Markov chain to evaluate the scheme for SUs described in section 2 and find stationary probability vector by following the censored Markov chain method.

4.1 Markov Chain for PUs' Operations

According to the assumptions in 3.1, it can be easily found that X_n (the the number of PV sources in talkspurt period during the nth uplink subframe) follows a discrete time Markov chain. Let $p_{i,j}$ be the one step transition probability of this Markov chain. Then we obtain $p_{i,j}$ from $\mathbf{P_{voice}}$ in Equation (1) as follows.

$$p_{i,j} = \sum_{k=\max\{0,i+j-N_p\}}^{\min\{i,j\}} \alpha^k (1-\alpha)^{i-k} (1-\beta)^{j-k} \beta^{N_p-i-j+k} \quad , \quad i, j \in [0, N_p] \quad (2)$$

4.2 Markov Chain for a Tagged SU's Operation

Our Markov chain for a tagged SU has states (i, j, k) where the backoff stage i, backoff counter j of the tagged SU and the number k of busy PV sources in the frame. We assume a saturated condition, which means each SU always has packets to send. An unsaturated situation can be modeled and analyzed similarly by adding an *idle state* to current Markov chain.

Let q_k be the collision probability of a tagged SU when the number of busy PV sources is k, given that the tagged SU transmits a packet. The one-step transition probability matrix of the 3-dimensional Markov chain will be expressed in terms of q_k, and conversely q_k will be expressed in terms of steady state probabilities of this Markov chain. (see Equation (5)). The state transition probabilities for this 3-dimensional Markov chain are listed below:

– If $N_p \geq N_{slot}$

$$
\begin{aligned}
&P_{(i,j,k)(i,j,k')} = p_{k,k'} && i \in [0,m],\ j \in [0, W_i - 1],\ k \in [N_{slot}, p],\ k' \in [0, N_p] \\
&P_{(i,j,k)(i,j-1,k')} = p_{k,k'} && i \in [0,m],\ j \in [1, W_i - 1],\ k \in [0, N_{slot} - 1],\ k' \in [0, N_p] \\
&P_{(i,0,k)(0,j',k')} = (1-q_k)p_{k,k'}/W_0 && i \in [0, m-1],\ j' \in [0, W_0 - 1],\ k \in [0, N_{slot} - 1],\ k' \in [0, N_p] \\
&P_{(m,0,k)(0,j',k')} = p_{k,k'}/W_0 && j' \in [0, W_0 - 1],\ k \in [0, N_{slot} - 1],\ k' \in [0, N_p] \\
&P_{(i,0,k)(i+1,j',k')} = q_k p_{k,k'}/W_{i+1} && i \in [0, m-1],\ j' \in [0, W_{i+1} - 1],\ k \in [0, N_{slot} - 1],\ k' \in [0, N_p]
\end{aligned}
$$

– If $N_p < N_{slot}$

$$
\begin{aligned}
&P_{(i,j,k)(i,j-1,k')} = p_{k,k'} && i \in [0,m],\ j \in [1, W_i - 1],\ k \in [0, N_p],\ k' \in [0, N_p] \\
&P_{(i,0,k)(0,j',k')} = (1-q_k)p_{k,k'}/W_0 && i \in [0, m-1],\ j' \in [0, W_0 - 1],\ k \in [0, N_p],\ k' \in [0, N_p] \\
&P_{(m,0,k)(0,j',k')} = p_{k,k'}/W_0 && j' \in [0, W_0 - 1],\ k \in [0, N_p],\ k' \in [0, N_p] \\
&P_{(i,0,k)(i+1,j',k')} = q_k p_{k,k'}/W_{i+1} && i \in [0, m-1],\ j' \in [0, W_{i+1} - 1],\ k \in [0, N_p],\ k' \in [0, N_p]
\end{aligned}
$$

Here, W_i is the size of contention window of backoff stage i; m is the maximum backoff stage.

This matrix has following form:

$$
\mathbf{P} = \begin{pmatrix}
\mathbf{B_{0,0}} & \mathbf{A_{0,1}} & 0 & 0 & \cdots & 0 & 0 \\
\mathbf{B_{1,0}} & \mathbf{C_{1,1}} & \mathbf{A_{1,2}} & 0 & \cdots & 0 & 0 \\
\mathbf{B_{2,0}} & 0 & \mathbf{C_{2,2}} & \mathbf{A_{2,3}} & \cdots & 0 & 0 \\
\mathbf{B_{3,0}} & 0 & 0 & \mathbf{C_{3,3}} & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\mathbf{B_{m-1,0}} & 0 & 0 & 0 & \cdots & \mathbf{C_{m-1,m-1}} & \mathbf{A_{m-1,m}} \\
\mathbf{B_{m,0}} & 0 & 0 & 0 & \cdots & 0 & \mathbf{C_{m,m}}
\end{pmatrix}. \tag{3}
$$

$\mathbf{B_{i,0}}$, $\mathbf{A_{i,i+1}}$ and $\mathbf{C_{i,i}}$ can be specified by above one-step transition probabilities. Let $\Pi = (\pi_{\overline{0}}, \pi_{\overline{1}}, ..., \pi_{\overline{m}})$ be the steady state probability vector of this Markov chain. Denote $S = \{\overline{0}, \overline{1}, ..., \overline{m}\}$ where $\overline{i}$ is level i, i.e., $\overline{i} = \{(I, J, K) | I = i\}$. $\Pi = (\pi_{\overline{0}}, \pi_{\overline{1}}, ..., \pi_{\overline{m}})$ can be obtained by the censored Markov chain method whose detailed derivation is omitted.

5 Performance Measures

Let τ_k be the transmission probability of the tagged SU when the number of busy PV sources is k, and q_k be the collision probability of a tagged SU when the number of busy PV sources is k, given that the tagged SU transmits a packet. Then τ_k and q_k are given by

$$\tau_k = \frac{\sum_{i=0}^{m} \pi_{i,0,k}}{\sum_{i=0}^{m} \sum_{j=0}^{W_i-1} \pi_{i,j,k}}, \quad k \in [0, \min(N_p, N_{slot}) - 1] \tag{4}$$

$$q_k = 1 - (1 - \tau_k \cdot \frac{1}{N_{slot} - k})^{N_s - 1}, \quad k \in [0, \min(N_p, N_{slot}) - 1] \tag{5}$$

In Equation (5), $(1 - \tau_k \cdot \frac{1}{N_{slot}-k})^{N_s-1}$ represents the probability that all of the other $N_s - 1$ secondary users (except the tagged secondary user) do not choose the same slot as the tagged SU chose when there are k busy PV sources in that frame.

We define the throughput T of secondary users as the ratio of the average number of slots successfully used by SUs to the total number of slots in a long enough duration.

Given $\mathbf{P_c} = (p_{i,j})$, the one-step transition probability matrix of the number of busy PV sources, we can find $C = (c_k)$, the steady state probability that there are k busy PV sources by solving $C\mathbf{P_c} = C$. Given that there are k busy PV sources in a frame, the average number of slots successfully used by SUs in a frame is

$$N_s \tau_k (1 - \tau_k \frac{1}{N_{slot} - k})^{N_s - 1}. \tag{6}$$

From Equation (6), we obtain the throughput of SUs as follows:

$$T = \frac{\sum_{k=0}^{N_{slot}-k} c_k N_s \tau_k (1 - \tau_k \frac{1}{N_{slot}-k})^{N_s-1}}{N_{slot}}. \tag{7}$$

We define the packet loss probability P_L of SUs as the ratio of the average number of lost packets of SUs to the number of successfully transmitted or lost packets. Given that there are k busy PV sources in a frame, the average number of lost packets and successfully transmitted packets of a tagged SU in one frame is $\pi_{m,0,k} q_k$ and $\sum_{i=0}^{m} \pi_{i,0,k}(1 - q_k)$, respectively. By conditioning on the number of busy PV sources in one frame, we obtain the average number of lost packets and successfully transmitted packets in one frame. Therefore, the packet loss probability is given as follows.

$$P_L = \frac{\sum_{k=0}^{N_{slot}-k} c_k \pi_{m,0,k} q_k}{\sum_{k=0}^{N_{slot}-k} c_k [\sum_{i=0}^{m} \pi_{i,0,k}(1 - q_k) + \pi_{m,0,k} q_k]}. \tag{8}$$

6 Numerical Results

In this section, we evaluate the throughput and packet loss probability of secondary users in the proposed IEEE 802.16e/m cognitive radio system.

We consider the following parameters: the average length of talkspurt and silent periods are 352ms and 650 ms, respectively. The parameter values used in numerical examples are listed in Table 1.

Table 1 Parameter values used in numerical examples

Parameter	Value
N_{slot}	70
α	0.9857
β	0.9923
R_t	1

The theoretical results are obtained by the equations in Section 4. Our simulations are done by MATLAB program based on our protocol described in 3, and in each case, simulations are repeated 20 times with different seeds and 2000 frames in each time. In the following figures, solid lines stand for theoretical results and dotted lines stand for simulation results.

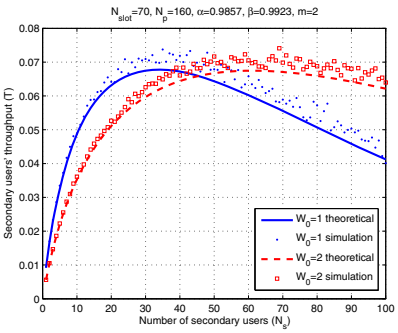


Fig. 1 Secondary users' throughput versus N_s ($N_p = 160$)

First notice that, the analytic results and simulation results match. Figure 1 describes the saturated throughput T versus the number of SUs when the initial contention window size W_0 is 1 and 2. It can be observed that if the number of secondary users is less than 45, the saturated throughput is larger with $W_0 = 1$ than that with $W_0 = 2$, while opposite if the number of secondary users is larger than 50. This is because larger initial contention window size is suitable for more SUs to contend.

Figure 2 and 3 depicts the packet loss probability of SUs versus the maximum backoff stage m and N_m, respectively, when the number of secondary users N_s is 20, 30 and 40. As expected, the packet loss probabilities decrease as m increases and as

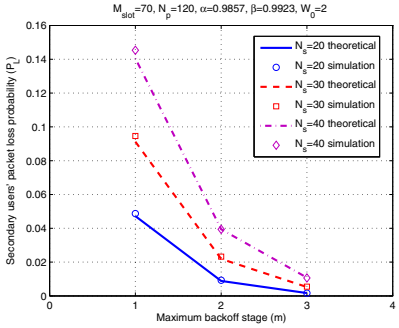


Fig. 2 Secondary users' packet loss probability versus m

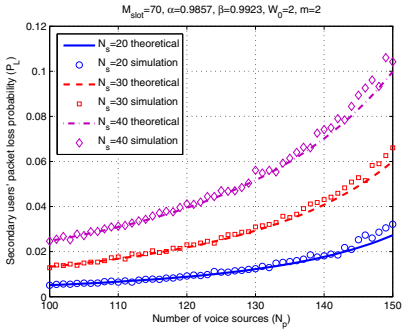


Fig. 3 Secondary users' packet loss probability versus N_p

N_p decreases. It is reasonable since larger m implies more opportunities to attempt to transmit and smaller N_p means averagely there are more idle slots in each frame for SUs to utilize.

References

1. Akyildiz, I.F., Lee, W.Y., Vuran, M.C., Mohanty, S.: Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey. Computer Networks **50**(13), 2127–2159 (2006)
2. IEEE std 802.16e-2006. IEEE standard for local and metropolitan area networks- part 16: Air interface for fixed and mobile broadband wireless access systems. amendment 2: Physical and medium access control layers for combined fixed and mobile operation in licensed bands and corrigendum (February 2006)
3. IEEE P802.16m/D3 (December 2009)
4. Fantacci, R., Tarchi, D.: A novel cognitive networking scenario for IEEE 802.16 networks. In: Proc. of IEEE GLOBECOM, Honolulu, Hawaii, USA (December 2009)
5. Chong, J.W., Sung, Y., Sung, D.K.: RawPEACH: Multiband CSMA/CA-Based Cognitive Radio Networks. J. Comm. Net. **11**(2) (April 2009)

6. Hwang, E., Kim, K.J., Lyakhov, A., Choi, B.D., Hwang, E., Kim, K.J., Lyakhov, A., Choo, B.D.: IEEE Trans. on Vehicular Technology **12**, 131–138 (2008)
7. Breuer, L., Baum, D.: An Introduction to Queueing Theory and Matrix-Analytic Methods. Springer, Berlin (2005)
8. Kemeny, J.G., Snell, J.L., Knapp, A.W.: Denumerable Markov Chains. Graduate Texts in Mathematics, 2nd edn., vol. 40. Springer, New York (1976)
9. Deng, S.: Traffic characteristics of packet voice. In: IEEE Int. Conf. Commun., vol. 3, pp. 1369–1374 (1995)

Part III
Queueing Models II

M/M/1/1 Retrial Queues with Setup Time

Tuan Phung-Duc

Abstract This paper considers single server retrial queues with setup time. In the basic model, if the server completes a service and there are no customers in the orbit, the server is turned off immediately. Arriving customers that see the server occupied join the orbit and repeat their attempt after some random time. The new feature of our models is that an arriving customer that sees the server off waits at the server and the server is turned on. The server needs some setup time to be active so as to serve the waiting customer. If the server completes a service and the orbit is not empty, it stays idle waiting for either a new customer or a customer from the orbit. For this model, we obtain explicit expressions for the generating functions of the joint queue length. We then consider an extended model where the server stays idle for a while before being turned off for which explicit solution is also obtained.

Keywords M/M/1/1 retrial queue · Setup time · Power-saving

1 Introduction

Power-saving in ICT systems is an important issue because ICT devices consume a large amount of energy. One simple method is to turn off an idle device and to switch it on again when some jobs arrive. This is because in the current technology idle devices still consume about 60% of their peak processing a job [2]. On the other hand, a quick response is crucial for delay sensitive applications. An off server needs some setup time in order to be active during which the server consumes energy but cannot process a job. Thus, there is a trade-off between power-consumption and delay performance. This trade-off can be analyzed using single server queueing models with setup times which are extensively studied in the literature [3, 12].

T. Phung-Duc (✉)
Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Ookayama, Meguro-ku, Tokyo, Japan
e-mail: tuan@is.titech.ac.jp

T.V. Do et al. (eds.), *Queueing Theory and Network Applications*,
Advances in Intelligent Systems and Computing 383,
DOI: 10.1007/978-3-319-22267-7_9

Retrial is a common phenomenon in ICT systems. Customers (jobs) that cannot occupy the server immediately upon arrival join an orbit and retry to enter the server after some random time. Although queues with retrial or setup time are separately investigated in the literature, this paper is the first attempt to combine these two features in one model. We first consider an M/M/1/1 retrial queue with setup time where the server is immediately turned off when the system (server and orbit) becomes empty. We then consider an extended model where the server waits for a while before being switched off. This idle time reduces the mean number of customers in the orbit and the mean waiting time but at the same time it increases the power consumption. Thus, there is a need for an appropriate setting of this idle time.

Our models are suitable for a downlink of a mobile station with a power saving mode. A mobile station receives data from a base station. Arriving messages are stored in the base station and the mobile station downloads these messages from the base station. Upon the completion of a download, if there are no messages in the base station the mobile station is turned off in order to save energy. However, when a message arrives, the base station sends a signal in order to wake up the mobile station. The mobile station needs some random setup time to be active so as to receive waiting messages.

A closely related work is due to Do [4] who considers an M/M/1/1 retrial queue with working vacation in which the server can still work at a different rate during the vacation period. In [4], the retrial rate is independent to the number of customers in the orbit. Artalejo [1] considers M/G/1/1 retrial queue with constant retrial rate and vacation. In contrast to the models in [1, 4], we consider the so-called classical retrial policy in which the retrial rate is proportional to the number of customers in the orbit. It should be noted that the classical retrial policy makes the underlying Markov chain non-homogeneous and thus its analysis is more challenging in comparison with the constant retrial rate policy. Multiserver queues with setup time and without retrials are analyzed in [8, 9, 10]. Analytical solutions for multiserver retrial queue and tandem retrial model could be found in [5, 6] and [7], respectively.

The rest of this paper is organized as follows. Section 2 presents the basic M/M/1/1 retrial queue with setup time and its analysis. Section 3 presents an extended model where the server stays idle for a while before being turned off and a summary of analytical results. Concluding remarks are presented in Section 5.

2 Model Without a Waiting Time

2.1 Model

We consider an M/M/1/1 retrial queue with setup time. Customers arrive at the server according to a Poisson process with rate λ. The service time of customers follows an exponentially distributed time with mean $1/\nu$. Customers that see the server busy upon arrival join the orbit and retry for service after some exponentially distributed time with mean $1/\mu$. When the system becomes empty, the server is turned off immediately. Customers that see the off server waits at the server and the server is

turned on. However, the server needs some setup time to be active so as to serve the waiting customer. We assume that the setup time is exponentially distributed with mean $1/\alpha$. Customers that see the server in setup state joins the orbit and behaves the same as other customers in the orbit.

Remark 1 Our model is different from other retrial models with vacations [1, 4] where arriving customers that see the server on vacation join the orbit. In our model, the setup time is activated upon an arrival of a new customer while the vacations in [1, 4] are independent of the arrivals.

2.2 Analysis

In this section, we present an analytical solution for the joint stationary distribution in terms of generating functions. Let $C(t)$ and $N(t)$ denote the state of the server and the number of customers in the orbit, respectively.

$$C(t) = \begin{cases} 0, & \text{the server is empty,} \\ 1, & \text{the server is busy,} \\ 2, & \text{the server is in setup process.} \end{cases}$$

It is easy to see that $\{X(t) = (C(t), N(t)); t \geq 0\}$ forms a Markov chain on the state space:

$$\mathcal{S} = \{(i, j); i = 0, 1, 2, j \in \mathbb{Z}_+\},$$

where $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$. We assume that the system is stable and thus $\lambda < \nu$.

We refer to Figure 1 for transitions among states. It should be noted that $(0, 0)$ represents the state where the server is turned off.

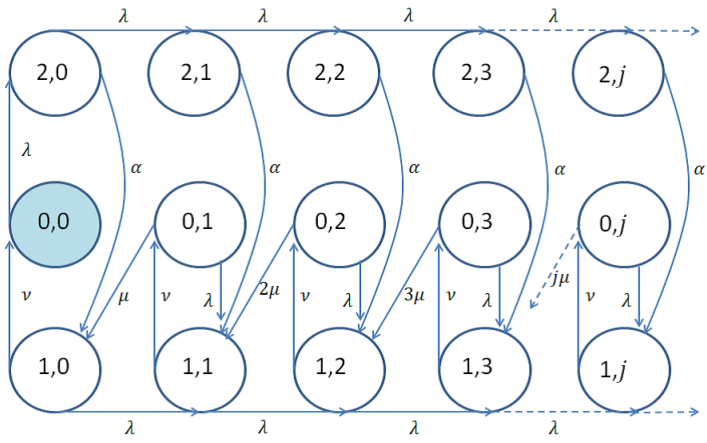


Fig. 1 Transitions among states

Let $\pi_{i,j} = \lim_{t\to\infty} \mathrm{P}(C(t) = i, N(t) = j)$ $((i, j) \in \mathcal{S})$ denote the joint stationary distribution of $\{X(t)\}$. In this section, we obtain explicit expressions for the generating functions of the joint stationary distribution $\pi_{i,j}$ $((i, j) \in \mathcal{S})$. We define the generating functions as follows.

$$\Pi_i(z) = \sum_{j=0}^{\infty} \pi_{i,j} z^j, \qquad i = 0, 1, 2.$$

The balance equations for states $(0, j)$ read as follows.

$$(\lambda + j\mu)\pi_{0,j} = \nu\pi_{1,j}, \qquad j \in \mathbb{Z}_+.$$

Multiplying this equation by z^j and summing up over $j \in \mathbb{Z}_+$, we obtain

$$\lambda\Pi_0(z) + \mu z\Pi_0'(z) = \nu\Pi_1(z). \tag{1}$$

Next, we consider balance equations for states $(1, j)$ $(j \in \mathbb{Z}_+)$. We have

$$(\lambda + \nu)\pi_{1,j} = \alpha\pi_{2,j} + \lambda\pi_{1,j-1} + (j + 1)\mu\pi_{0,j+1} + \lambda\pi_{0,j}(1 - \delta_{0,j}),$$

where $\pi_{1,-1} = 0$ and $\delta_{0,j}$ is the Kronecker delta, i.e. $\delta_{0,j} = 1$ if $j = 0$ and $\delta_{0,j} = 0$ if $j \neq 0$. Multiplying this equation by z^j and summing up over $j \in \mathbb{Z}_+$ yields

$$(\lambda + \nu)\Pi_1(z) = \alpha\Pi_2(z) + \lambda z\Pi_1(z) + \mu\Pi_0'(z) + \lambda(\Pi_0(z) - \pi_{0,0}). \tag{2}$$

Next, we consider balance equations for states $(2, j)$ $(j \in \mathbb{Z}_+)$.

$$(\lambda + \alpha)\pi_{2,0} = \lambda\pi_{0,0}, \qquad (\lambda + \alpha)\pi_{2,j} = \lambda\pi_{2,j-1}, \qquad j \geq 1.$$

Summing the first equation by z^0 and the second equation by z^j and summing over $j \in \mathbb{Z}_+$, we obtain

$$(\lambda + \alpha)\Pi_2(z) = \lambda z\Pi_2(z) + \lambda\pi_{0,0}, \tag{3}$$

leading to

$$\Pi_2(z) = \frac{\lambda\pi_{0,0}}{\lambda + \alpha - \lambda z}.$$

We also have the following equation by summing up (1), (2) and (3) and arranging the result.

$$\lambda(\Pi_1(z) + \Pi_2(z)) = \mu\Pi_0'(z). \tag{4}$$

It should be noted that (4) represents the balance between the flows in and out the orbit. Substituting $\Pi_1(z)$ and $\Pi_2(z)$ in terms of $\Pi_0(z)$ into (4), we obtain

$$\lambda\left(\frac{\lambda\Pi_0(z)+\mu z\Pi_0'(z)}{\nu}+\frac{\lambda\pi_{0,0}}{\lambda+\alpha-\lambda z}\right)=\mu\Pi_0'(z). \tag{5}$$

Arranging this equation we obtain

$$\Pi_0'(z)=\frac{\lambda^2}{\mu\nu}\frac{1}{1-\frac{\lambda z}{\nu}}\Pi_0(z)+\frac{\lambda^2}{\mu(\lambda+\alpha)}\frac{\pi_{0,0}}{(1-\frac{\lambda z}{\lambda+\alpha})(1-\frac{\lambda z}{\nu})}. \tag{6}$$

Remark 2 Taking the limit $\mu\to\infty$, (6) becomes $\Pi_0'(z)=0$ leading to $\Pi_0(z)=\pi_{0,0}$. As a result, our model reduces to the conventional M/M/1 queue with setup time (see e.g. Section 4.1 in [8]).

The differential equation (6) is solvable. First, we solve the homogeneous equation:

$$\Pi_0'(z)=\frac{\lambda^2}{\mu\nu}\frac{1}{1-\frac{\lambda z}{\nu}}\Pi_0(z).$$

The solution of this equation is given by

$$\Pi_0(z)=C_0\left(1-\frac{\lambda z}{\nu}\right)^{-\frac{\lambda}{\mu}},$$

for some constant C_0. This suggests us to find the solution for (6) of the form

$$\Pi_0(z)=C(z)\left(1-\frac{\lambda z}{\nu}\right)^{-\frac{\lambda}{\mu}}.$$

Substituting this function into (6), we obtain

$$C'(z)=\frac{\lambda^2}{\mu(\lambda+\alpha)}\frac{\pi_{0,0}}{(1-\frac{\lambda z}{\lambda+\alpha})}(1-\frac{\lambda z}{\nu})^{\frac{\lambda}{\mu}-1},$$

whose solution is given by

$$C(z)=C+\frac{\lambda^2\pi_{0,0}}{\mu(\lambda+\alpha)}\int_0^z\frac{(1-\frac{\lambda u}{\nu})^{\frac{\lambda}{\mu}-1}}{1-\frac{\lambda u}{\lambda+\alpha}}du,$$

where C is some constant. Because $\Pi_0(0)=\pi_{0,0}$, we have $C(0)=C=\pi_{0,0}$.

Thus, we have

$$\Pi_0(1)=\kappa_0\pi_{0,0}.$$

where

$$\kappa_0 = \left(1 - \frac{\lambda}{\nu}\right)^{-\frac{\lambda}{\mu}} \left(1 + \frac{\lambda^2}{\mu(\lambda+\alpha)} \int_0^1 \frac{(1-\lambda u/\nu)^{\frac{\lambda}{\mu}-1}}{1-\lambda u/(\lambda+\alpha)} du\right).$$

Furthermore, it follows from the differential equation (6) that

$$\Pi_0'(1) = \kappa_0' \pi_{0,0},$$

where

$$\kappa_0' = \frac{\lambda^2}{\mu(\nu-\lambda)} \left(\kappa_0 + \frac{\nu}{\alpha}\right).$$

It follows from (4) that

$$\Pi_1(1) + \Pi_2(1) = \frac{\mu}{\lambda} \kappa_0' \pi_{0,0}.$$

Furthermore, because $\Pi_0(1) + \Pi_1(1) + \Pi_2(1) = 1$, we have

$$\pi_{0,0} = \frac{1}{\kappa_0 + \frac{\mu}{\lambda}\kappa_0'}.$$

Differentiating equation (6) at $z = 1$ yields

$$\Pi_0''(1) = \kappa_0'' \pi_{0,0},$$

where

$$\kappa_0'' = \frac{\lambda}{\mu}\left(\frac{\rho^2\kappa_0}{(1-\rho)^2} + \frac{\rho\kappa_0'}{1-\rho} + \frac{\rho\lambda(\nu+\alpha-\lambda)}{(1-\rho)^2\alpha^2}\right), \qquad \rho = \frac{\lambda}{\nu}.$$

Thus, the mean number of customers in the system is given by

$$\mathrm{E}[N] = (\kappa_0' + \frac{\mu}{\lambda}\kappa_0'')\pi_{0,0}.$$

3 Model with an Idle Time

3.1 Model

In this section, we extend the model in Section 2 by adding a new feature. In particular, we assume that when the system becomes empty the server is not immediately turned off but stays idle for some random time. In this idle period, an arriving customer receives the service immediately. We assume that the idle time is exponentially distributed with mean $1/\beta$. Let $C(t)$ denote the state of the server (defined as in the previous section) and $N(t)$ denote the number of customers in the orbit. Let

$$X(t) = \begin{cases} O, & \text{the server is turned off,} \\ (C(t), N(t)), & \text{otherwise.} \end{cases}$$

It is easy to see that $\{X(t); t \geq 0\}$ forms a Markov chain on the state space $\mathcal{S}$ given by

$$\mathcal{S} = O \cup \{0, 1, 2\} \times \mathbb{Z}_+.$$

We assume that $\lambda < \nu$ and thus the Markov chain is stable. Furthermore, we are going to find the stationary distribution defined as follows.

$$\pi_0 = \lim_{t\to\infty} \mathrm{P}(X(t) = O), \qquad \pi_{i,j} = \lim_{t\to\infty} \mathrm{P}(X(t) = (i, j)).$$

We refer to Figure 2 for transitions among states. The generating functions $\Pi_i(z)$ $(i = 0, 1, 2)$ are defined the same as in the previous section.

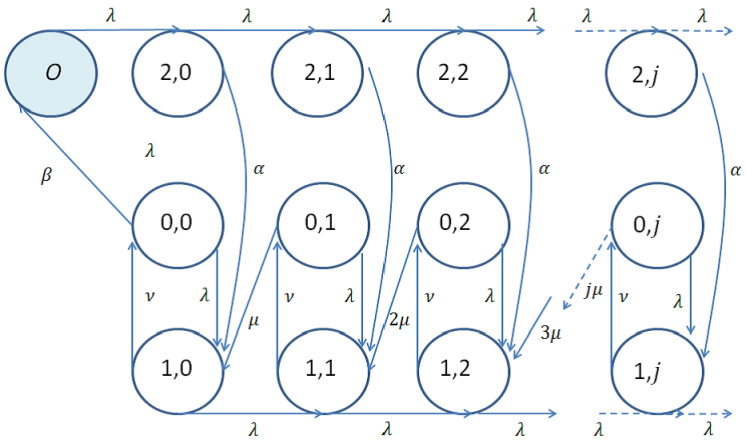


Fig. 2 Transition among states

3.2 Analysis

We have the following balance equations for states $(0, j)$ $(j \in \mathbb{Z}_+)$

$$\pi_{0,0}\beta = \lambda\pi_0, \tag{7}$$

$$(\lambda + \beta)\pi_{0,0} = \nu\pi_{1,0}, \qquad j = 0, \tag{8}$$

$$(\lambda + j\mu)\pi_{0,j} = \nu\pi_{1,j}, \qquad j \geq 1. \tag{9}$$

Multiplying (8) by z^0 and (9) by z^j and summing up over $j \in \mathbb{Z}_+$ we obtain

$$\beta\pi_{0,0} + \lambda\Pi_0(z) + \mu z\Pi_0'(z) = \nu\Pi_1(z). \tag{10}$$

Next we consider balance equations for states $(1, j)$ $(j \in \mathbb{Z}_+)$.

$$(\lambda + \nu)\pi_{1,j} = \lambda\pi_{1,j-1} + (j+1)\mu\pi_{0,j+1} + \alpha\pi_{2,j} + \lambda\pi_{0,j}.$$

where $\pi_{1,-1} = 0$. Multiplying this equation by z^j and summing up over $j \in \mathbb{Z}_+$, we obtain

$$(\lambda + \nu)\Pi_1(z) = \lambda z\Pi_1(z) + \mu\Pi_0'(z) + \alpha\Pi_2(z) + \lambda\Pi_0(z). \tag{11}$$

Finally, we consider balance equations for states $(2, j)$ $(j \in \mathbb{Z}_+)$.

$$(\lambda + \alpha)\pi_{2,0} = \lambda\pi_0, \qquad j = 0, \tag{12}$$

$$(\lambda + \alpha)\pi_{2,j} = \lambda\pi_{2,j-1}, \qquad j \geq 1. \tag{13}$$

Multiplying the first equation by z^0 and the second equation by z^j and summing up over $j \in \mathbb{Z}_+$, we obtain

$$(\lambda + \alpha)\Pi_2(z) - \lambda z\Pi_2(z) = \lambda\pi_0 \Leftrightarrow \Pi_2(z) = \frac{\lambda\pi_0}{\lambda + \alpha - \lambda z}. \tag{14}$$

As in Section 2, we also have the following equation (representing the balance between the flows in and out the orbit)

$$\lambda(\Pi_1(z) + \Pi_2(z)) = \mu\Pi_0'(z), \tag{15}$$

by summing up (10), (11) and (14) and arranging the result. Substituting $\Pi_1(z)$ and $\Pi_2(z)$ in terms of $\Pi_0(z)$ into the above equation and arranging the result yields

$$\Pi_0'(z) = \frac{\lambda^2}{\mu\nu}\frac{1}{1 - \frac{\lambda}{\nu}z}\Pi_0(z) + \frac{\lambda^2\pi_0(\lambda + \alpha + \nu - \lambda z)}{\mu\nu(\lambda + \alpha - \lambda z)(1 - \frac{\lambda z}{\nu})}. \tag{16}$$

It should be noted that (16) becomes $\Pi_0'(z) = 0$ as $\mu \to \infty$, i.e. $\Pi_0'(z) = \pi_{0,0}$. In this case our model reduces to the model with idle period and setup time (without retrial). The solution of (16) is given by

$$\Pi_0(z) = \pi_0\left(1 - \frac{\lambda z}{\nu}\right)^{-\frac{\lambda}{\mu}}\left(\frac{\lambda}{\beta} + \frac{\lambda^2}{\mu\nu}\int_0^z \left(1 - \frac{\lambda u}{\nu}\right)^{\frac{\lambda}{\mu}-1}\left(1 + \frac{\nu}{\lambda + \alpha - \lambda u}\right)du\right).$$

Thus, we have $\Pi_0(1) = \chi_0\pi_0$, where

$$\chi_0 = \left(1 - \frac{\lambda}{\nu}\right)^{-\frac{\lambda}{\mu}}\left(\frac{\lambda}{\beta} + \frac{\lambda^2}{\mu\nu}\int_0^1 \left(1 - \frac{\lambda u}{\nu}\right)^{\frac{\lambda}{\mu}-1}\left(1 + \frac{\nu}{\lambda + \alpha - \lambda u}\right)du\right).$$

Furthermore, it follows from the differential equation that

$$\begin{aligned}\Pi_0'(1) &= \frac{\lambda^2 \Pi_0(1)}{\mu(\nu-\lambda)} + \frac{\pi_0\lambda^2(\alpha+\nu)}{\mu\alpha(\nu-\lambda)} \\ &= \chi_0'\pi_0,\end{aligned}$$

where

$$\chi_0' = \frac{\lambda^2\chi_0}{\mu(\nu-\lambda)} + \frac{\lambda^2(\alpha+\nu)}{\mu\alpha(\nu-\lambda)}.$$

This expression together with the balance equation between the flow in and out the orbit (15) yield

$$\Pi_1(1) + \Pi_2(1) = \frac{\mu}{\lambda}\chi_0'\pi_0.$$

Because

$$\Pi_0(1) + \Pi_1(1) + \Pi_2(1) + \pi_0 = 1,$$

we have

$$\pi_0 = \frac{1}{1 + \chi_0 + \frac{\mu\chi_0'}{\lambda}}.$$

Thus, we also have explicit expressions for $\Pi_i(z)$ $(i = 0, 1, 2)$.

Differentiating equation (16) at $z = 1$ yields,

$$\Pi_0''(1) = \pi_0\chi_0'',$$

where

$$\chi_0'' = \frac{\lambda}{\mu}\left(\frac{\rho^2\chi_0}{(1-\rho)^2} + \frac{\rho\chi_0'}{1-\rho} + \frac{\rho\lambda(\nu+\alpha-\lambda)}{(1-\rho)^2\alpha^2} + \frac{\rho^2}{(1-\rho)^2}\right).$$

Thus, the mean number of customers in the system is given by

$$\mathrm{E}[N] = (\chi_0' + \frac{\mu}{\lambda}\chi_0'')\pi_0.$$

4 Performance Measures and Numerical Results

We consider two main performance measures: the probability that the server is off ($\pi_{0,0}$ in the model in Section 2 and π_0 in the model in Section 3) and the mean number of customers in the orbit. We would like to increase the former (i.e. decrease the probability of the states on which the server consumes power) in order to save energy while we also would like to decrease the mean number of customers in the orbit. Thus, we have a trade-off between the performance and power consumption. In order to see this trade-off we consider a cost function which is the product of the probability that the server is in either SETUP or ON or IDLE (not in OFF state)

and the mean number of customers in the orbit, i.e., $(1-\pi_{0,0})\mathrm{E}[N]$ in the model in Section 2 and $(1-\pi_0)\mathrm{E}[N]$ in the model in Section 3. It should be noted that the server consumes power in SETUP and ON and IDLE states.

In this section, we present some numerical results. We fix the parameters as follows: $\mu = 1$ and $\nu = 1$. We consider three cases where $\beta = 0.1, 1$ and 10 for the model with a waiting time (exponentially distributed with mean $1/\beta$). We first consider the case where $\rho = \lambda/\nu = 0.7$. Figure 3 shows the probability that the server is in OFF state against the setup rate. We observe that the π_0 increases with β in the model with waiting time. This is because a large β results in a short mean idle time $1/\beta$ and thus a large π_0. We also observe that $\pi_0 < \pi_{0,0}$ which is also intuitive due to the same reason as in the monotonicity of π_0 in β.

Furthermore, we observe from Figure 8 that the mean number of customers in the orbit $\mathrm{E}[N]$ decreases with β. This is intuitive because the server has more chance to be in the idle state during which it can serve an arriving customer immediately when β is small. We also observe that $\mathrm{E}[N]$ for the model with a waiting time is bounded by that for the model without a waiting time.

Finally, we consider the cost function against the setup rate α. We observe that when α is small, the cost function increases with β. This suggests that if the setup time is long, it is better to keep the idle time long. However, when the setup rate α is large enough, we observe the cost function decreases with β. This implies that if the setup is fast enough, it is better to keep only a short idle time so as to save power consumption.

Figures 4, 7 and 6 show the probability of OFF state, $\mathrm{E}[N]$ and the cost function for the case of $\rho = 0.1$. We observe the same trends as for the case of $\rho = 0.7$. Furthermore, the range of α at which the cost function of the model without waiting time outperforms that of the model with a waiting time is larger for the case of $\rho = 0.1$ in comparison with the case $\rho = 0.7$. This suggest that when the utilization is low and the setup time is large enough, it is better to switched off as soon as the server becomes idle.

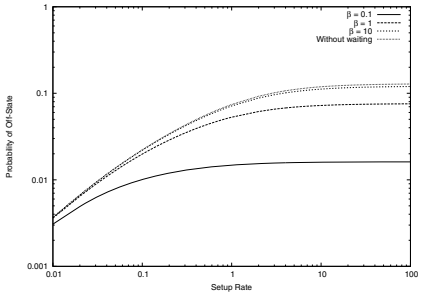


Fig. 3 Probability of OFF state against α ($\rho = 0.7$)

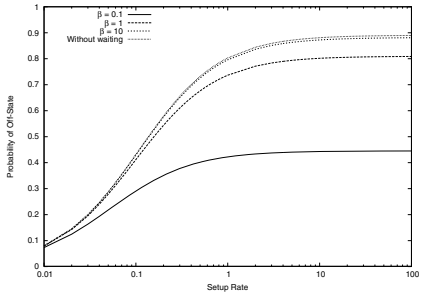


Fig. 4 Probability of OFF state against α ($\rho = 0.1$)

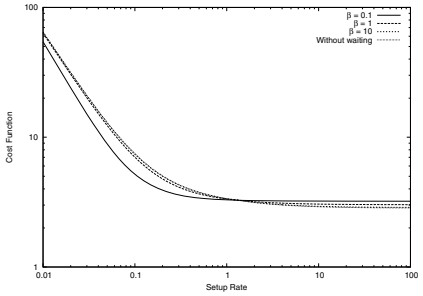


Fig. 5 Cost function againstα ($\rho = 0.7$)

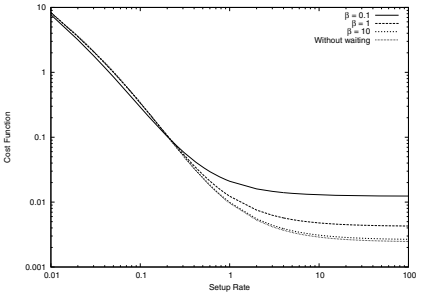


Fig. 6 Cost function againstα ($\rho = 0.1$)

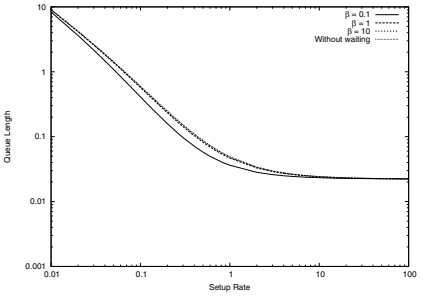


Fig. 7 Mean number of jobs in orbit against α ($\rho = 0.1$)

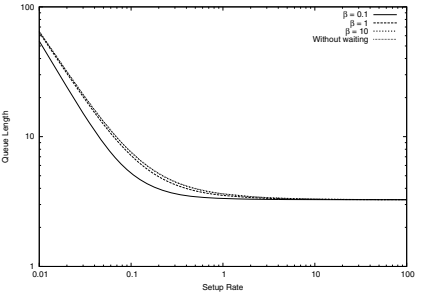


Fig. 8 Mean number of jobs in orbit against α ($\rho = 0.7$)

5 Concluding Remark

In this paper, we have proposed two retrial queueing models with setup time. In the first model, the server is immediately turned off when the system becomes empty while in the second model, the server stays idle for a while before being switched off. We have derived explicit expressions for the partial generating functions of the joint stationary probability of the state of the server and the number of customers in the orbit. From the generating function, we have obtained the mean number of customers in the orbit in an explicit form. We have demonstrated some numerical examples to show the effects of parameters on some performance measures. Models with general distributions for service time and setup time are left for future studies. Extension of the current model to the model with N-policy may be also another interesting topic.

Acknowledgments Tuan Phung-Duc was supported in part by Japan Society for the Promotion of Science, JSPS Grant-in-Aid for Young Scientists (B), Grant Number 2673001. The author would like to thank the anonymous referees for constructive comments which improve the presentation of the paper.

References

1. Artalejo, J.R.: Analysis of an M/G/1 queue with constant repeated attempts and server vacations. Computers and Operations Research **24**(6), 493–504 (1997)
2. Barroso, L.A., Holzle, U.: The case for energy-proportional computing. Computer **40**(12), 33–37 (2007)
3. Bischof, W.: Analysis of M/G/1-queues with setup times and vacations under six different service disciplines. Queueing Systems: Theory and Applications **39**(4), 265–301 (2001)
4. Do, T.V.: M/M/1 retrial queue with working vacations. Acta Informatica **47**(1), 67–75 (2010)
5. Phung-Duc, T., Masuyama, H., Kasahara, S., Takahashi, Y.: M/M/3/3 and M/M/4/4 retrial queues. Journal of Industrial and Management Optimization **5**(3), 431 (2009)
6. Phung-Duc, T., Masuyama, H., Kasahara, S., Takahashi, Y.: State-dependent M/M/c/c+ r retrial queues with Bernoulli abandonment. Journal of Industrial and Management Optimization **6**(3), 517–540 (2010)
7. Phung-Duc, T.: An explicit solution for a tandem queue with retrials and losses. Operational Research **12**(2), 189–207 (2012)
8. Phung-Duc, T.: Impatient customers in power-saving data centers. In: Sericola, B., Telek, M., Horváth, G. (eds.) ASMTA 2014. LNCS, vol. 8499, pp. 185–199. Springer, Heidelberg (2014)
9. Phung-Duc, T.: Server farms with batch arrival and staggered setup. In: Proceedings of the Fifth Symposium on Information and Communication Technology, pp. 240–247. ACM (2014)
10. Phung-Duc, T.: Exact solution for M/M/*c*/Setup queue (2014). Preprint: http://arxiv.org/abs/1406.3084
11. Phung-Duc, T.: Multiserver queues with finite capacity and setup time. In: Remke, A., Manini, D., Gribaudo, M. (eds.) ASMTA 2015. LNCS, vol. 9081, pp. 173–187. Springer, Heidelberg (2015)
12. Takagi, H.: Priority queues with setup times. Operations Research **38**(4), 667–677 (1990)

The Pseudo-fault Geo/Geo/1 Queue with Setup Time and Multiple Working Vacation

Zhanyou Ma, Pengcheng Wang and Wuyi Yue

Abstract In this paper, we consider a discrete time Geo/Geo/1 repairable queueing system with pseudo-fault, setup time, N-policy and multiple working vacations. We assume that the service interruption is caused by pseudo-fault and breakdown, and occurs only when the server is busy. Using quasi birth-and-death chain, we establish a two-dimensional Markov chain. We obtain the distribution of the steady-state queue length by using matrix-geometric solution method. Moreover, We provide several performance indices of the system in steady-state. Finally, we present numerical results to illustrate the effect of several parameters on the systems.

Keywords Pseudo-fault · Setup time · Multiple working vacation · Matrix-geometric solution

1 Introduction

The theory of discrete time queueing system has been well investigated and applied in a variety of directions, such as computer system, computer communications and manufacturing systems. Domestic and foreign scholars are interested in the discrete time queueing system which is first proposed in [1]. Ndreca et al. considered a GI/Geo/1 queueing model with priority and derived the distributions of the mean queue length and the mean waiting time in [2].

The working vacation policy was introduced into the queueing system firstly in [3]. Subsequently, working vacation policy was extended to discrete time queueing system models in [4]. Furthermore, N-policy and setup time were also studied in the

Z. Ma · P. Wang
College of Science, Yanshan University, Qinhuangdao 066004, China
e-mail: mzhy55@ysu.edu.cn, kdwpc@126.com

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

T.V. Do et al. (eds.), *Queueing Theory and Network Applications*,
Advances in Intelligent Systems and Computing 383,
DOI: 10.1007/978-3-319-22267-7_10

considered system. Yadin introduced N-policy into the queueing system first in [5]. A discrete-time queueing model with batch service and multiple working vacations was studied in [6].

Actually, there is no breakdown in classical queueing system. But mechanical failure and service interruption often occur while making a production or serving a customer. Failure policy was first introduced into the queueing model in [7]. Subsequently, Kulkarni et al. extended the queueing model with failure in [8]. Kalidass et al. analyzed an M/M/1 queueing system with unreliable server and derived steady-state distribution and analytical expressions for the transient state probabilities of the system length in [9].

The definition of system false fault state was presented in power systems, and the calculating formulas of false fault indices were given in [10]. The reliability evaluation algorithm was also proposed in order to improve the system performance.

However, the practical problems such as Internet, digital communications and industrial manufacturing are usually full of complexity, improving the system performance or increasing the stability is more important work. For this, in this paper, we first introduce the concept of pseudo-fault into the considered queueing model to provide theoretical basis for the analysis of network performance. Then, we present a discrete time Geo/Geo/1 repairable queueing system with setup time, N-policy, pseudo failures and multiple working vacations. Finally, we offer numerical results to illustrate the parameter effect on the performance measures.

2 The Mathematical Model

In this paper, $\forall x \in [0, 1]$, and let $\bar{x}$ be $1 - x$. We first describe a discrete time Geo/Geo/1 repairable queueing system with pseudo-fault, setup time, N-policy and multiple working vacations (MWV) as follows:

(1) A customer arrives at the system in (n^-, n). The inter-arrival time is an independently identically distributed (i.i.d.) sequence, which follows a geometric distribution with parameter p $(0 < p < 1)$, namely,

$$P\{T = j\} = p\bar{p}^{j-1}, \quad j = 1, 2, ...$$

(2) The starting and ending of service occur at epoch n. The service time S_b follows a geometric distribution with parameter μ_b $(0 < \mu_b < 1)$ in a regular busy period. The service time S_v follows a geometric distribution with parameter μ_v $(0 \le \mu_v < 1,\ \mu_v < \mu_b)$ in a working vacation period, namely,

$$P\{S_b = j\} = \mu_b\bar{\mu}_b^{j-1},\ j = 1, 2, ...,$$
$$P\{S_v = j\} = \mu_v\bar{\mu}_v^{j-1},\ j = 1, 2, ...$$

(3) The starting and ending of the vacation occur in (n, n^+), the vacation time V follows a geometric distribution with parameter θ $(0 < \theta < 1)$. The beginning

and ending of setup time occur in (n, n^+), the setup time U follows a geometric distribution with parameter β $(0 < \beta < 1)$, namely,

$$P\{V = j\} = \theta\bar{\theta}^{j-1},\ j = 1, 2, ...,$$
$$P\{U = j\} = \beta\bar{\beta}^{j-1},\ j = 1, 2, ...$$

(4) When the system becomes empty, the server will start a vacation. If a customer arrives in the system during a vacation period, it will be served with the service rate μ_v. When a working vacation ends and the number of the customers is less than N (a fixed positive integer) in the queue, another vacation is taken; otherwise, the server starts a setup period. At this time, if the server launches successfully, the service rate will be changed from μ_v to μ_b along with vacation end and a regular busy period beginning; if the server can not start, the server will continue to attempt to start until the system begins a regular busy period.

We assume that the server cannot serve customers in a setup period. During a regular busy period, if service interruptions do not appear in the queue, the server will still work with service rate μ_b continuously; if the service is interrupted, the queueing system will follow the specifical principle as the following (5).

(5) Service interruption occurs only in a regular busy period with probability q $(0 \leq q \leq 1,)$ in (n^-, n). Subsequently, breakdown and pseudo-fault occur with probabilities α $(0 \leq \alpha \leq 1)$ and $\bar{\alpha}$, respectively. If the pseudo-fault occurs, a vacation period is taken. If the breakdown appears, the repair period starts immediately.

The repair time R follows a geometric distribution with parameter γ $(0 < \gamma < 1)$. Suppose that the repair period begins in (n^-, n) and the repair period ends in (n, n^+). After repair period, the server is assumed as good as new and the server will continue to serve customers. The served time in this case is still valid.

(6) We assume that inter-arrival times, the probability of service interruptions, repair times, service times, setup time and vacation times are mutually independent. And the service discipline is first in first out (FIFO).

The queueing system model which we considered in this paper can be illustrated in Figure 1.

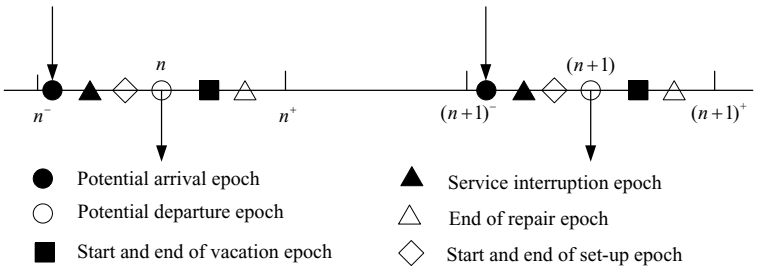


Fig. 1 The schematic diagram for the model description

3 Analysis

3.1 State Transition Probability Matrix

Let L_n^+ be the number of customers in the system at time n^+ and J_n be the server state at time n^+. Define that

$$J_n = \begin{cases} 0, \text{ the system is in a working vacation or pseudo-fault period at time } n^+ \\ 1, \text{ the system is in a setup period at time } n^+ \\ 2, \text{ the system is in a regular busy period at time } n^+ \\ 3, \text{ the system is in a breakdown period at time } n^+. \end{cases}$$

$\left\{\left(L_n^+,\ J_n\right),\ n \geq 0\right\}$ is a discrete time Markov chain in this system and its state space is given by

$$\Omega = \{(0,0)\} \cup \{(i,1),\ i \geq N\} \cup \{(i,j),\ i \geq 1,\ j = 0,2,3\}.$$

Using the lexicographical sequence for the states, the one-step state transition probability matrix of Markov chain can be written as follows:

$$\boldsymbol{P} = \begin{array}{c} 0 \\ 1 \\ 2 \\ \vdots \\ N-2 \\ N-1 \\ N \\ N+1 \\ \vdots \end{array} \begin{pmatrix} A_{00} & \boldsymbol{C}_{01} & & & & & & & \\ \boldsymbol{B}_{10} & \boldsymbol{A}_0 & \boldsymbol{C}_0 & & & & & & \\ & \boldsymbol{B}_0 & \boldsymbol{A}_0 & \boldsymbol{C}_0 & & & \boldsymbol{0} & & \\ & & \ddots & \ddots & \ddots & & & & \\ & & & \boldsymbol{B}_0 & \boldsymbol{A}_0 & \boldsymbol{C}_0 & & & \\ & & & & \boldsymbol{B}_0 & \boldsymbol{A}_0 & \boldsymbol{C}_1 & & \\ & & \boldsymbol{0} & & & \boldsymbol{B}_1 & \boldsymbol{A} & \boldsymbol{C} & \\ & & & & & & \boldsymbol{B} & \boldsymbol{A} & \boldsymbol{C} \\ & & & & & & & \ddots & \ddots & \ddots \end{pmatrix} \tag{1}$$

where $A_{00} = \bar{p}$, $\boldsymbol{C}_{01} = (p,0,0)$, $\boldsymbol{B}_{10} = (\bar{p}\mu_v, \bar{p}\mu_b, 0)^{\mathrm{T}}$, T denotes matrix transpose, and

$$\boldsymbol{A}_0 = \begin{pmatrix} \bar{p}\bar{\mu}_v + p\mu_v & 0 & 0 \\ (\bar{p}\bar{\mu}_b + p\mu_b)q\bar{\alpha} & (\bar{p}\bar{\mu}_b + p\mu_b)\bar{q} & (\bar{p}\bar{\mu}_b + p\mu_b)q\alpha \\ 0 & \bar{p}\gamma & \bar{p}\bar{\gamma} \end{pmatrix},$$

$$\boldsymbol{B}_0 = \begin{pmatrix} \bar{p}\mu_v & 0 & 0 \\ \bar{p}\mu_b q\bar{\alpha} & \bar{p}\mu_b\bar{q} & \bar{p}\mu_b q\alpha \\ 0 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{C}_0 = \begin{pmatrix} p\bar{\mu}_v & 0 & 0 \\ p\bar{\mu}_b q\bar{\alpha} & p\bar{\mu}_b\bar{q} & p\bar{\mu}_b q\alpha \\ 0 & p\gamma & p\bar{\gamma} \end{pmatrix},$$

$$\boldsymbol{B}_1 = \begin{pmatrix} \bar{p}\mu_v & 0 & 0 \\ 0 & 0 & 0 \\ \bar{p}\mu_b q\bar{\alpha} & \bar{p}\mu_b\bar{q} & \bar{p}\mu_b q\alpha \\ 0 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{C}_1 = \begin{pmatrix} p\bar{\mu}_v\bar{\theta} & p\bar{\mu}_v\theta & 0 & 0 \\ p\bar{\mu}_b q\bar{\alpha} & 0 & p\bar{\mu}_b\bar{q} & p\bar{\mu}_b q\alpha \\ 0 & 0 & p\gamma & p\bar{\gamma} \end{pmatrix},$$

$$\boldsymbol{A}=\begin{pmatrix} (\bar{p}\bar{\mu}_v+p\mu_v)\bar{\theta} & (\bar{p}\bar{\mu}_v+p\mu_v)\theta & 0 & 0 \\ 0 & \bar{p}\bar{\beta} & \bar{p}\beta & 0 \\ (\bar{p}\bar{\mu}_b+p\mu_b)q\bar{\alpha} & 0 & (\bar{p}\bar{\mu}_b+p\mu_b)\bar{q} & (\bar{p}\bar{\mu}_b+p\mu_b)q\alpha \\ 0 & 0 & \bar{p}\gamma & \bar{p}\bar{\gamma} \end{pmatrix},$$

$$\boldsymbol{B}=\begin{pmatrix} \bar{p}\mu_v\bar{\theta} & \bar{p}\mu_v\theta & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \bar{p}\mu_b q\bar{\alpha} & 0 & \bar{p}\mu_b\bar{q} & \bar{p}\mu_b q\alpha \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{C}=\begin{pmatrix} p\bar{\mu}_v\bar{\theta} & p\bar{\mu}_v\theta & 0 & 0 \\ 0 & p\bar{\beta} & p\beta & 0 \\ p\bar{\mu}_b q\bar{\alpha} & 0 & p\bar{\mu}_b\bar{q} & p\bar{\mu}_b q\alpha \\ 0 & 0 & p\gamma & p\bar{\gamma} \end{pmatrix}.$$

3.2 The Steady-State Analysis

If Markov chain $\{(L_n^+,\ J_n),\ n \geq 0\}$ is positive recurrent, let $(L,\ J)$ be the limit of the stationary distribution of $(L_n^+,\ J_n)$ and its distribution is given as follows:

$$\boldsymbol{\Pi} = (\pi_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, ...)$$

where $\pi_0 = \pi_{00},\ \boldsymbol{\pi}_i = (\pi_{i0}, \pi_{i2}, \pi_{i3}),\ 1 \leq i < N,\ \boldsymbol{\pi}_i = (\pi_{i0}, \pi_{i1}, \pi_{i2}, \pi_{i3}),\ i \geq N$, and

$$\pi_{ij} = P\{L = i,\ J = j\} = \lim_{n\to\infty} P\left\{L_n^+ = i, J_n = j\right\},\ (i, j) \in \Omega.$$

Theorem 1. *A necessary and sufficient condition for the Markov chain* $\{(L_n^+,\ J_n),\ n \geq 0\}$ *to be positive recurrent is that the matrix quadratic equation:*

$$\boldsymbol{R}^2\boldsymbol{B} + \boldsymbol{R}\boldsymbol{A} + \boldsymbol{C} = \boldsymbol{R} \tag{2}$$

has a minimal non-negative solution $\boldsymbol{R}$ *and the spectral radius* $\mathrm{SP}(\boldsymbol{R}) < 1$. *Also the* $3N + 2$ *dimensional stochastic matrix*

$$B[\boldsymbol{R}] = \begin{matrix} 0 \\ 1 \\ 2 \\ \vdots \\ N-1 \\ N \end{matrix} \begin{pmatrix} A_{00} & \boldsymbol{C}_{01} & & & & \\ \boldsymbol{B}_{10} & \boldsymbol{A}_0 & \boldsymbol{C}_0 & & & \\ & \boldsymbol{B}_0 & \boldsymbol{A}_0 & \boldsymbol{C}_0 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \boldsymbol{B}_0 & \boldsymbol{A}_0 & \boldsymbol{C}_1 \\ & & & & \boldsymbol{B}_1 & \boldsymbol{A}+\boldsymbol{R}\boldsymbol{B} \end{pmatrix} \tag{3}$$

has a left-invariant vector. When Markov chain is positive recurrent, its stationary distribution satisfies:

$$\begin{cases} \boldsymbol{\pi}_i = \boldsymbol{\pi}_N \boldsymbol{R}^{i-N}, \quad i \geq N+1 \\ (\pi_0, \boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_N) = (\pi_0, \boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_N)\, B\,[\boldsymbol{R}] \\ \pi_0 + \sum\limits_{k=1}^{N-1} \boldsymbol{\pi}_k \boldsymbol{e}_1 + \boldsymbol{\pi}_N (\boldsymbol{I} - \boldsymbol{R})^{-1} \boldsymbol{e}_2 = 1 \end{cases}$$

where $\boldsymbol{e}_1 = (1, 1, 1)^{\mathrm{T}}$, $\boldsymbol{e}_2 = (1, 1, 1, 1)^{\mathrm{T}}$.

The proof of Theorem 1 can be obtained by using equilibrium equation $\boldsymbol{\Pi P} = \boldsymbol{\Pi}$ and matrix-geometric solution method presented in [11, 12, 13].

3.3 Approximation Algorithm of Rate Matrix $\boldsymbol{R}$

Because the matrixes $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ presented in this paper are relatively complex, and therefore getting analytic expression of the rate matrix $\boldsymbol{R}$ directly becomes more difficult, we derive the recursion expression of the rate matrix $\boldsymbol{R}$ and calculate the numerical solution by using MATLAB as a method we usually use. From Eq. (2), the recursion expression is derived as follows:

$$\boldsymbol{R}_{n+1} = (\boldsymbol{R}_n^2 \boldsymbol{B} + \boldsymbol{C})(\boldsymbol{I} - \boldsymbol{A})^{-1}, \quad n = 0, 1, ... \tag{4}$$

The numerical solution of rate matrix $\boldsymbol{R}$ is the output of $\boldsymbol{R}_{n+1}$.

4 Performance Measures

In this section, we obtain a serial of main performance measures of the considered queueing system in this paper. Queue indices and fault analysis indices are given as follows:

(1) The expected waiting queue length is given by

$$E(L_q) = \sum_{i=1}^{\infty} \sum_{j=0}^{3} (i-1)\pi_{ij}. \tag{5}$$

(2) The breakdown probability is indicated by P_{q1}, namely

$$P_{q1} = P\{L \geq 1, j = 3\} = \sum_{i=1}^{\infty} \pi_{i3}. \tag{6}$$

(3) The pseudo-fault probability is denoted by P_{q2}, namely

$$P_{q2} = \bar{\alpha} P_{q1}/\alpha = (1/\alpha - 1) \sum_{i=1}^{\infty} \pi_{i3}. \tag{7}$$

5 Numerical Results

In this section, we provide numerical results to describe the effect of parameters on performance measures. Specifically, Figures 2 and 3 show the curves of the different parameters and system indices by taking $q = 0.4$, $\gamma = 0.6$, $\theta = 0.8$ as an example.

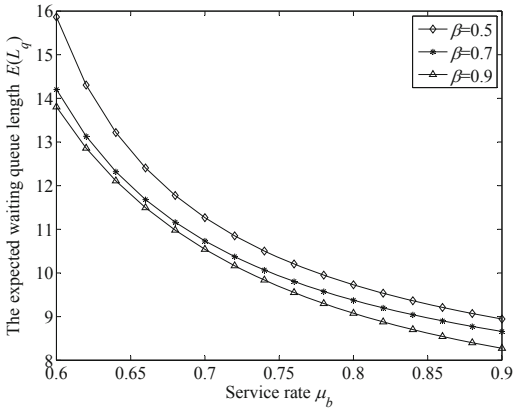


Fig. 2 Relation of $E(L_q)$ with μ_b

Figure 2 illustrates the effect of the waiting queue length $E(L_q)$ with μ_b and β when $p = 0.3$, $\mu_v = 0.2$, $\alpha = 0.7$, $N = 15$. If β is a constant, $E(L_q)$ decreases with the increasing value of μ_b. This is mainly because the service time reduces with the increase of the service rate μ_b, $E(L_q)$ turns small. When μ_b is fixed, $E(L_q)$ decreases with the increase of β. That is mainly because the larger the probability β is, the more the busy period becomes, therefore, $E(L_q)$ decreases.

Figure 3 shows the change of the probability of pseudo-fault period P_{q2} with the arrival rate p and breakdown rate α when $\mu_v = 0.2$, $\mu_b = \beta = 0.7$, $N = 15$. When α is fixed, P_{q2} increases with the increasing value of p. This is mainly because the probability of the regular busy period turns large with the increase of arrival rate p, and then the probability of pseudo-fault period P_{q2} increases. When p is fixed, P_{q2} decreases with the increasing value of α.

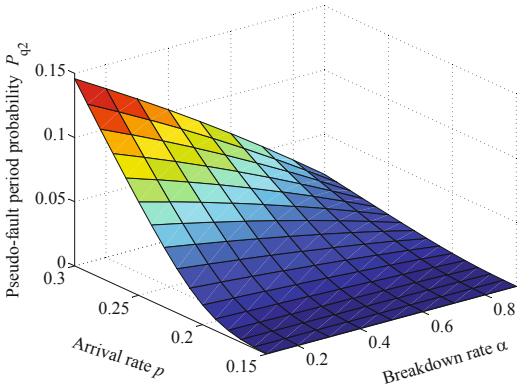


Fig. 3 Relation of P_{q2} with p and α

6 Conclusions

According to the analysis in this paper, we can draw a convincible conclusion: pseudo-fault really exists in the queue and it cannot be ignored in studying practical problem. Furthermore, the size of the impact of pseudo-fault and breakdown on system depends on the relative parameters such as μ_v, p and α, and so on. Therefore, we should take the corresponding solution for solving service interruption which is caused by different factors. For example, when the pseudo fault is dominant in service interruption, we should relax restrictions of the system or replace more advanced equipment. When the breakdown is prior, we should improve repair rate. As a result, how to cope with service interruption which is caused by different factors will play an important role in many aspects.

Acknowledgement This work was supported in part by the National Natural Science Foundation of China (No. 61472342), Hebei Province National Science Foundation (No. A2014203096), and was supported in part by MEXT, Japan.

References

1. Meisling, T.: Discrete-time Queuing Theory. Operations Research **6**(1), 96–105 (1958)
2. Ndreca, S., Scoppola, B.: Discrete Time GI/Geom/1 Queueing System with Priority. European Journal of Operational Research **189**(3), 1403–1408 (2008)
3. Servi, L., Finn, S.: M/M/1 Queues with Working Vacations (M/M/1/WV). Performance Evaluation **50**(1), 41–52 (2002)
4. Baba, Y.: Analysis of a GI/M/1 Queue with Multiple Working Vacations. Operations Research Letters **33**(2), 201–209 (2005)
5. Yadin, M., Naor, P.: Queueing Systems with a Removable Service Station. Operations Research **14**, 393–405 (1963)
6. Goswami, V., Mund, G.: Analysis of Discrete-time Batch Service Renewal Input Queue with Multiple Working Vacations. Computers & Industrial Engineering **61**(3), 629–636 (2011)
7. Avi-Itzhak, B., Naor, P.: Some Queuing Problems with the Service Station Subject to Breakdown. Operations Research **11**(3), 303–320 (1963)
8. Kulkarni, V., Choi, B.: Retrial Queues with Server Subject to Breakdowns and Repairs. Queueing Systems **7**(2), 191–208 (1990)
9. Kalidass, K., Gnanaraj, J., Gopinath, S., Kasturi, R.: Transient Analysis of an M/M/1 Queue with a Repairable Server and Multiple Vacations. International Journal of Mathematics in Operational Research **6**(2), 193–216 (2014)
10. Wan, G., Ren, Z., Wang, S.: Reliability Evaluation Algorithm of Power System using System False Fault Indices Pruning. Automation of Electric Power Systems **28**(23), 56–60 (2004)
11. Neuts, M.F.: Matrix-geometric Solution in Stochastic Model: An Algorithmic Application. The Johns Hopkins University Press, Baltimore (1981)
12. Latouche, G., Ramaswami, V.: Introduction to Matrix Analytic Methods in Stochastic Modeling. ASA-SIAM Series on Statistic and Applied Probability, Philadelphia (1999)
13. He, Q.M.: Fundamentals of Matrix-analytic Methods. Springer, New York (2014)

Analysis of an M/M/1 Retrial Queue with Speed Scaling

Tuan Phung-Duc and Wouter Rogiest

Abstract Recently, queues with speed scaling have received considerable attention due to their applicability to data centers, enabling a better balance between performance and energy consumption. This paper proposes a new model where blocked customers must leave the service area and retry after a random time, with retrial rate either varying proportionally to the number of retrying customers (linear retrial rate) or non-varying (constant retrial rate). For both, we study the case without and with setup time. In all four cases, we obtain an exact solution for the stationary queue length distribution. This document presents the resulting expressions as well as their derivation.

Keywords Data center · Energy efficiency · Speed scaling · Setup time · Retrial queue

1 Introduction

In current large-scale data centers, thousands of parallel servers are responsible for the processing of incoming jobs. While system performance is still measured by means of traditional measures like job latency, the overall energy consumption is a second important consideration. According to [4], data centers constitute about 40 % of the global ICT electricity consumption in 2012, or approximately 107 TWh. Concretely, a modern system needs mechanisms to handle the trade-off between performance and energy consumption [3].

T. Phung-Duc
Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Ookayama, Tokyo 152-8552, Japan
e-mail: tuan@is.titech.ac.jp

W. Rogiest (✉)
Department of Telecommunications and Information Processing, Ghent University, St.-Pietersnieuwstraat 41, 9000 Gent, Belgium
e-mail: wouter.rogiest@ugent.be

T.V. Do et al. (eds.), *Queueing Theory and Network Applications*,
Advances in Intelligent Systems and Computing 383,
DOI: 10.1007/978-3-319-22267-7_11

In response to this, speed scaling has been developed [6, 7, 15], slowing down server speed when the number of customers is low, and speeding up, in the converse case. As argued first in [7] (and later in [15]), this enables a better balance between performance and energy consumption. This is also argued in [19] in the context of data centers, and can be intuitively understood as follows. Assume that the speed of the system can be tuned by tuning the service rate ("speed scaling"). While power consumption rises more than proportionally with service rate (e.g., with the former approximately equal to the square of the latter [7]), this does not hold true for the mean number of customers in system. Specifically, the latter is approximately proportional to the mean service time (inverse of the service rate) in case of (very) low traffic load (with low arrival rate). Opposed to this, in case of high traffic load, speeding up can have a much larger than proportional impact on the number of customers in system, while the relation between service rate and power consumption remains the same. In other words, the added value per additional unit of power is higher when traffic load is high than when traffic load is low, creating a trade-off. In this sense, it is useful to work at lower speed when traffic load is low, and at higher speed in the converse case.

To the best of our knowledge, the first queueing model to address (a form of) speed scaling is [5], presenting the analysis of a single-server system with Poisson arrivals and a service rate that depends on the number of customers n according to a formula $\mu_n = n^c \mu_1$, where μ_1 is a model parameter describing the service rate for a customer arriving at an idle system. An important recent contribution with speed scaling is [15], which features the concept of *switching delay* discussed also below.

While [5, 15] study a classic model without retrials, some retrial queues have been studied which also relate to the current work. Specifically, while [14] does not discuss speed scaling as such, it presents a generic study of the broad class of retrial queues with state-dependent rates, sharing many of the assumptions of this contribution. However, it is important to note that, on the one hand, [14] does not include any of the expressions derived below, and that, on the other hand, the concept of a setup time is not treated in [14], whereas it plays a key role in this contribution. Specifically, Sect. 4 and 5 below are devoted to models with a setup time, an important and realistic model extension defined below, studied earlier in e.g. [1, 10, 11, 12, 13, 16, 17, 18]. Further, the mentioned switching delay of [15] is identical to the setup time as defined in this work. Summarizing, speed scaling has already been considered in settings with setup times, and also indirectly in settings with retrial queues, but never in the combination of both. Since both phenomena are found in realistic data centers, it is useful to quantify their impact by means of the formulas derived in this contribution.

This paper is organized as follows. In Sect. 2 and 3, a speed scaling model without setup time is considered, either with classical linear retrial rate (Sect. 2) or with constant retrial rate (Sect. 3). In Sect. 4 and 5, the speed scaling model extension with finite setup time is considered, again with either linear (Sect. 4) or constant (Sect. 5) retrial rate. Sect. 6 presents a note on practical implementation. Conclusions are drawn in Sect. 7.

2 Linear Retrial Rate Model

2.1 Assumptions

We consider a single server retrial queueing system where blocked customers leave the server and retry after independent and identically distributed (iid) retrial times. Retrials take place at rate $n\nu$, where n is the number of customers in orbit: A so-called linear retrial rate model. Further, as is common in retrial queue terminology, see e.g. [1, 8], during consecutive retrials, the customer is said to be in the orbit. However, different from a classical retrial queue, *speed scaling* takes place: The service rate of the server is linear to the total number of customers in the system. In particular, if there are n customers in the orbit the customer in the server (if any) is served at rate $(n+1)\mu$. Customers arrive at the system according to a Poisson process with rate λ.

2.2 Analysis

In this section, we present a recursive scheme to calculate the joint stationary distribution. Let $C(t)$ and $N(t)$ denote the number of active servers and the number of customers in the orbit, respectively. It is easy to see that $\{X(t) = (C(t), N(t)); t \geq 0\}$ forms a Markov chain on the state space:

$$\mathcal{S} = \{(i, j); i = 0, 1, j \in \mathbb{Z}_+\}.$$

Let $\pi_{i,j} = \lim_{t\to\infty} \mathrm{P}(C(t) = i, N(t) = j)$ $((i, j) \in \mathcal{S})$ denote the joint stationary distribution of $\{X(t)\}$.

In this section, we derive a recursion for calculating the joint stationary distribution $\pi_{i,j}$ $((i, j) \in \mathcal{S})$. The balance equations for states with $i = 0$ read as follows.

$$(\lambda + n\nu)\pi_{0,n} = (n+1)\mu\pi_{1,n}, \qquad n \in \mathbb{Z}_+, \tag{1}$$

$$(\lambda + (n+1)\mu)\pi_{1,n} = \lambda\pi_{0,n} + \lambda\pi_{1,n-1} + (n+1)\nu\pi_{0,n+1}, \qquad n \in \mathbb{Z}_+, \tag{2}$$

where $\mathbb{Z}_+$ denotes $\{1, 2, \ldots\}$. Using the notation of (5), we obtain the following system of equations for the partial generating functions $\Pi_0(z)$ and $\Pi_1(z)$.

$$\lambda\Pi_0(z) + \nu z\Pi_0'(z) = \mu z\Pi_1'(z) + \mu\Pi_1(z), \tag{3}$$

$$\lambda\Pi_1(z) + \mu z\Pi_1'(z) + \mu\Pi_1(z) = \lambda\Pi_0(z) + \lambda z\Pi_1(z) + \nu\Pi_0'(z). \tag{4}$$

Adding these two equations yields $\nu\Pi_0'(z) = \lambda\Pi_1(z)$. Substituting $\Pi_1(z)$ into the first equation we obtain

$$z\Pi_0''(z) + \frac{\lambda}{\mu}\left(\frac{\mu}{\lambda} - z\right)\Pi_0'(z) - \frac{\lambda^2}{\mu\nu}\Pi_0(z) = 0.$$

Coining the notation $p(x) = \Pi_0(\mu x/\lambda) = \Pi_0(x/\rho)$ $(\rho = \lambda/\mu)$, we obtain the following equation.

$$xp''(x) + (1-x)\,p'(x) - \frac{\lambda}{\nu}p(x) = 0.$$

This is the confluent hypergeometric differential equation whose solution is a confluent hypergeometric function, a special case of the hypergeometric function also encountered in the analysis of some retrial queue models without speed scaling, such as the one studied in [2]. The solution for this equation is given by following expression.

$$p(x) = \pi_{0,0} M(a,b,x) = \pi_{0,0} \sum_{n=0}^{\infty} \frac{a_{(n)} x^n}{b_{(n)} n!},$$

where

$$a = \frac{\lambda}{\nu}, \qquad b = 1,$$

and

$$a_{(0)} = 1, \qquad a_{(n)} = a(a+1)\cdots(a+n-1), \qquad n \geq 1,$$

where $M(a,b,x)$ denotes the confluent hypergeometric function. We then have

$$\Pi_0(z) = \pi_{0,0} p(\lambda z/\mu) = \pi_{0,0} \sum_{n=0}^{\infty} \frac{a_{(n)} (\lambda z/\mu)^n}{b_{(n)} n!} = \pi_{0,0} \sum_{n=0}^{\infty} \frac{a_{(n)} (\lambda z/\mu)^n}{n!^2},$$

where we used $b_{(n)} = n!$ in the second equality. Thus,

$$\pi_{0,n} = \pi_{0,0} \frac{a_{(n)} \rho^n}{b_{(n)} n!} = \pi_{0,0} \frac{a_{(n)}}{n!^2} \left(\frac{\lambda}{\mu}\right)^n.$$

Furthermore, we have

$$\Pi_1(z) = \frac{\nu}{\lambda} \Pi_0'(z) = \pi_{0,0} \frac{\lambda}{\mu} M(a+1, b+1, \lambda z/\mu),$$

where we have used

$$M'(a,b,x) = \frac{a}{b} M(a+1, b+1, x).$$

Formally, the unknown number $\pi_{0,0}$ is determined using the normalization condition:

$$\Pi_0(1) + \Pi_1(1) = 1.$$

yielding

$$\pi_{0,0} = \left(M(a, b, \lambda/\mu) + \frac{\lambda}{\mu} M(a+1, b+1, \lambda/\mu) \right)^{-1}.$$

Although this is an explicit expression for $\pi_{0,0}$, it still contains the confluent hypergeometric function, and thus, indirectly, infinite sums. This however poses no problem for the numerical calculation of $\pi_{0,0}$, since most scientific software packages are able to handle confluent hypergeometric functions directly.

3 Constant Retrial Rate Model

3.1 Assumptions

We consider a single server retrial queueing system where blocked customers leave the server and retry at a later time. As in the previous section, the retrial times are iid random variables. However, different from the previous section, the retrial rate is independent of the number of customers in the orbit and is given by $\nu(1 - \delta_{0,n})$ provided that there are n customers present in the orbit. Here, $\delta_{m,n}$ denotes the Kronecker delta, which returns 1 if $m = n$, and 0 otherwise. Again, *speed scaling* takes place: Service rate of the server is proportional to the total number of customers in the system. Just like in the linear retrial rate case studied in the previous section, if there are n customers in the orbit the customer in the server (if any) is served at rate $(n + 1)\mu$. Customers arrive at the system according to a Poisson process with rate λ.

3.2 Analysis

In this section, we present a recursive scheme to calculate the joint stationary distribution. Let $C(t)$ and $N(t)$ denote the number of active servers and the number of customers in the orbit, respectively. It is easy to see that $\{X(t) = (C(t), N(t)); t \geq 0\}$ forms a Markov chain on the state space:

$$\mathcal{S} = \{(i, j); i = 0, 1, j \in \mathbb{Z}_+\}.$$

Let $\pi_{i,j} = \lim_{t\to\infty} \mathrm{P}(C(t) = i, N(t) = j)$ $((i, j) \in \mathcal{S})$ denote the joint stationary distribution of $\{X(t)\}$.

In this section, we derive a recursion for calculating the joint stationary distribution $\pi_{i,j}$ $((i, j) \in \mathcal{S})$. The balance equations for states with $i = 0$ read as follows.

$$\begin{aligned}
(\lambda + \nu(1 - \delta_{0,n})))\pi_{0,n} &= (n+1)\mu\pi_{1,n}, \qquad n \in \mathbb{Z}_+, \\
(\lambda + (n+1)\mu)\pi_{1,n} &= \lambda\pi_{0,n} + \lambda\pi_{1,n-1} + \nu\pi_{0,n+1}, \qquad n \in \mathbb{Z}_+.
\end{aligned}$$

We define partial generating functions as follows.

$$\Pi_0(z) = \sum_{n=0}^{\infty} \pi_{0,n} z^n, \qquad \Pi_1(z) = \sum_{n=0}^{\infty} \pi_{1,n} z^n. \tag{5}$$

We obtain the following system of equations for generating functions.

$$\lambda \Pi_0(z) + \nu(\Pi_0'(z) - \pi_{0,0}) = \mu z \Pi_1'(z) + \mu \Pi_1(z), \tag{6}$$

$$\lambda \Pi_1(z) + \mu z \Pi_1'(z) + \mu \Pi_1(z) = \lambda \Pi_0(z) + \lambda z \Pi_1(z) + \frac{\nu}{z}(\Pi_0(z) - \pi_{0,0}). \tag{7}$$

Summing up these two equations yields

$$\lambda \Pi_1(z) = \frac{\nu(\Pi_0(z) - \pi_{0,0})}{z}$$

or

$$z\Pi_1(z) = \frac{\nu(\Pi_0(z) - \pi_{0,0})}{\lambda}.$$

Taking the first derivative of the latter equation with respect to z and substituting the result in the right-hand side of (6) yields

$$\lambda \Pi_0(z) + \nu(\Pi_0(z) - \pi_{0,0}) = \frac{\mu\nu}{\lambda} \Pi_0'(z)$$

or

$$\Pi_0'(z) = \frac{\lambda(\lambda + \nu)}{\mu\nu} \Pi_0(z) - \frac{\lambda}{\mu} \pi_{0,0}.$$

Solving this equation we obtain

$$\Pi_0(z) = \pi_{0,0} \left[\frac{\lambda}{\lambda + \nu} \exp(\gamma z) + \frac{\nu}{\lambda + \nu} \right],$$

where we coined the notation

$$\gamma = \frac{\lambda(\lambda + \nu)}{\mu\nu}.$$

We also find that

$$\Pi_1(z) = \frac{\nu}{\lambda + \nu} \frac{\exp(\gamma z) - 1}{z} \pi_{0,0}.$$

From the normalization condition,

$$\Pi_0(1) + \Pi_1(1) = 1,$$

we find that $\pi_{0,0} = \exp(-\gamma)$.

4 Linear Retrial Rate Model with Setup Time

In this section, we consider an extension of the model studied in Sect. 2, introducing the notion of a setup time. As is the case in many realistic systems, upon turning idle (i.e., empty server and empty orbit), the system may go into sleep mode (or hybernation mode) to save energy, returning to active mode when triggered by the arrival of a new customer. Moving from idle to active mode may happen instantaneously (as in the models of Sect. 2 and 3) or the system may be in setup mode during a finite time called the setup time. In this section and the following, we assume finite iid setup times with exponential distribution with parameter α. Further, we assume that the first customer in the busy period immediately goes to the server without joining the orbit. Arriving customers who find the server occupied (either setting up or actually serving) join the orbit and repeat their attempt after some random time. Below, the terms "busy" and "active" are interchangeable, as well as "idle" and "sleeping".

Let $C(t)$ denote the state of the server and $N(t)$ denote the number of customers in the orbit at time t.

$$C(t) = \begin{cases} 0, & \text{the server is idle,} \\ 1, & \text{the server is busy,} \\ 2, & \text{the server is in setup mode.} \end{cases}$$

Here, $\{X(t) = (C(t), N(t)); t \geq 0\}$ forms a Markov chain on the state space

$$\mathcal{S} = \{(i, j); i \in \{0, 1, 2\}, j \in \mathbb{Z}_+\},$$

where $C(t) = 0, 1, 2$ implies that the server is idle, busy or in setup mode, respectively. It is easy to see that the system is always stable due to the speed scaling. Let $\pi_{i,j} = \lim_{t\to\infty} \mathrm{P}(C(t) = i, N(t) = j$. Our goal is to explicitly express all $\pi_{i,j}$ in terms of $\pi_{0,0}$ which is uniquely determined using the normalization condition.

More specially, let (0,0) denote the state corresponding to sleep mode (with thus an idle server), while $(0, j)$ $(j \geq 1)$ denotes states for which the server is idle while there are j customers in the orbit. Further, the states $(1, j)$ $(j \geq 1)$ correspond to a server busy serving a customer with j customers present in the orbit. Finally, the states $(2, j)$ $(j \geq 1)$ correspond to one customer awaiting setup in the server with j customers present in the orbit. The balance equation for an idle server reads

$$(\lambda + n\nu)\pi_{0,n} = (n + 1)\mu\pi_{1,n},$$

which is identical to (1), the balance equation *without* setup time. As a result, the relation between the partial generating functions $\Pi_0(z)$ and $\Pi_1(z)$ (defined by (5)) also holds true here. Opposed to this, the balance equations for a busy server, with states $(1, j)$, explicitly involve the setup parameter α, as follows.

$$(\lambda + \mu)\pi_{1,0} = \nu\pi_{0,1} + \alpha\pi_{2,0}, \tag{8}$$

$$(\lambda + (n + 1)\mu)\pi_{1,n} = \lambda\pi_{0,n} + \lambda\pi_{1,n-1} + (n + 1)\nu\pi_{0,n+1} + \alpha\pi_{2,n}. \tag{9}$$

Introducing the generating function $\Pi_2(z) = \sum_{j=0}^{\infty} \pi_{2,j} z^j$, we then have

$$\lambda \Pi_1(z) + \mu \Pi_1(z) + \mu z \Pi_1'(z) = \lambda(\Pi_0(z) - \pi_{0,0}) + \lambda z \Pi_1(z) + \nu \Pi_0'(z) + \alpha \Pi_2(z).$$

The balance equations for a server setting up, with states $(2, j)$ are given by

$$(\lambda + \alpha)\pi_{2,0} = \lambda \pi_{0,0}, \tag{10}$$

$$(\lambda + \alpha)\pi_{2,j} = \lambda \pi_{2,j-1}, \qquad j = 1, 2 \ldots, \tag{11}$$

leading to

$$(\lambda + \alpha)\Pi_2(z) = \lambda z \Pi_2(z) + \lambda \pi_{0,0} \quad \Longleftrightarrow \quad \Pi_2(z) = \frac{\lambda \pi_{0,0}}{\lambda + \alpha - \lambda z}.$$

Taking the balance of flows in and out the orbit yields

$$\lambda(\Pi_1(z) + \Pi_2(z)) = \nu \Pi_0'(z).$$

Multiplying both sides by z and taking the derivative of both sides yields

$$\lambda[(z\Pi_1(z))' + (z\Pi_2(z))'] = \nu z \Pi_0''(z) + \nu \Pi_0'(z).$$

Substituting $(z\Pi_1(z))'$ in terms of $\Pi_0(z)$ we find the following differential equation.

$$\lambda \frac{\lambda \Pi_0(z) + \nu z \Pi_0'(z)}{\mu} + \lambda (z\Pi_2(z))' = \nu z \Pi_0''(z) + \nu \Pi_0'(z).$$

Reworking this equation, we obtain

$$z\Pi_0''(z) + (1 - \frac{\lambda}{\mu} z)\Pi_0'(z) - \frac{\lambda^2}{\mu \nu} \Pi_0(z) = \frac{\lambda}{\nu} (z\Pi_2(z))',$$

where

$$\Pi_2(z) = \frac{\lambda \pi_{0,0}}{\lambda + \alpha - \lambda z}.$$

This is a non-homogeneous confluent differential equation and its explicit solution seems difficult. But we can solve it by power expansion method.

In particular, substituting $\Pi_0(z) = \sum_{j=0}^{\infty} \pi_{0,j} z^j$ into the left hand side of the differential equation we obtain

$$\sum_{j=0}^{\infty} \left[(j+1)^2 \pi_{0,j+1} - \frac{\lambda}{\mu} \left(j + \frac{\lambda}{\nu} \right) \pi_{0,j} \right] z^j = \frac{\lambda^2 \pi_{0,0}}{\nu(\lambda + \alpha)} \sum_{j=0}^{\infty} (j+1) \left(\frac{\lambda}{\lambda + \alpha} \right)^j z^j,$$

where we have used

$$\Pi_2(z) = \frac{\lambda\pi_{0,0}}{\lambda+\alpha}\left(\sum_{j=0}^{\infty}\frac{\lambda z}{\lambda+\alpha}\right)^j, \tag{12}$$

and thus

$$(z\Pi_2(z))' = \frac{\lambda\pi_{0,0}}{\lambda+\alpha}\sum_{j=0}^{\infty}\left(\frac{\lambda}{\lambda+\alpha}\right)^j (j+1)z^j.$$

Comparing the coefficients of z^0 in both sides yields,

$$\pi_{0,1} = \frac{\lambda^2(\lambda+\mu+\alpha)}{\mu\nu(\lambda+\alpha)}\pi_{0,0}.$$

Assuming that $\pi_{0,j} = \beta_j\pi_{0,0}$ ($j \in \mathbb{Z}_+$), it follows from the comparison between the coefficients of z^j that

$$\begin{aligned}(j+1)^2\beta_{j+1} - \frac{\lambda}{\mu}\left(j+\frac{\lambda}{\nu}\right)\beta_j &= \frac{\lambda^2}{\nu(\lambda+\alpha)}(j+1)\left(\frac{\lambda}{\lambda+\alpha}\right)^j, \\ &= \frac{\lambda}{\nu}(j+1)\left(\frac{\lambda}{\lambda+\alpha}\right)^{j+1}\end{aligned}$$

where $\beta_0 = 1$. This equation leads to

$$\beta_{j+1} = \frac{\lambda}{\mu}\frac{(j+\lambda/\nu)}{(j+1)^2}\beta_j + \frac{\lambda}{\nu}\frac{(\lambda/(\lambda+\alpha))^{j+1}}{j+1},$$

where $\beta_0 = 1$. This equation allows to calculate $\pi_{0,j}$ in terms of $\pi_{0,0}$ for any j. Thus, using (1), we can also calculate $\pi_{1,j}$ in terms of $\pi_{0,0}$ for any j. Determining $\pi_{0,0}$ can be done by means of the recursion explained below in Sect. 6.

5 Constant Retrial Rate Model with Setup Time

In this section, we extend the model of Sect. 3 with the notion of a setup time, an iid random variable with exponential distribution with parameter α. Further, the state space is the same as in the previous section. Finally, while the steady-state distribution is obviously different, we use the same notation as in the previous section. The balance equations are as follows.

$$\begin{aligned}\lambda\pi_{0,0} &= \mu\pi_{1,0}, \\ (\lambda+\nu)\pi_{0,n} &= (n+1)\mu\pi_{1,n}, \qquad n \geq 1.\end{aligned}$$

Transforming this equation to z-domain yields,

$$(\lambda+\nu)\Pi_0(z)-\nu\pi_{0,0}=\mu(z\Pi_1'(z)+\Pi_1(z))$$

Balance of flows in and out the orbit yields

$$\lambda(\Pi_1(z)+\Pi_2(z))=\frac{\nu}{z}(\Pi_0(z)-\pi_{0,0}). \tag{13}$$

Multiplying both sides by z and taking the derivative of both sides arranging the result yields

$$\begin{aligned}\Pi_0'(z)&=\frac{\lambda(\lambda+\nu)}{\mu\nu}\Pi_0(z)-\frac{\lambda}{\mu}\pi_{0,0}+\frac{\lambda}{\nu}(z\Pi_2(z))',\\&=\gamma\Pi_0(z)+\pi_{0,0}Q(z),\end{aligned} \tag{14}$$

where

$$Q(z)=-\frac{\lambda}{\mu}+\frac{\lambda}{\nu}\left(\frac{\lambda z}{\lambda+\alpha-\lambda z}\right)',\qquad \gamma=\frac{\lambda(\lambda+\nu)}{\mu\nu}.$$

It should be noted that we have used

$$\Pi_2(z)=\frac{\lambda\pi_{0,0}}{\lambda+\alpha-\lambda z}.$$

The solution of the differential equation (14) has the form:

$$\Pi_0(z)=\pi_{0,0}\exp(\gamma z)\left(1+\int_0^z\exp(-\gamma u)Q(u)du\right).$$

Hence, formally, we have $\Pi_0(1)=\kappa_0\pi_{0,0}$ where

$$\kappa_0=\exp(\gamma)\left(1+\int_0^1\exp(-\gamma u)Q(u)du\right).$$

From (13), we also have $\Pi_1(1)+\Pi_2(1)=\kappa_1\pi_{0,0}$ where

$$\kappa_1=\frac{\nu}{\lambda}(\kappa_0-1).$$

From the normalization condition

$$\Pi_0(1)+\Pi_1(1)+\Pi_2(1)=1,$$

we can obtain

$$\pi_{0,0}=\frac{1}{\kappa_0+\kappa_1}.$$

We can obtain κ_0 (and thus, also κ_1 and $\pi_{0,0}$) using numerical integration which is readily available in almost all scientific software packages. Furthermore, $\pi_{0,0}$ can also be obtained directly by means of the recursion explained next.

6 Recursive Approach

From theoretical point of view, the results in the previous two sections are nice since they are related to some well-known differential equation. However, from practical point of view, it is more convenient to evaluate the stationary probabilities via some simple recursion.

Practically, the approach for the model of Sect. 4 is as follows. In a first step, we set $\pi_{0,0} = 1$. In a second step, we can calculate $\pi_{2,0}$ and then $\pi_{1,0}$. Using these results, we can calculate $\pi_{0,1}$ using the balance equation in and out the orbit, i.e.,

$$(n+1)\nu\pi_{0,n+1} = \lambda(\pi_{1,n} + \pi_{2,n}).$$

The probability $\pi_{2,n+1}$ is easily calculated in terms of $\pi_{0,0}$ for any n using (10) and (11).

So, we can again use the following balance equation in order to determine $\pi_{1,n+1}$.

$$(\lambda + (n+1)\nu)\pi_{0,n+1} = (n+2)\mu\pi_{1,n+1}.$$

The step from n to $n+1$ is taken in the same manner. As a result, we can calculate relative values of the $\pi_{i,n}$ ($i = 0, 1, 2$) for any value of n up to a certain value $n = N_0$ characterizing the accuracy (the larger the more accurate), and then normalize the result by ensuring that the sum of the obtained probabilities is 1.

A similar procedure can be applied for the models of Sect. 2, 3 and 5. As a result, we can calculate any desired performance measure with high accuracy, by setting N_0 sufficiently high.

7 Conclusions

In this contribution, we studied an M/M/1 retrial queue model with speed scaling. The analysis yielded an exact solution for the steady-state queue length distribution, and this for four different cases: two without setup times (either linear or constant retrial rate), and two with setup times (again, linear or constant retrial rate).

With these results available, future work is to study the trade-off between performance and energy consumption, inherent to speed scaling systems. Here, a first route is by means of the existing cost function used in [7, 15]; however, this may ideally be contrasted with alternative formulations of the mentioned trade-off.

Acknowledgments Tuan Phung-Duc was supported in part by Japan Society for the Promotion of Science, JSPS Grant-in-Aid for Young Scientists (B), Grant Number 2673001. Wouter Rogiest is Postdoctoral Fellow with the Research Foundation Flanders (FWO-Vlaanderen). Part of this

research has been funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office. The authors would like to thank the referees for their constructive comments which helped to improve this paper.

References

1. Artalejo, J.R., Economou, A., Lopez-Herrero, M.J.: Analysis of a multiserver queue with setup times. Queueing Systems **51**(1–2), 53–76 (2005)
2. Artalejo, J.R., Phung-Duc, T.: Markovian Retrial Queues with Two Way Communication. Journal of Industrial and Management Optimization **8**(4), 781–806 (2012)
3. Barroso, L.A., Holzle, U.: The case for energy-proportional computing. Computer **40**(12), 33–37 (2007)
4. Van Heddeghem, W., Lambert, S., Lannoo, B., Colle, D., Pickavet, M., Demeester, P.: Trends in worldwide ICT electricity consumption from 2007 to 2012. Computer Communications **50**(1), 64–76 (2014)
5. Conway, R.W., Maxwell, W.L.: A Queueing Model with State Dependent Service Rate. Journal of Industrial Engineering **12**, 132–136 (1961)
6. Yao, F., Demers, A., Shenker, S.: A scheduling model for reduced CPU energy. In: Proceedings 36th Annual Symposium on Foundations of Computer Science, pp. 374–382 (1995)
7. Wierman, A., Andrew, L., Tang, A.: Power-aware speed scaling in processor sharing systems. In: Proceedings of IEEE INFOCOM 2009, pp. 2007–2015 (2009)
8. Phung-Duc, T., Rogiest, W., Takahashi, Y., Bruneel, H.: Retrial queues with balanced call blending: analysis of single-server and multiserver model. Annals of Operations Research (2014). doi:10.1007/s10479-014-1598-2
9. Phung-Duc, T.: An explicit solution for a tandem queue with retrials and losses. Operational Research **12**(2), 189–207 (2012)
10. Phung-Duc, T.: Impatient customers in power-saving data centers. In: Sericola, B., Telek, M., Horváth, G. (eds.) ASMTA 2014. LNCS, vol. 8499, pp. 185–199. Springer, Heidelberg (2014)
11. Phung-Duc, T.: Server farms with batch arrival and staggered setup. In: Proceedings of the Fifth Symposium on Information and Communication Technology, pp. 240–247. ACM (2014)
12. Phung-Duc, T.: Exact solution for M/M/*c*/Setup queue (2014). Preprint: http://arxiv.org/abs/1406.3084
13. Phung-Duc, T.: Multiserver queues with finite capacity and setup time. In: Remke, A., Manini, D., Gribaudo, M. (eds.) ASMTA 2015. LNCS, vol. 9081, pp. 173–187. Springer, Heidelberg (2015)
14. Parthasarathy, P.R., Sudhesh, R.: Time-dependent analysis of a single-server retrial queue with state-dependent rates. Operations Research Letters **35**(5), 601–611 (2007)
15. Lu, X., Aalto, S., Lassila, P.: Performance-energy trade-off in data centers: impact of switching delay. In: Proceedings of 22nd IEEE ITC Specialist Seminar on Energy Efficient and Green Networking (SSEEGN), pp. 50–55 (2013)
16. Gandhi, A., Harchol-Balter, M., Adan, I.: Server farms with setup costs. Performance Evaluation **67**, 1123–1138 (2010)
17. Gandhi, A., Doroudi, S., Harchol-Balter, M., Scheller-Wolf, A.: Exact analysis of the M/M/k/setup class of markov chains via recursive renewal reward. In: Proceedings of the ACM SIGMETRICS, pp. 153–166 (2013)
18. Gandhi, A., Doroudi, S., Harchol-Balter, M., Scheller-Wolf, A.: Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward. Queueing Systems **77**(2), 177–209 (2014)
19. Mitrani, I.: Managing performance and power consumption in a server farm. Annals of Operations Research **202**(1), 121–134 (2013)

Part IV
Network Models

Mathematical Model and Performance Evaluation of AMI Applied to Mobile Environment

Shunsuke Matsuzawa, Satoru Harada, Kazuya Monden, Yukihiro Takatani and Yutaka Takahashi

Abstract Advanced Metering Infrastructure (AMI) will play an important role in order to realize the vision of a smart grid. An integral part of AMI is a set of smart meters, which measure and transmit power consumption data periodically with fixed intervals. If the bandwidth allocated to AMI is much more than the capacity necessary for collecting power consumption data, the remaining part of the bandwidth can be used for other types of communication. In this paper, we propose a mathematical model and evaluate the communication performance of AMI taking into account the variable number of smart meters, caused by the turn on/off of meters and mobility of meters across the AMI network coverage. We construct a discrete time Markov chain whose states are defined by the numbers of on-state meters and request-holding meters, deriving some performance measures.

Keywords AMI · Mobile environment · Random access · Markov chain

1 Introduction

For the purpose of realization of a smart grid, Advanced Metering Infrastructure (AMI) has been attracting considerable attention. An integral part of AMI is a set of smart meters. Smart meters measure and transmit power consumption data. Monitoring power demand is useful for setting more flexible electricity bill, power demand control, and peak shift [1]. In addition to collection of power consumption data, smart meters can be used for other types of communication such as gas and water consumption, demand response for electricity, and inquiries to electric companies [2].

S. Matsuzawa(✉) · Y. Takahashi
Graduate School of Informatics, Kyoto University, Yoshida Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
e-mail: matsuzawa@sys.i.kyoto-u.ac.jp, takahashi@i.kyoto-u.ac.jp

S. Harada · K. Monden · Y. Takatani
Hitachi, Ltd., Yokohama Research Laboratory, Yokohama, Japan

T.V. Do et al. (eds.), *Queueing Theory and Network Applications*,
Advances in Intelligent Systems and Computing 383,
DOI: 10.1007/978-3-319-22267-7_12

Although meter reading operations have been regarded as difficult due to augmenting concern about preserving security, smart meters may resolve the problem [3]. Live information of residents can be inferred from collected data, which is helpful in elderly support services. In order to put AMI into practice, however, it is necessary to reduce communication and electricity costs [4].

Wireless multihop networks will be mainly adopted for AMI from the perspective of communication costs and scalability. Smart meters autonomously compose wireless multihop networks, which enables communication with gateways even if radio waves do not reach directly [5]. Besides, smart meters can select another path when a part of a network is damaged.

Power consumption data is collected periodically with fixed intervals called check cycles. If the bandwidth allocated for AMI is much more than the capacity necessary for collecting power consumption data, the remaining part of the bandwidth can be used for other types of communication such as gas and water consumption, demand response for electricity, and inquiries to electric companies. A hybrid protocol [6, 7, 8] will be used in AMI to integrate collection of power consumption data and irregular data communication in the same network. For instance, we can apply a polling protocol to periodical data collection and a random access protocol to infrequent or occasional data communication. Smart meters are polled sequentially one by one for their regular transmission in a polling mode, while each smart meter transmits data autonomously in a random access mode, as a result, the transmission may fail.

We assume that the bandwidth allocated for a random access mode is divided into time slots and that the length of time slots is determined by the maximum number of hops within the network. A large number of studies have been made on random access protocols based on slotted ALOHA [9, 10, 11].

A previous work [12] assesses scheduling methods to integrate polling and random access protocols in the same bandwidth. However, there has been no study that takes into account variable numbers of smart meters within the AMI network, called on-state meters. The variability is caused by the turn on/off of meters, and mobility of meters across the AMI network coverage.

In this paper, we evaluate the communication performance of AMI in terms of performance measures such as throughput, the probability of successful transmissions, and transmission delay. Firstly, we construct a discrete time Markov chain whose states are defined by the numbers of on-state meters and request-holding meters. Secondly, we derive the stationary state distribution and throughput. Finally, we derive the probability of successful transmissions and transmission delay by using Laplace-Stieltjes transform of the probability distribution function.

This paper is organized as follows. Section 2 presents a mathematical model of AMI. In Section 3, we derive performance measures. Section 4 shows numerical examples of these performance measures. Finally, we conclude the paper in Section 5.

2 Mathematical Model

We describe a mathematical model of AMI in this section. There are N_{RA} smart meters in the AMI network. Let T denote the length of a check cycle, that is, power consumption data is collected with intervals of T. We then decompose T into $T = T_{\mathrm{P}} + T_{\mathrm{RA}}$, where T_{P} (resp., T_{RA}) denotes the length of a polling mode (resp., random access mode). We also define a period between the beginning of a random access mode and the end of the next polling mode as a frame.

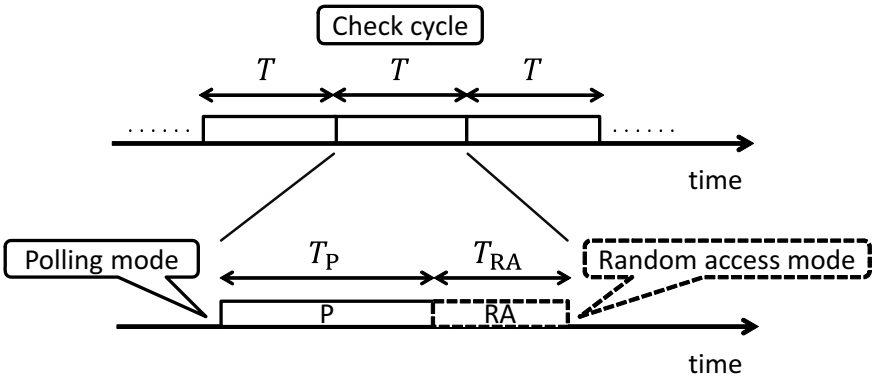


Fig. 1 Communication schedule for AMI

In a random access mode, smart meters decide autonomously when to start their transmission. Smart meters holding transmission attempts randomly and independently choose time slots and try to transmit data. If a time slot is chosen by two or more smart meters, the transmissions in the time slot fail. Smart meters which failed in their transmissions wait until the next random access mode begins. Let t_{RA} and L denote the length of a time slot and the number of time slots for a random access mode, respectively. We then have $T_{\mathrm{RA}} = Lt_{\mathrm{RA}}$. Moreover, t_{RA} satisfies $t_{\mathrm{RA}} = H_{\max} t_{\mathrm{H}}$, where $H_{\max}$ and t_{H} denote the maximum number of hops in the network and time necessary for one hop, respectively.

We assume that smart meters are either in on-state or off-state and that the states may change immediately after a random access mode begins. We call on-state meters which have (resp., do not have) transmission attempts request-holding meters (resp., non-request meters). The number of frames until an on-state meter (resp., an off-state meter) changes its state into off-state (resp., on-state) follows a geometric distribution with a parameter a (resp., b). Non-request meters can generate transmission attempts with probability λ immediately before a random access mode starts, while the requests are canceled when their states change into off-state.

3 Analysis

In this section, we construct a discrete time Markov chain whose states are defined by the numbers of on-state meters and request-holding meters at the beginning of random access modes. We then derive the stationary state distribution and obtain some performance measures.

3.1 Construction of Markov Chain

Let $N_{\mathrm{ON}}^{(n)}$ (resp., $N_{\mathrm{RQ}}^{(n)}$) denote the number of on-state meters (resp., request-holding meters) at the beginning of nth random access mode starts. From the assumptions for the model, $\{(N_{\mathrm{ON}}^{(n)}, N_{\mathrm{RQ}}^{(n)})\}$ is a discrete time Markov chain with finite state space $\mathbb{S}_1$, where

$$\mathbb{S}_1 = \{(i,k) | i \in \{0,1,2,\ldots,N_{\mathrm{RA}}\}, k \in \{0,1,2,\ldots,i\}\}.$$

Note here that this Markov chain is irreducible and aperiodic. Therefore, we have the unique stationary state distribution vector $\boldsymbol{\pi} := (\pi_k(i))_{(i,k)\in\mathbb{S}_1^2}$. By definition, $\boldsymbol{\pi}$ is given by

$$\boldsymbol{\pi}\boldsymbol{P} = \boldsymbol{\pi}, \ \sum_{i=0}^{N_{\mathrm{RA}}} \sum_{k=0}^{i} \pi_k(i) = 1,$$

where $\boldsymbol{P}$ denotes the transition probability matrix of the Markov chain $\{(N_{\mathrm{ON}}^{(n)}, N_{\mathrm{RQ}}^{(n)})\}$ and is derived in the following.

Let us now define the transition probability as follows,

$$p_{k,l}(i,j) := \Pr(N_{\mathrm{ON}}^{(n+1)} = j, N_{\mathrm{RQ}}^{(n+1)} = l \mid N_{\mathrm{ON}}^{(n)} = i, N_{\mathrm{RQ}}^{(n)} = k), \ (i,k) \times (j,l) \in \mathbb{S}_1^2.$$

We assume that i_1 request-holding meters and i_2 non-request meters change their states into off-state immediately after the random access mode begins, and then $j + i_1 + i_2 - i$ off-state meters change their states into on-state. Further, $l - k_1$ transmission attempts are generated if k_1 request-holding meters fail in their transmissions. Thus, we have

$$\begin{aligned} p_{k,l}(i,j) = \sum_{i_1=0}^{k} \sum_{i_2=\max(0,i-j-i_1)}^{i-k} \sum_{k_1=0}^{\min(l,k-i_1)} & S_{\mathrm{OFF}}(k,i_1) S_{\mathrm{OFF}}(i-k,i_2) \\ & \cdot S_{\mathrm{ON}}(N_{\mathrm{RA}} - i, j + i_1 + i_2 - i) q_{k-i_1,k_1} B_{l-k_1}(j-k_1), \end{aligned}$$

where $S_{\mathrm{OFF}}(i,j)$ (resp., $S_{\mathrm{ON}}(i,j)$) denotes the probability of j out of i on-state meters (resp., off-state meters) changing their states into off-state (resp., on-state)and

where $B_k(i)$ is the probability of transmission attempts being generated in k out of i non-request meters. These are obtained as follows.

$$S_{\mathrm{OFF}}(i,j) = \begin{cases} \binom{i}{j} a^j (1-a)^{i-j}, & 0 \le j \le i, \\ 0, & \text{otherwise}, \end{cases}$$

$$S_{\mathrm{ON}}(i,j) = \begin{cases} \binom{i}{j} b^j (1-b)^{i-j}, & 0 \le j \le i, \\ 0, & \text{otherwise}, \end{cases}$$

$$B_k(i) = \begin{cases} \binom{i}{k} \lambda^k (1-\lambda)^{i-k}, & 0 \le k \le i, \\ 0, & \text{otherwise}. \end{cases}$$

Besides, $q_{k,l}$ denotes the probability of $k-l$ out of k request-holding meters succeeding in their transmissions. See Appendix A for the derivation of $q_{k,l}$.

3.2 Throughput

We define throughput, θ, as the number of successful transmissions per a time slot in a random access mode. The mean number of successful transmissions is given by

$$\sum_{i=0}^{N_{\mathrm{RA}}} \sum_{k=0}^{i} \sum_{i_1=0}^{k} \sum_{k_1=0}^{k-i_1} \pi_k(i) S_{\mathrm{OFF}}(k, i_1) q_{k-i_1,k_1} \cdot (k - i_1 - k_1).$$

Therefore, We have

$$\theta = \frac{1}{L} \sum_{i=0}^{N_{\mathrm{RA}}} \sum_{k=0}^{i} \sum_{i_1=0}^{k} \sum_{k_1=0}^{k-i_1} \pi_k(i) S_{\mathrm{OFF}}(k, i_1) q_{k-i_1,k_1} \cdot (k - i_1 - k_1).$$

3.3 Probability of Successful Transmissions and Transmission Delay

In this subsection, we consider the probability of successful transmissions and transmission delay. Transmission delay is defined as time from when a transmission attempt is generated until when its transmission is completed successfully.

Let $(N_{\mathrm{ON}}, N_{\mathrm{RQ}})$ denote the state at the beginning of a random access mode immediately after a tagged transmission attempt is generated and $\alpha_k(i)$ denote the probability of $(N_{\mathrm{ON}}, N_{\mathrm{RQ}}) = (i,k)$. Further, we define D and $\mathcal{G}$ as the transmission delay of the transmission and the event of the request being successfully transmitted, respectively. We also define a distribution function denoted by

$$F^{(i,k)}(t) = \Pr(D \le t, \mathcal{G} \mid (N_{\mathrm{ON}}, N_{\mathrm{RQ}}) = (i,k)),\ (i,k) \in \mathbb{S}_2,$$

where

$$\mathbb{S}_2 = \{(i,k) | i \in \{1,2,\ldots,N_{\mathrm{RA}}\}, k \in \{1,2,\ldots,i\}\}.$$

Let $F^{(i,k)*}(s)$ denote the Laplace-Stieltjes transform (LST) of $F^{(i,k)}(t)$. The probability of successful transmissions, P_{s}, is then given by $P_{\mathrm{s}} = \boldsymbol{\alpha}\boldsymbol{F}^*(0)$, where $\boldsymbol{\alpha} := (\alpha_k(i))_{(i,k)\in\mathbb{S}_2}$ is a column vector, and $\boldsymbol{F}^*(s) := (F^{(i,k)*}(s))_{(i,k)\in\mathbb{S}_2}$ is a row vector. Similarly, we have

$$\mathrm{E}[D \mid \mathcal{G}] = -\frac{1}{P_{\mathrm{s}}}\boldsymbol{\alpha}\boldsymbol{F}^{*\prime}(0),\ \mathrm{E}\left[D^2 \mid \mathcal{G}\right] = \frac{1}{P_{\mathrm{s}}}\boldsymbol{\alpha}\boldsymbol{F}^{*\prime\prime}(0).$$

Moreover, the variance of transmission delay is given by

$$\mathrm{Var}[D \mid \mathcal{G}] = \mathrm{E}[D^2 \mid \mathcal{G}] - (\mathrm{E}[D \mid \mathcal{G}])^2.$$

We now derive $\alpha_k(i)$. The mean number of transmission attempts generated immediately before a random access mode starts is given by

$$\begin{aligned}
&\sum_{j_2=1}^{N_{\mathrm{RA}}}\sum_{l_2=1}^{j_2}\sum_{j_1=0}^{N_{\mathrm{RA}}}\sum_{l_1=0}^{j_1}\sum_{i_1=0}^{l_1}\sum_{i_2=\max(0,j_1-j_2-i_1)}^{j_1-l_1}\sum_{k_1=0}^{\min(l_2,l_1-i_1)} \pi_{l_1}(j_1)S_{\mathrm{OFF}}(l_1,i_1)\\
&\quad\cdot\ S_{\mathrm{OFF}}(j_1-l_1,i_2)S_{\mathrm{ON}}(N_{\mathrm{RA}}-j_1, j_2+i_1+i_2-j_1)\\
&\quad\cdot\ q_{l_1-i_1,k_1}B_{l_2-k_1}(j_2-k_1)\cdot(l_2-k_1), \qquad (1)
\end{aligned}$$

where j_1 (resp., l_1) denotes the number of on-state meters (resp., request-holding meters) at the beginning of a random access mode and where j_2 (resp., l_2) denotes the number of on-state meters (resp., request-holding meters) at the beginning of the next random access mode. Suppose $j_2 = i$ and $l_2 = k$ in (1), we have

$$\begin{aligned}
&\sum_{j_1=0}^{N_{\mathrm{RA}}}\sum_{l_1=0}^{j_1}\sum_{i_1=0}^{l_1}\sum_{i_2=\max(0,j_1-j_2-i_1)}^{j_1-l_1}\sum_{k_1=0}^{\min(k,l_1-i_1)} \pi_{l_1}(j_1)S_{\mathrm{OFF}}(l_1,i_1)S_{\mathrm{OFF}}(j_1-l_1,i_2)\\
&\quad\cdot\ S_{\mathrm{ON}}(N_{\mathrm{RA}}-j_1, i+i_1+i_2-j_1)q_{l_1-i_1,k_1}B_{k-k_1}(j_2-k_1)\cdot(k-k_1). \qquad (2)
\end{aligned}$$

As a result, $\alpha_k(i)$ can be derived by dividing (2) by (1).

We next derive $F^{(i,k)*}(s)$ and its derivatives. Suppose that the transmission of the tagged transmission attempt is successful during the random access mode immediately after its generation, then let d ($0 \le d \le T_{\mathrm{RA}}$) denote a random variable of transmission delay. From the assumption for our model, d follows a discrete uniform distribution with $t_{\mathrm{RA}}, 2t_{\mathrm{RA}} \ldots, Lt_{\mathrm{RA}}$. Let $d^*(s)$ denote the LST of the distribution function of d, we then have

$$d^{*\prime}(0) = -\mathrm{E}[d] = -\frac{1}{L}\sum_{m=1}^{L} m t_{\mathrm{RA}} = -\frac{L+1}{2} t_{\mathrm{RA}},$$

$$d^{*\prime\prime}(0) = \mathrm{E}\left[d^2\right] = \frac{1}{L}\sum_{m=1}^{L} m^2 t_{\mathrm{RA}}^2 = \frac{(L+1)(2L+1)}{6} t_{\mathrm{RA}}^2.$$

In the case of $(N_{\mathrm{ON}}, N_{\mathrm{RQ}}) = (i, 1),\ i \in \{1, 2, \ldots, N_{\mathrm{RA}}\}$, the request-holding meter succeeds in its transmission if its state remains on-state, so $F^{(i,1)*}(s) = (1 - a)d^*(s)$. Thus, we have

$$\begin{aligned} F^{(i,1)*}(0) &= 1 - a, \\ F^{(i,1)*\prime}(0) &= -(1-a) \cdot \frac{L+1}{2} t_{\mathrm{RA}}, \\ F^{(i,1)*\prime\prime}(0) &= (1-a) \cdot \frac{(L+1)(2L+1)}{6} t_{\mathrm{RA}}^2. \end{aligned}$$

We next consider the case of $(N_{\mathrm{ON}}, N_{\mathrm{RQ}}) = (i, k) \in \mathbb{S}_3$, where

$$\mathbb{S}_3 = \{(i,k) \mid i \in \{2, 3, \ldots, N_{\mathrm{RA}}\}, k \in \{2, 3, \ldots, i\}\}.$$

We define $r_k(i)$ as the probability that the tagged transmission attempt is successful during the random access mode. We also define $\tilde{r}_{k,l}(i, j)$ as the probability that the transmission fails and that the next state of the Markov chain is (j, l). It is necessary for a request-holding meter which failed in its transmission to wait for another T. Therefore, we have the following equation for $(i, k) \in \mathbb{S}_3$.

$$F^{(i,k)*}(s) = r_k(i) d^*(s) + \exp\{-sT\} \sum_{j=2}^{N_{\mathrm{RA}}} \sum_{l=2}^{j} \tilde{r}_{k,l}(i, j) F^{(j,l)*}(s). \tag{3}$$

Here, $r_k(i)$ is given by

$$r_k(i) = \begin{cases} (1-a)a^{k-1}, & L = 1, \\ (1-a) \displaystyle\sum_{i_1=0}^{k-1} S_{\mathrm{OFF}}(k, i_1) \left(\frac{L-1}{L}\right)^{k-i_1-1}, & L \geq 2. \end{cases}$$

Further, we have

$$\begin{aligned} \tilde{r}_{k,l}(i, j) = (1-a) \sum_{i_1=0}^{k-1} \sum_{i_2=0}^{i-k} \sum_{k_1=1}^{k-i_1-1} \sum_{k_2=0}^{k-i_1-1-k_1} & S_{\mathrm{OFF}}(k-1, i_1) S_{\mathrm{OFF}}(i-k, i_2) \\ & \cdot S_{\mathrm{ON}}(N_{\mathrm{RA}} - i, j + i_1 + i_2 - i) \cdot \frac{L \binom{k-i_1-1}{k_1} g_{k-i_1-1-k_1, k_2}^{(L-1)}}{L^{k-i_1}} \\ & \cdot B_{l-1-k_1-k_2}(j - 1 - k_1 - k_2), \end{aligned}$$

where $g_{k,l}^{(L)}$ denotes the number of combinations in which l out of k request-holding meters fail in their transmissions in L time slots. See Appendix A for the derivation of $g_{k,l}^{(L)}$. We then rewrite (3) as

$$\hat{\boldsymbol{F}}^*(s) = d^*(s)\boldsymbol{r} + \exp\{-sT\}\tilde{\boldsymbol{R}}\hat{\boldsymbol{F}}^*(s), \tag{4}$$

where

$$\hat{\boldsymbol{F}}^*(s) := (F^{(i,k)*}(s))_{(i,k)\in\mathbb{S}_3},\ \boldsymbol{r} := (r_k(i))_{(i,k)\in\mathbb{S}_3},\ \tilde{\boldsymbol{R}} := (\tilde{r}_{k,l}(i,j))_{(i,k)\times(j,l)\in\mathbb{S}_3^2}.$$

Since $\exp\{-sT\}\tilde{\boldsymbol{R}}$ is an inferior probability matrix and irreducible, $\hat{\boldsymbol{F}}^*(s)$ is given by

$$\hat{\boldsymbol{F}}^*(s) = d^*(s)\left(\boldsymbol{I} - \exp\{-sT\}\tilde{\boldsymbol{R}}\right)^{-1}\boldsymbol{r}.$$

Thus, we have

$$\hat{\boldsymbol{F}}^*(0) = d^*(0)\left(\boldsymbol{I} - \tilde{\boldsymbol{R}}\right)^{-1}\boldsymbol{r}.$$

We differentiate the both sides of (4) with respect to s, which results in

$$\hat{\boldsymbol{F}}^{*\prime}(s) = d^{*\prime}(s)\boldsymbol{r} + \exp\{-sT\}\tilde{\boldsymbol{R}}\left(\hat{\boldsymbol{F}}^{*\prime}(s) - T\hat{\boldsymbol{F}}^*(s)\right). \tag{5}$$

Differentiating the above equation yields

$$\begin{aligned}\hat{\boldsymbol{F}}^{*\prime\prime}(s) = {} & d^{*\prime\prime}(s)\boldsymbol{r} + \exp\{-sT\}\tilde{\boldsymbol{R}}\left(\hat{\boldsymbol{F}}^{*\prime\prime}(s) - T\hat{\boldsymbol{F}}^{*\prime}(s)\right) \\ & -T\exp\{-sT\}\tilde{\boldsymbol{R}}\left(\hat{\boldsymbol{F}}^{*\prime}(s) - T\hat{\boldsymbol{F}}^*(s)\right).\end{aligned} \tag{6}$$

From (5) and (6), we have the following equations.

$$\begin{aligned}\hat{\boldsymbol{F}}^{*\prime}(s) &= \left(\boldsymbol{I} - \exp\{-sT\}\tilde{\boldsymbol{R}}\right)^{-1}\left(d^{*\prime}(s)\boldsymbol{r} - T\exp\{-sT\}\tilde{\boldsymbol{R}}\hat{\boldsymbol{F}}^*(s)\right), \\ \hat{\boldsymbol{F}}^{*\prime\prime}(s) &= \left(\boldsymbol{I} - \exp\{-sT\}\tilde{\boldsymbol{R}}\right)^{-1} \\ &\quad \cdot\left(d^{*\prime\prime}(s)\boldsymbol{r} - 2T\exp\{-sT\}\tilde{\boldsymbol{R}}\hat{\boldsymbol{F}}^{*\prime}(s) + T^2\exp\{-sT\}\tilde{\boldsymbol{R}}\hat{\boldsymbol{F}}^*(s)\right).\end{aligned}$$

As a result, we obtain

Table 1 Parameter setting

Number of smart meters N_{RA}	5-40
Length of a check cycle T (sec)	18
Length of a polling mode T_{P} (sec)	10.5
Length of a random access mode T_{RA} (sec)	7.5
Mean duration of on-state periods T/a (sec)	45-7200
Mean duration of off-state periods T/b (sec)	180,900,1800,18000
Mean interval of generations of transmission attempts T/λ (sec)	36
Time for one hop t_{H} (msec)	75
Maximum number of hops $H_{\max}$	2
Number of time slots of a random access mode L	5
Length of a time slot of a random access mode t_{RA} (sec)	1.5

$$
\begin{aligned}
\hat{\boldsymbol{F}}^{*\prime}(0) &= \left(\boldsymbol{I}-\tilde{\boldsymbol{R}}\right)^{-1}\left(d^{*\prime}(0)\boldsymbol{r}-T\tilde{\boldsymbol{R}}\hat{\boldsymbol{F}}^{*}(0)\right)\\
&= -\left(\boldsymbol{I}-\tilde{\boldsymbol{R}}\right)^{-1}\left(\frac{L+1}{2}t_{\mathrm{RA}}\boldsymbol{r}+T\tilde{\boldsymbol{R}}\hat{\boldsymbol{F}}^{*}(0)\right),\\
\hat{\boldsymbol{F}}^{*\prime\prime}(0) &= \left(\boldsymbol{I}-\tilde{\boldsymbol{R}}\right)^{-1}\left(d^{*\prime\prime}(0)\boldsymbol{r}-2T\tilde{\boldsymbol{R}}\hat{\boldsymbol{F}}^{*\prime}(0)+T^{2}\tilde{\boldsymbol{R}}\hat{\boldsymbol{F}}^{*}(0)\right)\\
&= \left(\boldsymbol{I}-\tilde{\boldsymbol{R}}\right)^{-1}\left\{\frac{(L+1)(2L+1)}{6}t_{\mathrm{RA}}^{2}\boldsymbol{r}-2T\tilde{\boldsymbol{R}}\hat{\boldsymbol{F}}^{*\prime}(0)+T^{2}\tilde{\boldsymbol{R}}\hat{\boldsymbol{F}}^{*}(0)\right\}.
\end{aligned}
$$

In numerical examples, we show the coefficient of variation (CV) of transmission delay defined as

$$
\mathrm{CV}=\frac{\sqrt{\mathrm{Var}\left[D\mid\mathcal{G}\right]}}{\mathrm{E}\left[D\mid\mathcal{G}\right]}.
$$

4 Numerical Example

In this section, we show some numerical examples of the performance measures derived in the previous section. We then investigate the effects of the durations of on-state and off-state periods and their ratio on the performance measures.

4.1 Parameter Setting

In numerical experiments, we use the basic parameter values shown in Table 1.

4.2 *Effect of the Ratio of the Durations of On-state and Off-state Periods*

In this subsection, we set $T/b = 1800$ (sec) and change $T/a : T/b$ in order to investigate the effect of the ratio of the duration of on-state periods to that of off-state periods on the performance measures.

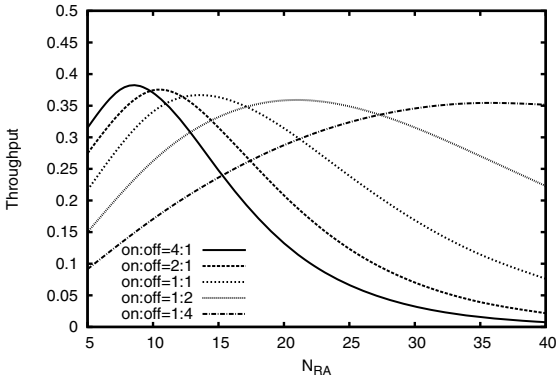


Fig. 2 Throughput vs. number of smart meters. ($T/b = 1800$)

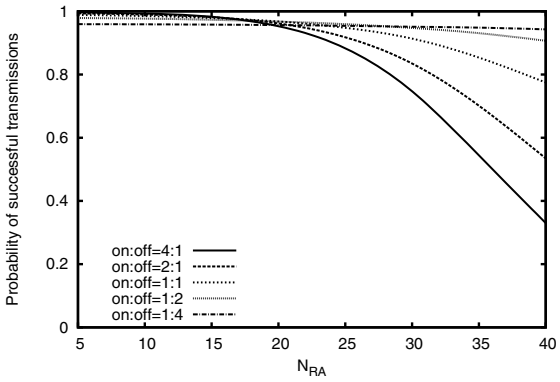


Fig. 3 Probability of successful transmissions vs. number of smart meters. ($T/b = 1800$)

Figure 2 represents the throughput against N_{RA} in cases of $(a/T)/(b/T) = 4, 2, 1, 1/2$, and $1/4$. In this figure, the number of smart meters at which the throughput takes the maximum value grows larger as the ratio decreases although the maximum throughput decreases. the number of transmissions per one time slot and can improve the throughput. Figure 3 shows the probability of successful transmissions against N_{RA} in cases of $(a/T)/(b/T) = 4, 2, 1, 1/2$, and $1/4$. It is observed from

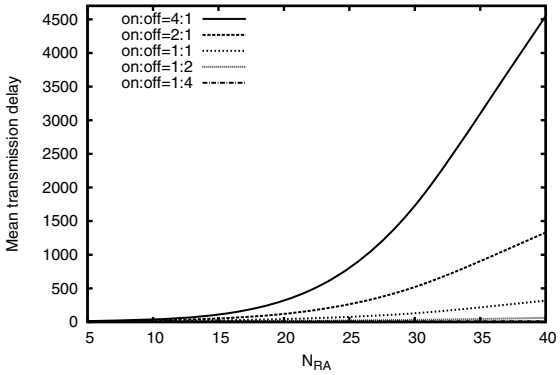


Fig. 4 Mean transmission delay vs. number of smart meters. ($T/b = 1800$)

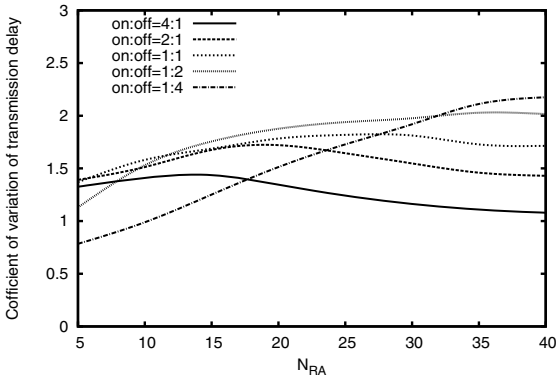


Fig. 5 Coefficient of variation of transmission delay vs. number of smart meters. ($T/b = 1800$)

this figure that the lower the ratio gets, the lower the deterioration rate of the probability of successful transmissions becomes. Figure 4 illustrates the mean transmission delay against N_{RA} in cases of $(a/T)/(b/T) = 4, 2, 1, 1/2$, and $1/4$. We observe from this figure that the increase rate of the mean transmission delay is low if the ratio is low. Figure 5 indicates the coefficient of variation of the transmission delay against N_{RA} in cases of $(a/T)/(b/T) = 4, 2, 1, 1/2$, and $1/4$. It is found from this figure that the number of smart meters which maximizes the coefficient of variation of the transmission delay becomes larger with the decrease of the ratio. These results imply that if the number of smart meters is large, the communication performance can be improved by a lower ratio of the duration of on-state periods to that of off-state periods. This is because the number of request-holding meters gets smaller and because collision probability in transmissions decreases when the ratio is low.

4.3 *Effect of the Durations of On-state and Off-state Periods*

In this subsection, we fix $(T/a)/(T/b) = 1/4$ and change T/b for the purpose of investigating the effect of durations of on-state and off-state periods on the performance measures. Note here that switching intervals between on-state and off-state are short when T/b is small.

Figure 6 shows throughput against N_{RA}. In this figure, the throughput is low when switching intervals are short to some extent, but switching intervals hardly affect the throughput if they are very long. Figure 7 illustrates the probability of successful transmissions against N_{RA}. It is observed from this figure that the probability of successful transmissions decreases as switching intervals get shorter. Figure 8 indicates the mean transmission delay against N_{RA}. It is found from this figure that

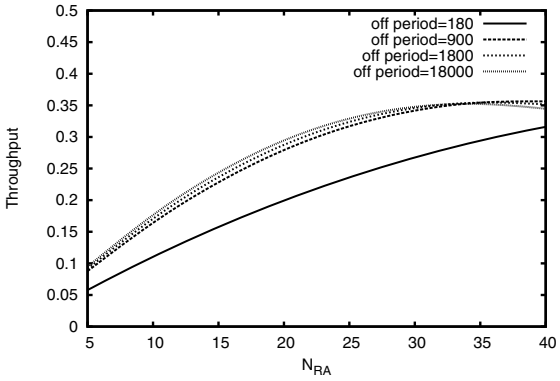


Fig. 6 Throughput vs. number of smart meters. $(T/a : T/b = 1 : 4)$

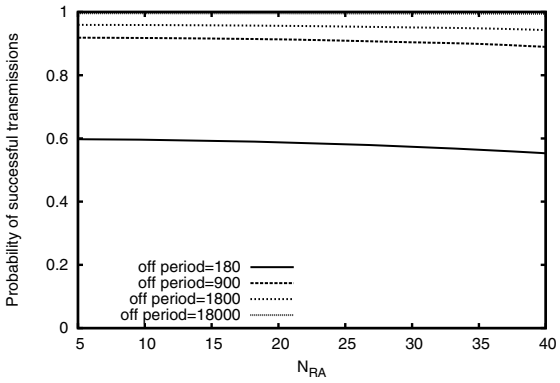


Fig. 7 Probability of successful transmissions vs. number of smart meters. $(T/a : T/b = 1 : 4)$

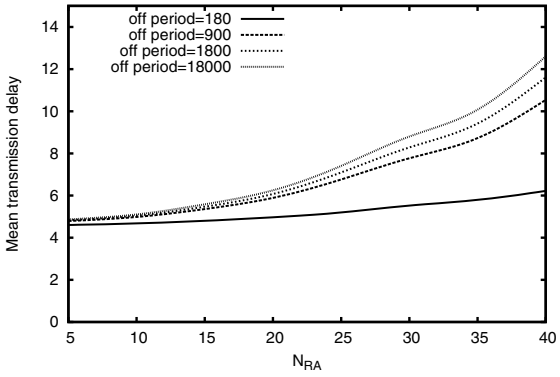


Fig. 8 Mean transmission delay vs. number of smart meters. ($T/a : T/b = 1 : 4$)

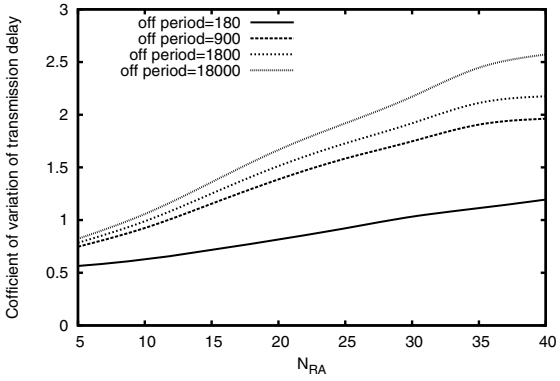


Fig. 9 Coefficient of variation of transmission delay vs. number of smart meters. ($T/a : T/b = 1 : 4$)

the increase rate of the mean transmission delay is low when switching intervals are short. Figure 9 represents the coefficient of variation of the transmission delay against N_{RA}. In , the shorter switching intervals get, the larger the coefficient of variation of the transmission delay becomes. Therefore, we can claim that the throughput is improved and the mean transmission delay is restrained although the probability of successful transmissions decreases when switching between on-state and off-state is frequent. This is because frequent switching increases the number of canceled transmission and generated transmission attempts.

5 Conclusion

In this paper, we evaluated the performance of AMI taking the mobile environment into consideration. We constructed a discrete time Markov chain whose states are defined by the numbers of on-state meters and request-holding meters, deriving the throughput, the probability of successful transmissions, the mean transmission delay, and the coefficient of variation of the transmission delay.

From numerical examples, we found that if the number of smart meters is large, the communication performance can be improved by a lower ratio of the duration of on-state periods to that of off-state periods although the coefficient of variation of the transmission delay becomes larger. It was also revealed that the probability of successful transmissions decreases, while the throughput is improved and the mean transmission delay is restrained when switching between on-state and off-state is frequent.

Electricity consumption of smart meters is much more than that of conventional electrical meters mainly because the former are equipped with communication functions. Further, in wireless networks, collisions among transmission attempts occur frequently and more electricity is consumed when the number of smart meters is large. We can make wireless networks scalable to the increase of the number of smart meters and suppress electricity costs by setting the probability of retransmissions extremely low, while the mean transmission delay may increase. Our future work is to consider a trade-off between electricity costs and the performance in AMI networks.

Acknowledgments This work was supported by JSPS KAKENHI Grant Numbers 26280113 and 15K12152.

References

1. http://sangyo.jp/ri/sg/na/article/20110408.html (in Japanese)
2. Fatemieh, O., Chandra, R., Gunter, C.A.: Low cost and secure smart meter communications using the TV white spaces. In: IEEE International Symposium on Resilient Control Systems, ISRCS 2010, August 2010
3. http://www.sankei.com/economy/news/141219/ecn1412190025-n.html (in Japanese)
4. http://www.hitachi.co.jp/New/cnews/month/2014/12/1219.html (in Japanese)
5. NEC's Approach towards Advanced Metering Infrastructure (AMI). http://www.nec.com/en/global/techrep/journal/g12/n01/pdf/120119.pdf
6. Romano, F., Zoppi, L.: A combined reservation random access polling protocol for voice-data transmissions in a wireless packet network. IEEE Transactions on Vehicular Technology **48**(2), 652–662 (1999)
7. Li, C., Li, J., Cai, X.: Performance evaluation of IEEE 802.11 WLAN - high speed packet wireless data network for supporting voice service. In: IEEE Wireless Communications and Networking Conference (2004)
8. Rubin, I., Louie, M.Y.: Hybrid TDMA/Random-Access Scheme for Multiple-Access Communication Networks. Computers & Electrical Engineering **10**(3), 159–182 (1983)
9. Davis, D.H., Gronemeyer, S.A.: Performance of Slotted ALOHA Random Access with Delay Capture and Randomized Time of Arrival. IEEE Transactions on Communications **COM–28**(5), 703–710 (1980)

10. Sen, S., Dorsey, D.J., Guerin, R., Chiang, M.: Analysis of slotted ALOHA with multipacket messages in clustered surveillance networks. In: IEEE Military Communications Conference, October 2012
11. Yang, Y., Yum, T.P.: Delay Distributions of Slotted ALOHA and CSMA. IEEE Transactions on Communications **51**(11), 1846–1857 (2003)
12. Ogasawara, S.: Haiburiddo tsushinhoushiki o yusuru AMI no surimoderu to seinoukaiseki (in Japanese), Graduation Thesis, Kyoto University, January 2014

A Appendix: Derivation of $q_{k,l}$ and $g_{k,l}^{(L)}$

Firstly, we derive the number of combinations, $g_{k,l}^{(L)}$, in which l out of k request-holding meters fail in their transmissions in L time slots. If two or more smart meters choose the same time slot, then all the transmissions in the time slot fail. Thus, $k-l$ request-holding meters choose different time slots in this event. On the other hand, each of $L-k-l$ time slots is chosen by at least two out of l request-holding meters. If $k-l>L$ or $k-l=L, l>0$, there are no combinations of $k-l$ request-holding meters succeeding in their transmissions. We now consider the case of $k-l<L$. If $l=0$, the number of combination of choosing successful time slots is $\binom{L}{k-l}$, and that of k request-holding meters choosing those time slots is $k!/l!$. The case of $l=1$ does not satisfy the given condition because the transmission necessarily succeeds. We next consider the case of $l \geq 2$. Let m denote the number of time slots chosen by two or more request-holding meters, we then have $1 \leq m \leq \min(L-k+l, [l/2])$. Therefore, $g_{k,l}^{(L)}$ is given as follows.

$$g_{k,l}^{(L)} = \begin{cases} 0, & \begin{cases} k<l, \\ k-l>L, \\ k-l=L,\ l>0, \\ k-l<L,\ l=1, \end{cases} \\ k!, & k-l=L,\ l=0, \\ \binom{L}{k-l}\dfrac{k!}{l!}, & k-l<L,\ l=0, \\ \displaystyle\sum_{m=1}^{\min(L-k+l,[l/2])} \dfrac{\binom{L}{k-l}k!\binom{L-k+l}{m}h(l,m)}{l!}, & \text{otherwise}, \end{cases}$$

where $h(l,m)$ denotes the number of combinations in which each of m out of l time slots is chosen by at least two request-holding meters. Further, the number of combinations of k request-holding meters choosing L time slots is L^k, so we obtain

$$q_{k,l} = \begin{cases} 0, & \begin{cases} k < l, \\ k - l > L, \\ k - l = L,\ l > 0, \\ k - l < L,\ l = 1, \end{cases} \\ \dfrac{k!}{L^k}, & k - l = L,\ l = 0, \\ \dbinom{L}{k-l}\dfrac{k!}{l!L^k}, & k - l < L,\ l = 0, \\ \displaystyle\sum_{m=1}^{\min(L-k+l,[l/2])} \dfrac{\binom{L}{k-l}k!\binom{L-k+l}{m}h(l,m)}{l!L^k}, & \text{otherwise.} \end{cases}$$

We then derive $h(l, m)$. In the case of $2 \leq l, m = 1$, there exists only one combination because of there being one time slot. If $l = 1$, this event is infeasible since no time slots are chosen by two or more request-holding meters. Moreover, the case of $l < 2m$ does not meet the given condition. In what follows, we consider the case of $2 \leq l, 2m \leq l$. There are m^l combinations of l request-holding meters choosing m time slots. We suppose that n_1 out of m time slots are not chosen by any request-holding meters, then the number of combinations of choosing those time slots is $\binom{m}{n_1}$. If each of n_2 time slots out of $m - n_1$ is chosen by only one request-holding meter, we obtain the number of combinations of choosing the time slots as $\binom{m-n_1}{n_2}$. There exist $j!/n_2!$ combinations for n_2 request-holding meters, and the number of combinations of $l - n_2$ request-holding meters choosing $m - n_1 - n_2$ time slots is given by $h(l - n_2, m - n_1 - n_2)$. Here, $1 \leq n_1 + n_2 \leq m - 1$ holds for n_1 and n_2, we then have

$$h(l,m) = m^l - \sum_{n_1=0}^{m-1} \sum_{n_2=\max(0,1-n_1)}^{m-1-n_1} \binom{m}{n_1}\binom{m-n_1}{n_2}\frac{j!}{n_2!}h(l-n_2, m-n_1-n_2).$$

As a result, $h(l, m)$ is given by the following recurrence relation.

$$h(l,m) = \begin{cases} 0, & 0 \leq l < 2m, \\ 1, & 2 \leq l,\ m = 1, \\ m^l - \displaystyle\sum_{n_1=0}^{m-1} \sum_{n_2=\max(0,1-n_1)}^{m-1-n_1} \dfrac{m!l!h(l-n_2, m-n_1-n_2)}{n_1!n_2!(m-n_1-n_2)!(l-n_2)!}, & \text{otherwise.} \end{cases}$$

Retrial Queue for Cloud Systems with Separated Processing and Storage Units

Tuan Phung-Duc

Abstract This paper considers a retrial queueing model for cloud computing systems where the processing unit (server) and the storage unit (buffer) are separated. Jobs that cannot occupy the server upon arrival are stored in the buffer from which they are sent to the server after some random time. After completing a service the server stays idle for a while waiting for either a new job or a job from the buffer. After the idle period, the server starts searching for a job from the buffer. We assume that the search time cannot be disregarded during which the server cannot serve a job. We model this system using a retrial queue with search for customers from the orbit and obtain an explicit solution in terms of partial generating functions. We present a recursive scheme for computing the stationary probability of all the states.

Keywords Retrial queue · Search time · Two-way communication · Cloud systems

1 Introduction

Retrial queueing systems are ubiquitous in our daily life. The are characterized by the fact that a customer who cannot receive service immediately upon arrival joins a virtual orbit and repeats its attempt after some random time. Almost all the papers in the retrial queueing literature assume that the server only waits for either a new customer or a repeated one from the orbit [9]. However, there are some situations in which the server has some initiative searching for blocked customers. We assume that after a service the server stays idle for a while and starts searching for blocked customers. In the idle time, if either a new customer or a repeated customer comes, it receives the service immediately. After the idle time, the server performs a search whose duration follows the exponential distribution. During the searching time, the server cannot serve a customer, i.e., customers that arrive during the searching time

T. Phung-Duc (✉)
Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Ookayama, Meguro-ku, Tokyo, Japan
e-mail: tuan@is.titech.ac.jp

T.V. Do et al. (eds.), *Queueing Theory and Network Applications*,
Advances in Intelligent Systems and Computing 383,
DOI: 10.1007/978-3-319-22267-7_13

of the server join the orbit. After the searching time the server gets a customer from the orbit if any, otherwise it stays idle again.

The model is motivated from cloud computing systems where the processing unit and the storage unit are separated. The processing unit has the capacity to serve only one job at a time. Jobs that arrive when the server is busy are stored in a buffer from which they are sent to the server. On completing a service the server stays idle for a while and then picks a job from the buffer which takes some time. We refer this time to as a search time. This system can be modeled using a retrial queue with search for customers for which we obtain an explicit solution. Analytical solutions for some Markovian retrial queues could be found in [11, 12, 14].

Some closely related works are as follows. Artalejo et al. [3] consider a retrial queue with search for customers from orbit. In particular, after completing a service, the server either immediately picks a customer from the orbit if any with probability p or stays idle with probability $1 - p$. This is similar to our model in the sense that the server picks a customer from the orbit. However there is no idle time and searching time (the searching time is zero) in this model [3]. Dudin et al. [8] consider the same model as in [3] with BMAP input and search for customers. However, the search mechanism is started just after the service completion. Some other extensions are found in [6, 7]. Artalejo and Phung-Duc [4, 5] consider a model with two-way communication where after the idle time the server initiates an outgoing call whose duration is exponentially distributed. This can be considered as the searching time in our model. However, after an outgoing call, the server stays idle, i.e., no customer from the orbit is picked up. In all the works above, the idle time and the searching time are separately considered. This paper is the first which proposes a search mechanism which is initiated after some idle time of the server. Other related works are due to Artalejo and Gomez-Corrall [1] and Artalejo and Atencia [2] where the retrial rate is a linear function of the number of customers in the orbit.

The rest of the paper is organized as follows. Section 2 describes the queueing model in details while Section 3 is devoted to the analysis of the model. In section 4, we present a special case where the searching time is negligible. Concluding remarks are presented in section 5.

2 Model

Incoming jobs arrive at the server according to a Poisson process with rate λ. Service time of incoming customers follows the exponential distribution with mean $1/\nu_1$. After the completion of a service the server stays idle for an exponentially distributed time with mean $1/\alpha$. During this idle time, an arriving customer (either a new customer or a repeated one) is immediately served. After the idle time, the server starts searching for a customer in the orbit. The searching time follows the exponential distribution with mean $1/\nu_2$. Arriving customers who see the server busy (serving a customer or searching) join the orbit from which each customer retries to enter the server after some exponentially distributed time with mean $1/\mu$. To the best of our knowledge, this model has not been analyzed in the literature.

3 Analysis

Let $C(t)$ denote the state of the server at time $t \geq 0$.

$$C(t) = \begin{cases} 0, & \text{the server is idle,} \\ 1, & \text{the server is serving a job,} \\ 2, & \text{the server is searching for a customer.} \end{cases}$$

Let $N(t)$ denote the number of customers in the orbit at time $t \geq 0$. We then have the fact that $\{X(t) = (C(t), N(t)), t \geq 0\}$ forms a Markov chain on the state space $\mathcal{S} = \{0, 1, 2\} \times \{0, 1, 2, \dots\}$. See Figure 1 for the transitions among states.

We assume that the system is stable, i.e., the stationary distribution exists. The necessary and sufficient condition for the stability is $\lambda < \nu_1$ which will be obtained later in the analysis.

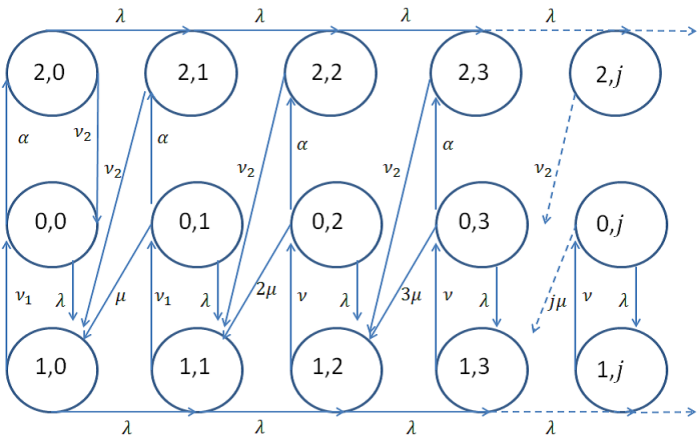


Fig. 1 Transitions among states

Letting $\pi_{i,j} = \lim_{t\to\infty} \mathrm{P}(C(t) = i, N(t) = j)$, the balance equations for states (i, j) are given as follows.

$$(\lambda + \alpha)\pi_{0,0} = \nu_1\pi_{1,0} + \nu_2\pi_{2,0}, \tag{1}$$

$$(\lambda + \alpha + j\mu)\pi_{0,j} = \nu_1\pi_{1,j}, \quad j \geq 1, \tag{2}$$

$$(\lambda + \nu_1)\pi_{1,j} = (j+1)\mu\pi_{0,j+1} + \nu_2\pi_{2,j+1} + \lambda\pi_{1,j-1} + \lambda\pi_{0,j}, \quad j \geq 0,$$

$$(\lambda + \nu_2)\pi_{2,j} = \alpha\pi_{0,j} + \lambda\pi_{2,j-1}, \quad j \geq 0, \tag{3}$$

where $\pi_{i,-1} = 0$ $(i = 1, 2)$. Let $\Pi_i(z)$ denote the generating function of $\pi_{i,j}$, i.e. $\Pi_i(z) = \sum_{j=0}^{\infty} \pi_{i,j} z^j$ $(i = 0, 1, 2)$. Transforming the above balance equations to generating functions we obtain,

$$(\lambda+\alpha)\Pi_0(z)+\mu z\Pi_0'(z)=\nu_1\Pi_1(z)+\nu_2\pi_{2,0}, \tag{4}$$

$$(\lambda+\nu_1)\Pi_1(z)=\mu\Pi_0'(z)+\frac{\nu_2}{z}(\Pi_2(z)-\pi_{2,0}) + \lambda z\Pi_1(z)+\lambda\Pi_0(z), \tag{5}$$

$$(\lambda+\nu_2)\Pi_2(z)=\alpha\Pi_0(z)+\lambda z\Pi_2(z). \tag{6}$$

Summing the above equations and arranging the result yields

$$\lambda(\Pi_1(z)+\Pi_2(z))=\mu\Pi_0'(z)+\frac{\nu_2(\Pi_2(z)-\pi_{2,0})}{z}. \tag{7}$$

This equation represents the balance between the flows coming into and out the orbit. From (4) and (6), we obtain

$$\Pi_1(z)=\frac{(\lambda+\alpha)\Pi_0(z)+\mu z\Pi_0'(z)-\nu_2\pi_{2,0}}{\nu_1}, \tag{8}$$

$$\Pi_2(z)=\frac{\alpha\Pi_0(z)}{\lambda+\nu_2-\lambda z}. \tag{9}$$

Substituting these two expressions into the orbit balance equation (7) and arranging the result yields

$$\Pi_0'(z)=A(z)\Pi_0(z)+B(z), \tag{10}$$

where

$$A(z)=\frac{\frac{\lambda(\lambda+\alpha)}{\nu_1}+\frac{\alpha(\lambda-\nu_2/z)}{\lambda+\nu_2-\lambda z}}{\mu\left(1-\frac{\lambda z}{\nu_1}\right)}, \qquad B(z)=\frac{\pi_{2,0}\nu_2}{\mu z}.$$

We decompose $A(z)$ as follows.

$$A(z)=\frac{a}{z}+\frac{b}{1-\frac{\lambda z}{\nu_1}}+\frac{c}{1-\frac{\lambda z}{\lambda+\nu_2}},$$

where a, b and c are given by

$$a=-\frac{\alpha\nu_2}{\mu(\lambda+\nu_2)}, \quad b=\frac{\lambda^2(\lambda+\alpha+\nu_2-\nu_1)}{\mu\nu_1(\lambda+\nu_2-\nu_1)}, \quad c=\frac{\lambda^2\alpha\nu_1}{(\lambda+\nu_2)^2\mu(\nu_1-\lambda-\nu_2)}.$$

We first solve the non-homogeneous differential equation

$$\Pi_0'(z)=A(z)\Pi_0(z),$$

which is transformed to

$$\frac{\Pi_0'(z)}{\Pi_0(z)} = \frac{a}{z} + \frac{b}{1-\frac{\lambda z}{\nu_1}} + \frac{c}{1-\frac{\lambda z}{\lambda+\nu_2}}.$$

The solution of this differential equation is given by

$$\Pi_0(z) = Cz^a \left(\frac{\nu_1-\lambda}{\nu_1-\lambda z}\right)^{\frac{b\nu_1}{\lambda}} \left(\frac{\nu_2}{\lambda+\nu_2-\lambda z}\right)^{\frac{c(\lambda+\nu_2)}{\lambda}},$$

where C is a constant number. As usual, we find the solution for our original differential equation (10) in the following form.

$$\Pi_0(z) = C(z)z^a \left(\frac{\nu_1-\lambda}{\nu_1-\lambda z}\right)^{\frac{b\nu_1}{\lambda}} \left(\frac{\nu_2}{\lambda+\nu_2-\lambda z}\right)^{\frac{c(\lambda+\nu_2)}{\lambda}},$$

where $C(z)$ is an unknown function. Substituting this into the original differential equation (10) yields

$$C'(z)z^a \left(\frac{\nu_1-\lambda}{\nu_1-\lambda z}\right)^{\frac{b\nu_1}{\lambda}} \left(\frac{\nu_2}{\lambda+\nu_2-\lambda z}\right)^{\frac{c(\lambda+\nu_2)}{\lambda}} = \frac{\pi_{2,0}\nu_2}{\mu z},$$

or equivalently

$$C'(z) = \frac{\pi_{2,0}\nu_2}{\mu} z^{-(a+1)} \left(\frac{\nu_1-\lambda}{\nu_1-\lambda z}\right)^{-\frac{b\nu_1}{\lambda}} \left(\frac{\nu_2}{\lambda+\nu_2-\lambda z}\right)^{-\frac{c(\lambda+\nu_2)}{\lambda}}.$$

Therefore, we have

$$C(z) = C_0 - \frac{\pi_{2,0}\nu_2}{\mu} \int_z^1 u^{-(a+1)} \left(\frac{\nu_1-\lambda}{\nu_1-\lambda u}\right)^{-\frac{b\nu_1}{\lambda}} \left(\frac{\nu_2}{\lambda+\nu_2-\lambda u}\right)^{-\frac{c(\lambda+\nu_2)}{\lambda}} du,$$

where C_0 is a constant number. Because $\Pi_0(z)$ is analytic at $z = 0$ and $a < 0$, we must have $C(0) = 0$ implying that

$$C_0 = \frac{\pi_{2,0}\nu_2}{\mu} \int_0^1 u^{-(a+1)} \left(\frac{\nu_1-\lambda}{\nu_1-\lambda u}\right)^{-\frac{b\nu_1}{\lambda}} \left(\frac{\nu_2}{\lambda+\nu_2-\lambda u}\right)^{-\frac{c(\lambda+\nu_2)}{\lambda}} du.$$

The final solution for $\Pi_0(z)$ is given by

$$\Pi_0(z) = \frac{\pi_{2,0}\nu_2}{\mu} z^a \left(\frac{\nu_1-\lambda}{\nu_1-\lambda z}\right)^{\frac{b\nu_1}{\lambda}} \left(\frac{\nu_2}{\lambda+\nu_2-\lambda z}\right)^{\frac{c(\lambda+\nu_2)}{\lambda}} \times \int_0^z u^{-(a+1)} \left(\frac{\nu_1-\lambda}{\nu_1-\lambda u}\right)^{-\frac{b\nu_1}{\lambda}} \left(\frac{\nu_2}{\lambda+\nu_2-\lambda u}\right)^{-\frac{c(\lambda+\nu_2)}{\lambda}} du. \quad (11)$$

From (7), (9) and (10), we obtain

$$\Pi_1(1) + \Pi_2(1) = \left(\frac{\mu}{\lambda}A(1) + \frac{\alpha}{\lambda}\right)\Pi_0(1). \tag{12}$$

We also have the normalization condition:

$$\Pi_0(1) + \Pi_1(1) + \Pi_2(1) = 1. \tag{13}$$

From (12) and (13), we obtain

$$\Pi_0(1) = \frac{\nu_2(1-\frac{\lambda}{\nu_1})}{\alpha+\nu_2},$$

where the expression of $A(1)$ in terms of given parameters is used.

It follows from (8) and (9) that

$$\Pi_2(1) = \frac{\alpha(1-\frac{\lambda}{\nu_1})}{\alpha+\nu_2}, \qquad \Pi_1(1) = \frac{\lambda}{\nu_1}.$$

Therefore, from the expression for $\Pi_0(z)$, we obtain the expression for $\pi_{2,0}$ as follows.

$$\pi_{2,0} = \frac{\mu(1-\frac{\lambda}{\nu_1})}{(\lambda+\nu_2)\int_0^1 u^{-(a+1)}\left(\frac{\nu_1-\lambda}{\nu_1-\lambda u}\right)^{-\frac{b\nu_1}{\lambda}}\left(\frac{\nu_2}{\lambda+\nu_2-\lambda u}\right)^{-\frac{c(\lambda+\nu_2)}{\lambda}} du}. \tag{14}$$

From this expression, we obtain the fact that the stability condition for the model is $\lambda < \nu_1$.

3.1 Recursive Formulae

Now, we are going to derive a recursive scheme for the stationary distribution. From the orbit balance equation, we obtain

$$\lambda(\pi_{1,j} + \pi_{2,j}) = (j+1)\mu\pi_{0,j+1} + \nu_2\pi_{2,j+1}.$$

From this equation and (3) with $j := j+1$, we obtain,

$$\lambda\pi_{1,j} + \frac{\lambda^2}{\lambda+\nu_2}\pi_{2,j} = \left((j+1)\mu + \frac{\alpha\nu_2}{\lambda+\nu_2}\right)\pi_{0,j+1}.$$

Therefore, we have the following recursive scheme for the stationary distribution.

$$\pi_{0,j} = \frac{\lambda[(\lambda+\nu_2)\pi_{1,j-1} + \lambda\pi_{2,j-1}]}{j\mu(\lambda+\nu_2)+\alpha\nu_2}, \qquad j \geq 1,$$
$$\pi_{1,j} = \frac{(\lambda+\alpha+j\mu)\pi_{0,j}}{\nu_1}, \qquad j \geq 1,$$
$$\pi_{2,j} = \frac{\alpha\pi_{0,j} + \lambda\pi_{2,j-1}}{\lambda+\nu_2}, \qquad j \geq 1,$$

where $\pi_{0,0}$, $\pi_{1,0}$ and $\pi_{2,0}$ are given in advance. In particular, $\pi_{2,0}$ is obtained by (14) and $\pi_{0,0}$ is obtained from (3) with $j = 0$ while $\pi_{1,0}$ is obtained by summing up (1) and (3) with $j = 0$, i.e., $\pi_{1,0} = \lambda(\pi_{0,0} + \pi_{2,0})/\nu_1$. It should be noted that the second and the third equations follow from (2) and (3), respectively.

Remark 1 This recursive formulae allow to calculate any probability $\pi_{i,j}$. Furthermore, the recursive scheme can be implemented in both numerical and symbolic manners.

Remark 2 Taking the derivatives at $z = 1$ for the differential equation (10) we can obtain $\Pi_0^{(n)}(1)$ for any n. Since $\Pi_1(z)$ and $\Pi_2(z)$ are expressed in terms of $\Pi_0(z)$, we can also calculate $\Pi_1^{(n)}(1)$ and $\Pi_2^{(n)}(1)$ for any n.

4 Limiting Case

We investigate the case where $\nu_2 \to \infty$ meaning that a call in the orbit is picked to the server after an exponentially distributed idle time with mean $1/\alpha$. This is equivalent to the linear retrial rate policy presented in [1]

In particular, we observe that when $\nu_2 \to \infty$,

$$a = -\frac{\alpha}{\mu}, \qquad b = \frac{\lambda^2}{\mu\nu_1}, \qquad c = 0.$$

Furthermore,

$$\lim_{\nu_2\to\infty} \Pi_2(z) = 0,$$

meaning that the searching states do not exist. We have

$$\lim_{\nu_2\to\infty} \frac{\nu_2\pi_{2,0}}{\mu} = \lim_{\nu_2\to\infty} \frac{\nu_2(1-\frac{\lambda}{\nu_1})}{(\lambda+\nu_2)\int_0^1 u^{-(a+1)}\left(\frac{\nu_1-\lambda}{\nu_1-\lambda u}\right)^{-\frac{b\nu_1}{\lambda}}\left(\frac{\nu_2}{\lambda+\nu_2-\lambda u}\right)^{-\frac{c(\lambda+\nu_2)}{\lambda}} du}$$
$$= \frac{1-\frac{\lambda}{\nu_1}}{\int_0^1 u^{-(a+1)}\left(\frac{\nu_1-\lambda}{\nu_1-\lambda u}\right)^{-\frac{b\nu_1}{\lambda}} du}.$$

Thus, it follows from (11) that

$$\Pi_0(z) = \left(1 - \frac{\lambda}{\nu_1}\right) z^{-\frac{\alpha}{\mu}} \left(\frac{\nu_1 - \lambda}{\nu_1 - \lambda z}\right)^{\frac{\lambda}{\mu}} \frac{\int_0^z u^{\frac{\alpha}{\mu}-1} \left(\frac{\nu_1-\lambda}{\nu_1-\lambda u}\right)^{\frac{\lambda}{\mu}} du}{\int_0^1 u^{\frac{\alpha}{\mu}-1} \left(\frac{\nu_1-\lambda}{\nu_1-\lambda u}\right)^{\frac{\lambda}{\mu}} du}.$$

Substituting (10) into (8), we obtain

$$\Pi_1(z) = \frac{(\lambda + \alpha + \mu z A(z))\Pi_0(z)}{\nu_1}.$$

5 Concluding Remarks

In this paper, we present a new queueing model for cloud computing systems where the processing unit and the storage unit are separated. The model is explicitly analyzed in terms of generating functions. Furthermore, we have presented a simple recursive scheme allowing to calculate the stationary distribution. We also consider one special case of our model which has appeared in the literature. For future work, we would like to extend our model to a multiserver setting which may call for a level-dependent QBD formulation [13]. It might be also interesting to consider the corresponding model with constant retrial rate as in [15].

Acknowledgments Tuan Phung-Duc was supported in part by Japan Society for the Promotion of Science, JSPS Grant-in-Aid for Young Scientists (B), Grant Number 2673001. The author would like to thank the anonymous referees for constructive comments which improve the presentation of the paper.

References

1. Artalejo, J.R., Gomez-Corral, A.: Steady state solution of a single-server queue with linear repeated request. Journal of Applied Probability **34**, 223–233 (1997)
2. Artalejo, J.R., Atencia, I.: On the single server retrial queue with batch arrivals. Sankhya **66**, 140–158 (2004)
3. Artalejo, J.R., Joshua, V.C., Krishnamoorthy, A.: An M/G/1 retrial queue with orbital search by the server. In: Artalejo, J.R., Krishnamoorthy, A., (eds.) Advances in Stochastic Modelling, pp. 41–54. Notable Publications Inc., NJ (2002)
4. Artalejo, J.R., Phung-Duc, T.: Markovian retrial queues with two way communication. Journal of Industrial and Management Optimization **8**(4), 781–806 (2012)
5. Artalejo, J.R., Phung-Duc, T.: Single server retrial queues with two way communication. Applied Mathematical Modelling **37**(4), 1811–1822 (2013)
6. Chakravarthy, S.R., Krishnamoorthy, A., Joshua, V.C.: Analysis of a multi-server retrial queue with search of customers from the orbit. Performance Evaluation **63**(8), 776–798 (2006)
7. Deepak, T.G., Dudin, A.N., Joshua, V.C., Krishnamoorthy, A.: On an $M^X/G/1$ retrial system with two types of search of customers from the orbit. Stochastic Analysis and Applications **31**(1), 92–107 (2013)
8. 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**(1), 169–179 (2004)

9. Falin, G., Templeton, J.G.: Retrial Queues. Chapman and Hall (1997)
10. Krishnamoorthy, A., Deepak, T.G., Joshua, V.C.: An M/G/1 retrial queue with nonpersistent customers and orbital search. Stochastic Analysis and Applications **23**(5), 975–997 (2005)
11. Phung-Duc, T., Masuyama, H., Kasahara, S., Takahashi, Y.: M/M/3/3 and M/M/4/4 retrial queues. Journal of Industrial and Management Optimization **5**(3), 431–451 (2009)
12. Phung-Duc, T., Masuyama, H., Kasahara, S., Takahashi, Y.: State-dependent M/M/c/c+ r retrial queues with Bernoulli abandonment. Journal of Industrial and Management Optimization **6**(3), 517–540 (2010)
13. Phung-Duc, T., Masuyama, H., Kasahara, S., Takahashi, Y.: A simple algorithm for the rate matrices of level-dependent QBD processes. In: Proceedings of the 5th International Conference on Queueing Theory and Network Applications, pp. 46–52. ACM (2010)
14. Phung-Duc, T.: An explicit solution for a tandem queue with retrials and losses. Operational Research **12**(2), 189–207 (2012)
15. Phung-Duc, T., Rogiest, W., Takahashi, Y., Bruneel, H.: Retrial queues with balanced call blending: analysis of single-server and multiserver model. Annals of Operations Research (2014). doi:10.1007/s10479-014-1598-2

Performance Analysis and Optimization of a Queueing Model for a Multi-skill Call Center in M-Design

Dequan Yue, Chunyan Li and Wuyi Yue

Abstract This paper studies a queuing model of a multi-skill call center in M-design. In this model, there are two types of customers and three groups of servers who have different skills. Servers in Group 1 can only serve type 1 customers, servers in Group 2 can only serve type 2 customers, and servers in Group 3 can serve both type 1 and type 2 customers. We obtain the state-transition rates by using results from M/M/c/c and M/M/c queueing systems. Then, we establish equations for the steady-state probabilities of the system. Finally, we obtain the computational formula for the service level and we present an optimization of a staffing problem.

Keywords Multi-skill call center · Queuing model · Steady-state probabilities · Service level · Optimization

D. Yue(✉)
College of Science, Yanshan University, Qinhuangdao 066004, China
email: ydq@ysu.edu.cn

C. Li
School of Economics and Management, Yanshan University, Qinhuangdao 066004, China

C. Li
Zhijiang College of Zhejiang University of Technology, Hangzhou 310024, China
email: llccyy1980@126.com

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

T.V. Do et al. (eds.), *Queueing Theory and Network Applications*,
Advances in Intelligent Systems and Computing 383,
DOI: 10.1007/978-3-319-22267-7_14

1 Introduction

Call centers are becoming increasingly important in the global business environment. Correspondingly, as the importance and complexity of modern call centers grow, there is a proliferation of literature relating to them, typically focusing on queueing models. In a queueing model of a call center, the call agents and calls correspond to servers and customers, respectively. For important related surveys, we refer to Koole and Mandelbaum [1] and Gans et al. [2]. An introduction to staffing problems with relevant bibliographic references can be found in Aksin et al. [3].

Multi-skill call centers have emerged and have recently been studied in the literature. A multi-skill call center handles several types of calls, and each agent has a selected number of skills. The agents are distinguished by the set of call types they can handle. A typical example is an international call center where incoming calls are in different languages, see Gan et al. [2].

Perry and Nilsson [4] considered a multi-skill call center with two classes of calls that are served by a single pool of agents. They determined the required number of agents and an assignment policy to satisfy a target for the expected waiting times of callers. Such multi-skill call centers are referred to as V-models or V-designs. Bhulai and Koole [5] proposed scheduling policies and showed that the policy is optimal for equal service time distributions. Gans and Zhou [6] also studied the same V-design model using a linear programming approach. They obtained results for the case of unequal service rates.

Örmeci [7] studied a dynamic admission control for a multi-skill call center in M-design where there are two classes of calls and three stations: one dedicated to each class, and one shared station. He showed that serving a call in its assigned station, whenever possible, is optimal. In this paper, we study an M-design model for a multi-skill call center by using a queueing model. We focus on the performance analysis and optimization for this M-design model of a multi-skill call center.

The rest of the paper is organized as follows. In Section 2, we describe the M-design model for a multi-skill call center. In Section 3, we obtain the state-transition rates by using results of M/M/c/c and M/M/c queueing systems. Then, we establish equations for the steady-state probabilities of the system. In Section 4, we obtain the computational formula for the service level and present a staffing problem. Section 5 concludes the paper.

2 System Model

In this paper, we study an M-design model for a multi-skill call center where there are two types of calls and three groups of servers.

1. **Arrival Process:** There are two types of calls (or customers). The calls of Type 1 and Type 2 arrive according to a Poisson process with rates λ_1 and λ_2,

respectively. Arriving calls are lined in two queues. Queue 1 and Queue 2 consist of calls of Type 1 and Type 2, respectively. There are infinite waiting spaces for both queues.

2. **Service Process:** There are three groups of servers (agents). Group 1, Group 2 and Group 3 consist of N_1 servers, N_2 servers and N_3 servers, respectively. Group 1 and Group 2 are specialized servers who can only serve customers of Type 1 and Type 2 calls, respectively. Group 3 is made up of flexible servers who can serve customers of both Type 1 and Type 2. The service times of servers in Group 1, Group 2 and Group 3 are all exponentially distributed with parameters μ_1, μ_2 and μ_3, respectively.
3. **Routing Policy:** A arriving customer of Type 1 (or Type 2) have priority to be served by a server in Group 1 (or Group 2) if there are free servers in Group 1 (or Group 2) and free servers in Group 3. If all servers in Group 1 (Group 2) are busy, the customer will be serviced by a free server in Group 3. If all servers are busy in Group 3, the customer must wait in Queue 1 or Queue 2.
4. **Queuing Discipline:** A free server in Group 1 (Group 2) serve the waiting customers in Queue 1 (Queue 2) according to a First-come First-served (FCFS) discipline, and a free server in Group 3 serve the waiting customers in Queue 1 or Queue 2 according to FCFS discipline. If all servers in Group 1 and Group 2 are busy and there are waiting customers both in Queue 1 and Queue 2, a free server in Group 3 will select randomly (with equal probability) a customer of Type 1 and Type 2 for service.

3 The Steady-State Probabilities

In this section, we firstly define the states of the model. Then, we derive the state-transition rates by using results of M/M/c/c and M/M/c queueing systems. Finally, we obtain equations for the steady-state probabilities of the system. It is assumed that the system is stationary.

3.1 The State Space

The M-design model has three skill groups. Each group has three states: an idle state (denoted by 1), that being the case where at least one agent is idle; a busy state (denoted by 2), the case where all agents in the group are busy and there no calls waiting for service rendered by this group; and an overload state (denoted by 3), the case where all agents in the group are busy and there is at least one call waiting for service by this group. Theoretically, the system has 27 states. However, due to the routing policy assumed in Section 2, the 15 states above marked in boldface do not exist. Therefore, the state space actually consists of 12 states, which are given by

$$E = \{(111),(121),(122),(132),(211),(212),(222),(221),(232),(312),(322),(332)\}$$

Let $n_1(n_2)$ be the number of customers waiting for service including those being serviced by servers of Group 1 (2), and n_3 be the number of customers being serviced by servers of Group 3. The i th state in E is denoted by $S_i, i=1,2,...,12$.

3.2 The Calculation of the State-Transition Rates

The transition of states occurs due to either the arrival of a call or the completion of a service.

1. The state-transition due to the arrival of calls.

Consider the state S_1. The trigger for the transfer from state S_1 to state S_3 is a call of Type 1. The transition rate q_{1-3} from state S_1 to state S_3 is given as follows:

$$q_{1-3} = \lambda_1 P(n_1 = N_1 - 1) \tag{1}$$

where $P(n_1 = N_1 - 1)$ is the probability that there are $N_1 - 1$ calls of Type 1 needing to be serviced by the agents in Group 1.

Note that if the process is in the state S_1 then the number of calls of either Type 1 or Type 2 is less than the number of the agents either in Group 1 or Group 2, i.e., $n_1 < N_1$ and $n_2 < N_2$. In this case, each queue behaves like an M/M/c/c loss queuing system. Thus, using the results of the M/M/c/c loss queuing system, we have

$$P(n_1 = N_1 - 1) = \frac{\rho_1^{N_1 - 1}}{(N_1 - 1)! \sum_{j=0}^{N_1} \frac{\rho_1^j}{j!}} \tag{2}$$

where $\rho_1 = \frac{\lambda_1}{\mu_1}$. Similarly, the other transition rates q_{i-j} caused by the arrival of calls are given as follows:

$q_{1-2} = q_{3-5} = q_{6-8} = q_{9-11} = \lambda_2 P(n_2 = N_2 - 1)$,

$q_{1-3} = q_{2-5} = q_{4-8} = q_{7-10} = \lambda_1 P(n_1 = N_1 - 1)$,

$q_{2-4} = \lambda_2 P^1(n_3 = N_3 - 1)$, $q_{3-6} = \lambda_1 P^2(n_3 = N_3 - 1)$,

$q_{5-8} = (\lambda_1 + \lambda_2) P^3(n_3 = N_3 - 1)$,

$q_{4-7} = q_{8-10} = q_{11-12} = \lambda_2$, $q_{6-9} = q_{8-11} = q_{10-12} = \lambda_1$

where

$$P(n_2 = N_2 - 1) = \frac{\rho_2^{N_2-1}}{(N_2-1)!\sum_{j=0}^{N_2}\frac{\rho_2^j}{j!}}, \tag{3}$$

$$P^1(n_3 = N_3 - 1) = \frac{\rho_4^{N_3-1}}{(N_3-1)!\sum_{j=0}^{N_3}\frac{\rho_4^j}{j!}}, \tag{4}$$

$$P^2(n_3 = N_3 - 1) = \frac{\rho_3^{N_3-1}}{(N_3-1)!\sum_{j=0}^{N_3}\frac{\rho_3^j}{j!}}, \tag{5}$$

$$P^3(n_3 = N_3 - 1) = \frac{(\rho_3+\rho_4)^{N_3-1}}{(N_3-1)!\sum_{j=0}^{N_3}\frac{(\rho_3+\rho_4)^j}{j!}} \tag{6}$$

where $\rho_2 = \frac{\lambda_2}{\mu_2}$, $\rho_3 = \frac{\lambda_1}{\mu_3}$, $\rho_4 = \frac{\lambda_2}{\mu_3}$.

5. The state-transition due to the completion of a service of a call.

Consider the state S_2. The trigger of the transfer from state S_2 to state S_1 is due to a service completion of a call of Type 2. If the state process is in state S_2, then all N_2 agents are busy. Thus, the transition rate from state S_2 to state S_1 is $q_{2-1} = N_2\mu_2$. Similar analysis gives the other transition rates q_{i-j} caused by a service completion which are given as follows:

$$q_{2-1} = q_{5-3} = q_{8-6} = q_{11-9} = N_2\mu_2, \quad q_{3-1} = q_{5-2} = q_{8-4} = q_{10-7} = N_1\mu_1,$$

$$q_{4-2} = q_{6-3} = q_{8-5} = N_3\mu_3, \quad q_{7-4} = q_{10-8} = (N_2\mu_2 + N_3\mu_3)P(n_2 = N_2 + 1),$$

$$q_{9-6} = q_{11-8} = (N_1\mu_1 + N_3\mu_3)P(n_1 = N_1 + 1),$$

$$q_{12-11} = (N_2\mu_2 + \frac{1}{2}N_3\mu_3)P(n_2 = N_2 + 1),$$

$$q_{12-10} = (N_1\mu_1 + \frac{1}{2}N_3\mu_3)P(n_1 = N_1 + 1)$$

where the probabilities of $P(n_1 = N_1 + 1)$ and $P(n_2 = N_2 + 1)$ can be obtained by using the results of the M/M/c queuing system which are given as follows:

$$P(n_1 = N_1 + 1) = \frac{\rho_1^{N_1+1}}{N_1(N_1)!}P_0^1, \tag{7}$$

$$P(n_2 = N_2 + 1) = \frac{\rho_2^{N_2+1}}{N_2(N_2)!}P_0^2 \tag{8}$$

where

$$P_0^1 = \left[\sum_{j=0}^{N_1-1}\frac{\rho_1^j}{j!} + \frac{N_1\rho_1^{N_1}}{N_1!(N_1-\rho_1)}\right]^{-1}, \tag{9}$$

$$P_0^2 = \left[\sum_{j=0}^{N_2-1}\frac{\rho_2^j}{j!} + \frac{N_2\rho_2^{N_2}}{N_2!(N_2-\rho_2)}\right]^{-1}. \tag{10}$$

3.3 The Equations for the Steady-State Probabilities

Let $P_i, i = 1, 2, ..., 12$ be the steady-state probabilities of the state process. Then, we can obtain the equations for the steady-state probabilities of the system as follows:

$$P_1(q_{1-2} + q_{1-3}) = P_2q_{2-1} + P_3q_{3-1},$$

$$P_2(q_{2-1} + q_{2-4} + q_{2-5}) = P_1q_{1-2} + P_4q_{4-2} + P_5q_{5-2},$$

$$P_3(q_{3-1} + q_{3-5} + q_{3-6}) = P_1q_{1-3} + P_5q_{5-3} + P_6q_{6-3},$$

$$P_4(q_{4-2} + q_{4-7} + q_{4-8}) = P_2q_{2-4} + P_7q_{7-4} + P_8q_{8-4},$$

$$P_5(q_{5-2} + q_{5-3} + q_{5-8}) = P_2q_{2-5} + P_3q_{3-5} + P_8q_{8-5},$$

$$P_6(q_{6-3} + q_{6-8} + q_{6-9}) = P_3q_{3-6} + P_8q_{8-6} + P_9q_{9-6},$$

$$P_7(q_{7-4} + q_{7-10}) = P_4q_{4-7} + P_{10}q_{10-7},$$

$$P_8(q_{8-4}+q_{8-5}+q_{8-6}+q_{8-10}+q_{8-11})=P_4q_{4-8}+P_5q_{5-8}+P_6q_{6-8}+P_{10}q_{10-8}+P_{11}q_{11-8},$$

$$P_9(q_{9-6}+q_{9-11})=P_6q_{6-9}+P_{11}q_{11-9},$$

$$P_{10}(q_{10-7}+q_{10-8}+q_{10-12})=P_7q_{7-10}+P_8q_{8-10}+P_{12}q_{12-10},$$

$$P_{11}(q_{11-8}+q_{11-9}+q_{11-12})=P_8q_{8-11}+P_9q_{9-11}+P_{12}q_{12-11},$$

$$P_{12}(q_{12-10}+q_{12-11})=P_{10}q_{10-12}+P_{11}q_{11-12},$$

$$\sum_{i=1}^{12}P_i=1\cdot$$

All the steady-state probabilities can be obtained by solving these equations. However, the calculation of these probabilities are very cumbersome. In next section, these probabilities are calculated numerically by using Matlab software.

4 Optimization Problem

In this section, we first consider the calculation of the service level. Then, we consider the staffing problem.

4.1 The Calculation of the Service Level

The service level is defined as the percentage of the serviced calls in a given fixed waiting time. Actually, the 80/20 principle is a general rule, that is to say at least 80 percent of the calls should be serviced within a 20 second waiting time. We can derive the service level using the steady-state probabilities.

Let P_{sl}^1 and P_{sl}^2 be the probabilities that the call of Type 1and Type 2 is serviced in a fixed time T_1 and T_2, respectively. Let $P_{ns}^1=1-P_{sl}^1$ and $P_{ns}^2=1-P_{sl}^2$.

Consider a call of Type 1. Calls of Type 1 form a queue when the process is in states S_9, S_{11}, S_{12}. It can be seen that the service rate for calls of Type 1 in each state of S_9 and S_{11} is $N_1\mu_1+N_3\mu_3$, and the service rate in state S_{12} is $N_1\mu_1+\frac{1}{2}N_3\mu_3$, Thus, we get the probability P_{ns}^1 that a call of Type 1 can not be serviced in time T_1 as follows:

$$P_{ns}^1=P_9\sum_{i=k_1}^{\infty}P(n_1=i)+P_{11}\sum_{i=k_1}^{\infty}P(n_1=i)+P_{12}\sum_{i=k_2}^{\infty}P(n_1=i) \qquad (11)$$

where

$$k_1 = N_1 + N_3 + T_1(N_1\mu_1 + N_3\mu_3), \tag{12}$$

$$k_2 = N_1 + \frac{1}{2}N_3 + T_1(N_1\mu_1 + \frac{1}{2}N_3\mu_3). \tag{13}$$

Similarly, we get the probability P_{ns}^2 that a call of Type 2 can not be served in time T_2 as follows:

$$P_{ns}^2 = P_7\sum_{i=k_3}^{\infty}P(n_2 = i) + P_{10}\sum_{i=k_3}^{\infty}P(n_2 = i) + P_{12}\sum_{i=k_4}^{\infty}P(n_2 = i) \tag{14}$$

where

$$k_3 = N_2 + N_3 + T_2(N_2\mu_2 + N_3\mu_3), \tag{15}$$

$$k_4 = N_2 + \frac{1}{2}N_3 + T_2(N_2\mu_2 + \frac{1}{2}N_3\mu_3). \tag{16}$$

Remark 1. The probability $P(n_1 = i)$ [$P(n_2 = i)$] in Eq. (11)[Eq. (14)] is the probability that there are i customers in the M/M/ N_1 [N_2] queuing system with arrival rate λ_1 [λ_2] and service rate μ_1 [μ_2] which is given in [8]. Their expressions are omitted.

Table 1 gives numerical results for the service levels P_{sl}^1 and P_{sl}^2 for λ_1=5, μ_1=0.5, λ_2=4, μ_2=0.3, μ_3=0.2, T_1=20, T_2=30.

Table 1 The numerical results of the service levels P_{sl}^1 and P_{sl}^2

N_1	N_2	N_3	P_{sl}^1	P_{sl}^2
20	20	20	1.0000	1.0000
20	15	25	1.0000	0.9765
15	20	25	1.0000	1.0000
15	20	10	0.9988	0.9995
15	15	15	0.9994	0.9547
20	15	10	1.0000	0.8693
15	15	10	0.9984	0.9135
11	15	10	0.8200	0.8983
11	14	11	0.8028	0.6677
11	16	9	0.8075	0.9551

4.2 Optimization of a Staffing Problem

Let C_1, C_2 and C_3 be the costs for each server in Group 1, Group 2 and Group 3, respectively. In order to minimize the cost, we try to find the optimal number of servers N_1, N_2 and N_3 subject to the constraint conditions. The optimization of a staffing problem can be expressed as follows:

$$\begin{aligned} \min \quad & C_1N_1 + C_2N_2 + C_3N_3 \\ \text{s.t.} \quad & P_{sl}^1 \geq \alpha_1, \\ & P_{sl}^2 \geq \alpha_2, \\ & N_1, N_2, N_3 \in Z^+ \end{aligned}$$

where α_1 and α_2 are the given service rate of the call of Type 1 and the call of Type 2, Z^+ denote the set of positive integer.

5 Conclusions

This paper has studied a queuing model of the M-design multi-skill call center. We have obtained the transition rates of states by using results from an M/M/c/c and M/M/c queueing system and then established equations for the steady-state probabilities of the system. We have derived the computational formula for the service level and presented an optimization of a staffing problem. In this work, we studied an exponential model in a multi-skill call center. A further extension for future research would be to study non-exponential models or models with impatient customers.

Acknowledgments This work is supported in part by the National Natural Science Foundation of China (No. 71071133) and the MEXT, Japan.

References

1. Koole, G., Mandelbaum, A.: Queueing models of call centers: an introduction. Annals of Operations Research **113**, 41–59 (2002)
2. Gans, N., Koole, G., Mandelbaum, A.: Telephone call centers: tutorial, review, and research prospects. Manufacturing & Service Operation Management **5**(2), 79–141 (2003)
3. Aksin, Z., Armony, M., Mehrotra, V.: The modern call-center: a multi-disciplinary perspective on operations management research. Production and Operations Management **16**, 655–688 (2007)

4. Perry, M., Nilsson A.: Performance modeling of automatic call distributors: assignable grade of service staffing. In: Proceedings of the 14th International Switching Symposium, pp. 294–298 (1992)
5. Bhulai, S., Koole, G.: A queueing model for call blending in call centers. IEEE Transactions on Automatic Control **48**, 1434–1438 (2003)
6. Gans, N., Zhou, Y.: A call-routing problem with service-level constraints. Operations Research **51**(2), 255–271 (2003)
7. Örmeci, L.E.: Dynamic admission control in a call center with one shared and two dedicated service facilities. IEEE Transactions on Automatic Control **49**(7), 1157–1161 (2004)
8. Gross, D., Harris, C.M.: Fundamentals of Queueing Theory, 2nd edn. Wiley, New York (1985)

Multi-server Queue with Job Service Time Depending on a Background Process

Tomoyuki Sakata and Shoji Kasahara

Abstract One of approaches to reducing energy consumption in a data center is to power down a group of servers. In this paper, we consider a power management scheme for distributed parallel processing over clusters of servers, where part of servers in each cluster are turned off in power-saving mode. We model the system as a multi-server queue in which the service time of a job depends on the state of a background process at the beginning of the job service. We analyze the joint distribution of the number of jobs in system and the state of the background process, deriving the mean job-response time and mean amount of energy consumption. In numerical examples, we investigate how the mean job-response time and energy consumption are affected by energy saving level and the number of clusters.

1 Introduction

Recently, cloud computing has been attracted considerable attention, and various kinds of computing services such as virtual machines and MapReduce software framework are provided by data centers. A data center contains a large number of server machines, resulting in high energy consumption. With the increase in cloud computing demand, the number of data centers is growing rapidly, and the amount of energy consumption for data centers is extremely huge. Therefore, considerable research efforts have been devoted to developing schemes which can save energy without degrading job-processing performance.

There exist much literature which concerns strategies of efficient energy saving for data centers. A typical scheme for reducing energy consumption in data centers is to manage the power of server machines according to demand for computing resources. In [5], the authors propose a scheme of server-power management for a data center,

T. Sakata · S. Kasahara (✉)
Graduate School of Information Science, Nara Institute of Science and Technology,
Takayama 8916-5, Ikoma, Nara 6300192, Japan
e-mail: {sakata.tomoyuki.so7,kasahara}@is.naist.jp

T.V. Do et al. (eds.), *Queueing Theory and Network Applications*,
Advances in Intelligent Systems and Computing 383,
DOI: 10.1007/978-3-319-22267-7_15

with which the number of running servers is varied according to the number of jobs in system. If the number of waiting jobs exceeds a predefined threshold, all the servers are turned on. If the number of servers busy for job processing are below another threshold, a given number of servers are turned off. In [5], the trade-off between job-waiting time and energy consumption is analyzed by an M/M/c queue with threshold-based on/off control.

In [1], the authors propose Berkeley Energy Efficient MapReduce (BEEMR), an efficient power management scheduling for MapReduce-type job processing. In BEEMR, servers in a data center are divided into two grouping zones, an interactive zone and a batch zone. The servers in the interactive zone are always turned on and serve small-sized jobs. On the other hand, the servers in the batch zone process huge-sized jobs that are insensitive to the response time. BEEMR controls the power of servers in the batch zone so that the amount of power consumption is reduced. The authors in [1] investigate the performance of BEEMR by simulation and on-site practical experiments. In [2], the performance of BEEMR is investigated by queueing theoretical approach.

In this paper, we consider a power-management scheme for data centers with server clusters. We focus on a data center accommodating a large number of server clusters, each of which consists of several server machines, providing parallel distributed computing service. The data center alternates two power-operation modes: normal-operation and power-saving. The periods of normal-operation mode and power-saving one are determined a priori, and the system alternates between two modes independently of the number of jobs in the system. The power of server machines is managed in a cluster-based manner. When the data center is in normal-operation mode, all the servers of all clusters are powered on. When the data center switches to power-saving mode, a part of servers in each cluster are powered off after completing the existing job. In both operation modes, a job is served by a cluster according to processor sharing discipline. Therefore, the service rate of a cluster in power-saving mode is smaller than that in normal mode.

In order to investigate the performance of the cluster-based power management scheme, we model the data center as a multi-server queueing system with job service depending on the state of a background process. Here, a server of the queueing model corresponds to a cluster of machines in the data center. The service time of a job depends on the state of the background process at the beginning of the job's service. We construct a trivariate continuous-time Markov chain for the system, deriving the steady-state probability vector by matrix geometric method. We consider the mean job-response time and mean amount of energy consumption as performance measures, and investigate how the performance measures are affected by system parameters such as the number of clusters and the parameter indicating energy-saving level.

This paper is organized as follows. In Sect. 2, we describe the queueing model considered in this paper, and analyze the steady-state distribution by matrix analytic method in Sect. 3. Numerical examples are shown in Sect. 4, and finally in Sect. 5 our conclusion is presented.

2 Queueing Model

We assume that the number of servers is c, and that the buffer capacity is infinite. Jobs arrive at the system according to a Poisson process with rate λ. The service time of a job depends on the state of a background process when the job service starts. The background process is continuous-time Markov chain with two states, Slow (S) and Fast (F), independent of the arrival process. The state S describes the power-saving mode in which a part of worker machines are turned off for energy saving, and hence the resulting service rate of a server is low. When the state of the background process is F, on the other hand, all the worker machines composing a server are turned on and the resulting service rate of the server is greater than that in state S. The state-transition rate from S to F and that from F to S is given by α_S and α_F, respectively.

When a job enters a server for its service, its service time depends on the state of the background process. If the background process is in state S (resp. F) at the service initiation point, the service time of the job follows an exponential distribution with rate μ_S (resp. μ_F). In the following, $\mu_S < \mu_F$. We also assume that when the background process switches from S to F (and vice versa), the service rate of the existing job remains the same as that at its service initiation point. Hereafter, a job served with rate μ_S and that with rate μ_F are called S job and F job, respectively.

3 Analysis

We define $N(t)$ as the number of jobs in the system at time t. Let $S(t)$ and $F(t)$ denote the numbers of S jobs and F jobs at t, respectively. We denote $J(t)$ ($\in \{S, F\}$) as the state of the background process at t. $F(t)$ can be expressed with $N(t)$ and $S(t)$ by

$$F(t) = \min(N(t), c) - S(t).$$

From the assumptions, $\{(N(t), S(t), J(t)) : t \geq 0\}$ is a trivariate continuous-time Markov chain with state space $\mathbb{F}$, where $\mathbb{F}$ is given by

$$\mathbb{F} = \mathbb{N} \cup \{0\} \times \{0, 1, \ldots, c\} \times \{S, F\}.$$

Let $\boldsymbol{Q}$ denote the infinitesimal generator of the Markov chain $\{N(t), S(t), J(t) : t \geq 0\}$, whose states are arranged in lexicographic order. Then, $\boldsymbol{Q}$ is given by

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{B}_0 & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \cdots \\ \boldsymbol{B}_1 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \cdots \\ \boldsymbol{O} & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & \boldsymbol{O} & \boldsymbol{O} & \cdots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & \boldsymbol{O} & \cdots \\ & & & \ddots & \ddots & \ddots & \end{bmatrix}. \tag{1}$$

In what follows, we describe the details of block matrices $\boldsymbol{B}$, $\boldsymbol{B}_0$, $\boldsymbol{B}_1$, $\boldsymbol{A}_0$, $\boldsymbol{A}_1$, and $\boldsymbol{A}_2$ in (1). Hereafter, i is an integer such that the inequality $k(k+1) < i \leq (k+1)(k+2)$ holds for any $k \in [0, c]$, and $[x]$ is the largest integer not greater than x.

(a) $c(c+1) \times c(c+1)$ matrix $\boldsymbol{B}$

(i) For odd i,

$$[\boldsymbol{B}]_{ij} = \begin{cases} \alpha_S, & j = i+1, \\ \lambda, & j = i+2k+4, \\ d(k+1, i+1)\,\mu_F, & j = i-2k, \\ -d(k, i-1)\,\mu_S, & j = i-2k-2, \\ -\alpha_S - \lambda - d(k+1, i+1)\,\mu_F + d(k, i-1)\,\mu_S, & j = i, \\ 0, & \text{otherwise}, \end{cases}$$

where $d(n, m) = \dfrac{n(n+1) - m}{2}$.

(ii) For even i,

$$[\boldsymbol{B}]_{ij} = \begin{cases} \alpha_F, & j = i-1, \\ \lambda, & j = i+2k+2, \\ d(k+1, i)\,\mu_F, & j = i-2k, \\ -d(k, i-2)\mu_S, & j = i-2k-2, \\ -\alpha_F - \lambda - d(k+1, i)\,\mu_F + d(k, i-2)\mu_S, & j = i, \\ 0, & \text{otherwise}. \end{cases}$$

(b) $c(c+1) \times 2(c+1)$ matrix $\boldsymbol{B}_0$

(i) For odd i,

$$[\boldsymbol{B}_0]_{ij} = \begin{cases} \lambda, & j = i - c(c-1) + 2 \text{ and } j \neq 1, \\ 0, & \text{otherwise}. \end{cases}$$

(ii) For even i,

$$[\boldsymbol{B}_0]_{ij} = \begin{cases} \lambda, & j = i - c(c-1), \\ 0, & \text{otherwise}. \end{cases}$$

(c) $2(c+1) \times c(c+1)$ matrix $\boldsymbol{B}_1$

$$[\boldsymbol{B}_1]_{ij} = \begin{cases} \left(c - \left[\dfrac{i-1}{2}\right]\right)\mu_F, & j = i + c(c-1), \\ \left[\dfrac{i-1}{2}\right]\mu_S, & j = i + c(c-1) - 2, \\ 0, & \text{otherwise}. \end{cases}$$

(d) $2(c+1) \times 2(c+1)$ matrix $\boldsymbol{A}_0$

$$[\boldsymbol{A}_0]_{ij} = \begin{cases} \lambda, & j = i, \\ 0, & j \neq i. \end{cases}$$

(e) $2(c+1) \times 2(c+1)$ matrix $\boldsymbol{A}_1$

(i) For odd i,

$$[\boldsymbol{A}_1]_{ij} = \begin{cases} \alpha_S, & j = i+1, \\ -\alpha_S - \lambda - \left[\dfrac{i-1}{2}\right]\mu_S - \left(c - \left[\dfrac{i-1}{2}\right]\right)\mu_F, & j = i, \\ 0, & \text{otherwise.} \end{cases}$$

(ii) For even i,

$$[\boldsymbol{A}_1]_{ij} = \begin{cases} \alpha_F, & j = i-1, \\ -\alpha_F - \lambda - \left[\dfrac{i-1}{2}\right]\mu_S - \left(c - \left[\dfrac{i-1}{2}\right]\right)\mu_F, & j = i, \\ 0, & \text{otherwise.} \end{cases}$$

(f) $2(c+1) \times 2(c+1)$ matrix $\boldsymbol{A}_2$

(i) For odd i,

$$[\boldsymbol{A}_2]_{ij} = \begin{cases} \dfrac{i-1}{2}\mu_S, & j = i, \\ \left(c - \dfrac{i-1}{2}\right)\mu_F, & j = i+2, \\ 0, & \text{otherwise.} \end{cases}$$

(ii) For even i,

$$[\boldsymbol{A}_2]_{ij} = \begin{cases} \left(c - \dfrac{i-2}{2}\right)\mu_F, & j = i, \\ \dfrac{i-2}{2}\mu_S, & j = i-2, \\ 0, & \text{otherwise.} \end{cases}$$

We define the steady-state probability as

$$\pi(i, j, k) = \lim_{t \to \infty} \Pr\{N(t) = i, S(t) = j, J(t) = k\}, \quad (i, j, k) \in \mathbb{F}.$$

We also define the following notations.

$$\begin{aligned} \boldsymbol{\pi}_{-1} = (&\pi(0,0,S), \pi(0,0,F), \pi(1,0,S), \pi(1,0,F), \\ &\pi(1,1,S), \pi(1,1,F), \ldots, \pi(c-1,0,S), \pi(c-1,0,F), \\ &\pi(c-1,1,S), \pi(c-1,1,F), \ldots, \pi(c-1,c-1,S), \pi(c-1,c-1,F)), \\ \boldsymbol{\pi}_i = (&\pi(c+i,0,S), \pi(c+i,0,F), \pi(c+i,1,S), \pi(c+i,1,F), \ldots, \\ &\pi(c+i,c,S), \pi(c+i,c,F)), \qquad i \geq 0. \end{aligned}$$

Let $\boldsymbol{\pi} = (\boldsymbol{\pi}_{-1}, \boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)$. $\boldsymbol{\pi}$ is the steady-state probability vector which satisfies $\boldsymbol{\pi}\boldsymbol{Q} = \boldsymbol{0}$ and $\boldsymbol{\pi}\boldsymbol{e} = 1$.

From (1), this continuous-time Markov chain is a quasi birth-and-death process. The steady-state probability vector $\boldsymbol{\pi}$ can be calculated by matrix-analytic method [3].

In terms of the system stability, we have the following theorem.

Theorem 1. *([4], p. 411, (9.36)) We assume that* $2(c+1)$ *dimensional square matrix* $\boldsymbol{A} = \boldsymbol{A}_0 + \boldsymbol{A}_1 + \boldsymbol{A}_2$ *satisfies* $\boldsymbol{\pi}_A \boldsymbol{A} = \boldsymbol{0}$ *and* $\boldsymbol{\pi}_A \boldsymbol{e}_1 = 1$.

Then, the stability condition for the system is

$$\boldsymbol{\pi}_A \boldsymbol{A}_0 \boldsymbol{e}_1 < \boldsymbol{\pi}_A \boldsymbol{A}_2 \boldsymbol{e}_1.$$

In our case, we can conjecture the following stability condition.

$$\lambda < \frac{c\mu_S\mu_F\left(\alpha_S^2 + 2\alpha_S\alpha_F + \alpha_F^2 + \alpha_S\mu_F + \alpha_F\mu_S\right)}{(\alpha_S + \alpha_F)(\alpha_S\mu_S + \alpha_F\mu_F + \mu_S\mu_F)}. \tag{2}$$

Finally, we consider two performance measures: the mean job-response time and mean amount of energy consumption. Let $\boldsymbol{a}$ denote the $1 \times c(c+1)$ vector whose ith element is given by

$$[\boldsymbol{a}]_i = k, \quad k(k+1) < i \leq (k+1)(k+2),$$

for $k = 0, 1, 2, \ldots, c-1$. The mean number of jobs in system is then given by

$$\mathrm{E}[L] = \boldsymbol{a}\boldsymbol{\pi}_{-1} + \sum_{i=0}^{\infty}(c+i)\boldsymbol{\pi}_i\boldsymbol{e}_1. \tag{3}$$

Using Little's law, the mean job-response time $\mathrm{E}[T]$ is given by $\mathrm{E}[T] = \mathrm{E}[L]/\lambda$.

Let E denote the mean amount of energy consumption per unit time. E can be expressed as

$$E = \sum_{(i,j,k)\in\mathbb{F}} \pi(i,j,k)\{j\mu_S + \min(c-j, i-j)\mu_F + \max(0, c-i)\kappa\mu_k\},$$

where κ is the ratio of the amount of energy consumption of a single idle server to that of a single busy server.

4 Numerical Examples

We define ζ as $\zeta = \mu_S/\mu_F$. ζ is related to the ratio of the service rate of power-saving mode to that of normal-operation mode. A large ζ indicates that the number of worker machines turned off in power-saving mode is small. In other words, a large

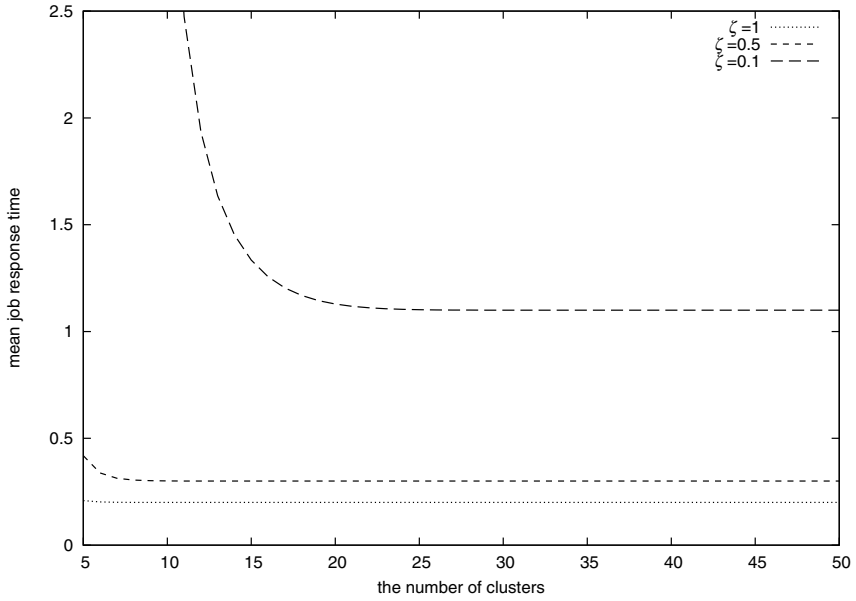


Fig. 1 c vs. E[T].

ζ implies a small amount of energy saving for the system. Note that when the system is in power-saving mode, all servers keep running for $\zeta = 1$, whereas all the servers are turned off for $\zeta = 0$.

In the following, we set $\kappa = 170/240$ according to [5]. We only consider the case of $\lambda = 12$ and $\alpha_S = \alpha_F = 1$ due to page limitation.

Figure 1 illustrates the mean job-response time E[T]. The horizontal axis is the number of clusters c, and E[T]'s for $\zeta = 0.1, 0.5, 1$ are plotted. Note that $\zeta = 1$ under $\alpha_S = \alpha_F$ corresponds to the case of an M/M/c with service rate μ_F. We observe from the figure that when c increases, E[T]'s for the three cases decrease and converge to some constants. We also observe that E[T] increases with the decrease in ζ. A remarkable point here is that the discrepancy between $\zeta = 1$ and $\zeta = 0.5$ is significantly smaller than that between $\zeta = 1$ and $\zeta = 0.1$. When $\zeta = 0.5$, a half of servers in a cluster are powered off in power-saving mode. This figure shows that energy-saving level of $\zeta = 0.5$ does not degrade the job-response time.

Figure 2 represents E against the number of clusters c. In this figure, the amount of energy consumption increases linearly for any ζ, as expected. We also observe that E for $\zeta = 1$ is the largest for any c, and that E becomes small with the decrease in ζ. Note that the power-saving level of $\zeta = 0.5$ effectively reduces E even for a small c. This result suggests that the power-saving level of $\zeta = 0.5$ is effective both for reducing the energy consumption and for keeping the job-response time small.

We also investigated the job-response time and energy consumption in cases of different λ's, observing the same tendency as Figures 1 and 2. This suggests that

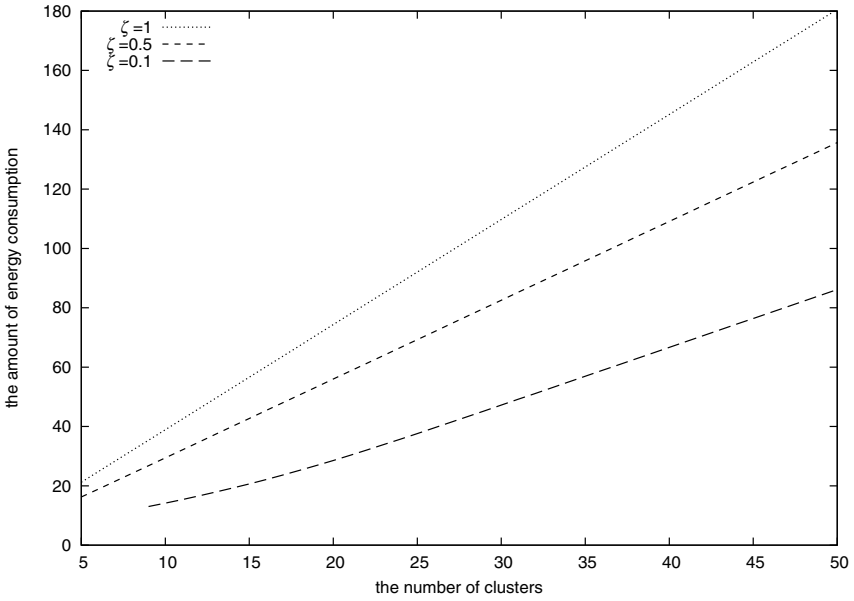


Fig. 2 c vs. E.

turning a half of servers in a cluster off is effective for keeping the mean job-response time small.

5 Conclusion

In this paper, we considered a queueing model for data centers with BEEMR-like energy-saving management mechanism. We modeled a data center as a multiple-server queue with job service depending on a background process. Using matrix-analytic method, we derived the joint distribution of the number of jobs in system and the state of the background process, yielding the mean job-response time and mean amount of energy consumption as performance measures. Numerical examples showed that the amount of energy consumption grows linearly with the increase in the number of clusters. We also confirmed that turning a half of servers in a cluster off is effective for keeping the mean job-response time small.

In our model, the system alternates between normal-operation mode and power-saving one independently of the number of jobs in system. For future work, we consider more practical scheme of power management, with which the system operation mode changes according to the number of jobs in the system.

Acknowledgement This research was supported in part by SCAT Foundation, and Japan Society for the Promotion of Science under Grant-in-Aid for Scientific Research (B) No. 15H04008.

References

1. Chen, Y., Alspaugh, S., Borthakur, D., Katz, R.: Energy efficiency for large-scale MapReduce workloads with significant interactive analysis. In: Proc. The European Professional Society on Computer Systems, pp. 43–56, April 2012
2. Kato, M., Masuyama, H., Kasahara, S., Takahashi, Y.: Performance analysis of energy-saving server scheduling mechanism for large-scale data centers. In: Proc. The 9th International Conference on Queueing Theory and Network Applications (QTNA2014), Bellingham, USA, pp. 28–35, 18–21, August 2014
3. Latouche, G., Ramaswami, V.: Introduction to Matrix Analytic Methods in Stochastic Modeling. ASA-SIAM (1999)
4. Nelson, R.: Probability, Stochastic Processes, and Queueing Theory. Springer Verlag (2000)
5. Schwarts, C., Pries, R., Tran-Gia, P.: A queuing analysis of an energy-saving mechanism in data centers. In: Proc. International Conference on Information Networking, pp. 70–75, February 2012

A Mixed Discrete-Time Delay/Retrial Queueing Model for Handover Calls and New Calls Competing for a Target Channel

Rein Nobel

Abstract To study the performance of handover calls approaching a target cell in combination with arrivals of new calls competing for the same cell, a mixed discrete-time delay/retrial model with one server and with priorities for the delayed customers is discussed. The handover calls are modeled as high-priority customers and the new calls as low-priority customers. The priority is non-preemptive. Upon arrival high-priority customers are put in a queue which is served on a first come first served basis. The behavior of the low-priority customers is modeled as in a retrial queue. Arrivals are in batches and all customers are served individually according to generally distributed and independent service times. The joint steady-state distribution of the queue length of the high priority customers and the orbit size of the low-priority customers is studied using probability generating functions. Several performance measures will be calculated, such as the mean queue length of the handover calls and the orbit size of the new calls. Also the covariance between the queue length and the orbit size will be studied, among others.

Keywords Handover calls · Discrete-time retrial queue · Priority customers · Generating functions

1 Introduction

In mobile telephony the problem how to handle handover calls is a important topic. When a mobile phone user is moving from one cell [the source] to another cell [the target] then his ongoing call has to be switched from the channel of the source cell to a channel of the target cell. Because neighboring cells cover overlapping regions, usually a so-called soft handover protocol is used, i.e. the ongoing call joins a queue,

R. Nobel (✉)
Department of Econometrics, Vrije University, Amsterdam, The Netherlands
e-mail: r.d.nobel@vu.nl

T.V. Do et al. (eds.), *Queueing Theory and Network Applications*,
Advances in Intelligent Systems and Computing 383,
DOI: 10.1007/978-3-319-22267-7_16

waiting for a free channel at the target cell, but the call continues to use the channel of the source cell until a channel at the target cell is available. Meanwhile also new calls will try to get a free channel at the target cell. To avoid unnecessary interruptions of the ongoing calls waiting for a free channel at the target cell, priority is given to the handover calls over the new calls. When all channels are busy new calls are temporarily rejected and the new calls have to be initiated anew some time later.

To model this protocol of soft handover calls at a target cell in combination with the arrival of new calls at this cell we study a mixed delay/retrial model in discrete time with one server [channel]. More specifically, we consider a one-server queueing model in discrete time with two types of customers. Time is divided in slots, and all events [arrivals, start of a service and departures] are considered to occur at the slot boundaries only. The high-priority customers [handover calls] arrive in batches following a general probability distribution. Upon arrival a batch of high-priority customers is put in a waiting line from which the customers are served one by one on a first come first served basis. The low-priority customers [new calls] also arrive in batches (primary arrivals), possibly following a different probability distribution, and when upon arrival a batch of low-priority customers sees the server busy, all incoming low-priority customers are sent into orbit, a virtual waiting space from which they will try to reenter the system individually some random time later (secondary arrivals). The service times of the high-priority and the low-priority customers are all independent and follow [possibly] a different general distribution. To resolve the conflict of simultaneous arrivals and departures we have chosen for the *late arrival set up with delayed access*, i.e. arrivals have precedence over departures and a service of newly arrived customers can only start at the time slot following the slot of the arrival at the earliest. Also the modeling assumption is made that the time slot after *any* departure the server always stays idle, even when high-priority customers are waiting in line. A new high-priority customer will start service the next slot when the queue of high-priority customers is not empty or a batch of high-priority customers will have arrived during the idle slot. In that case all possibly arrived low-priority customers are sent (back) into orbit. Otherwise, i.e. no high-priority customers present at the end of the idle slot, the server starts the service of a low-priority customer, randomly chosen from the mixed batch of primary and secondary low-priority arrivals. When neither high-priority customers are present at the end of the idle slot, nor low-priority customers will have arrived during the idle slot, the server stays idle also the following slot. All customers are served one by one, and in case a low-priority customer is taken into service all other primary and secondary low-priority customers having arrived in the same slot are sent (back) into orbit.

As is well-known *retrial models* have received much less attention in the literature than the well-known queueing models such as delay-models and loss-models, mainly because the arrival stream of the customers consists of two types, the primary arrivals who enter for the first time, and the secondary arrivals from the orbit, making the 'arrival intensity' dependent of the number of customers in the orbit. Also overtaking takes place, i.e. customers are not served according to a specific queueing discipline, which severely complicates the study of the waiting-time distribution of a customer, here defined as the total time that the customer spends in the orbit. It is probably fair

to say that the unpopularity of the research on retrial models is partly due to their intractability, because from a practical point of view retrial models often describe a more realistic picture of many queueing situations than any of the other type of models. Notwithstanding the mathematical difficulties encountered in the study of retrial systems some models, with the $M/G/1$ retrial queue in a prominent position, have been analyzed thoroughly, and we refer to the monographs of FALIN AND TEMPLETON [4] and ARTALEJO AND GÓMEZ-CORRAL [1] for an overview of the main results.

Although most papers on retrial queues discuss models in continuous time, as a consequence of the revolutionary developments in the computer and telecommunication technology, at the end of the past century people started to study also retrial models in discrete time. LI AND YANG [5], [6] and [9] made a start. NOBEL AND MORENO [8] were the first to study a discrete-time classical retrial queueing model with the so-called *late-arrival* setup, i.e. precedence is given to arrivals over departures. We recall that in a classical retrial model an idle server accepts exactly one customer for service from the batch of all the incoming customers [a mixture of primary customers and customers arriving from the orbit] and sends all the other newly arrived customers (back) to the orbit. As a consequence of the late-arrival setup, after a departure the server always stays idle for at least one time slot, due to the fact that the most recently arrived customers have seen the server still busy and therefore they have been sent into the orbit.

In this paper we will extend the classical discrete-time one-server retrial model of NOBEL AND MORENO [8] by adding a second type of customers [the handover calls] who will be put in a queue and are served one by one on a first come first served basis. These customers are given non-preemptive priority over the original customers [the new calls] who continue to act as retrial customers. In a previous paper (NOBEL AND MORENO [7]) the high-priority customers were lost when upon arrival they found the server busy. A model similar to our delay/retrial model has been studied in CHOI AND KIM [2], but they discuss only single arrivals and all customers follow the same service-time distribution. Further, they have chosen the early arrival setup. A continuous-time retrial model with priority customers has been studied by FALIN, ARTALEJO AND MARTIN [3], but in that paper only single arrivals have been considered. The model discussed in this paper can be seen both as an extension and as the discrete-time counterpart of that model.

We will study the joint steady-state distribution of the length of the queue of high-priority customers and the size of the orbit with low-priority customers. Not surprisingly, the mathematical analysis of our mixed delay/retrial model differs greatly from the analysis of the models discussed in the papers [2], [7] and [8].

Firstly, we will derive the generating function of the joint steady-state distribution of the number of low-priority customers in orbit, the number of high-priority customers in the queue and the residual service time of the customer in service [either a high-priority customer, or a low-priority customer]. This generating function will be used to calculate several performance measures, e.g. the mean queue length, the mean orbit size and the covariance of the queue length and the orbit size. In Section 2 we describe the model in detail. Section 3 discusses the steady-state distributions of

the orbit size and the queue length, among others. In Section 4 we derive an expression for the mean busy period. Numerical results will be presented in a forthcoming extended version of this paper.

2 Description of the Model

For a detailed description of the discrete-time setup with late arrivals and delayed access [LAS/DA] we refer to NOBEL AND MORENO [8]. Recall that in the classical retrial model the time slot after a departure the server always stays idle for at least one slot, due to the late-arrival setup with delayed access. For the mixed delay/retrial model to be discussed in this paper we make the technical assumption that the slot following a departure the server always stays idle, *also in case high-priority customers are waiting in the queue*. We can interpret this idle slot as a preparation time for the next service, but we admit that the main reason to include this idle slot following a departure is to enable tractability: a small price to pay for a deeper insight into this mixed delay/retrial model with priorities for the delayed customers.

We will now give the precise description of our discrete-time mixed delay/retrial queueing model with one server and priorities. During each time slot high-priority customers arrive in batches. The batch sizes are mutually independent and follow a general probability distribution $\left\{a_i^{(H)}\right\}_{i=0}^{\infty}$ with probability generating function (p.g.f.)

$$\mathcal{A}_H(y) = \sum_{i=0}^{\infty} a_i^{(H)} y^i.$$

In every time slot also low-priority customers arrive in batches. These batch sizes follow a general probability distribution $\left\{a_k^{(L)}\right\}_{k=0}^{\infty}$ with p.g.f.

$$\mathcal{A}_L(z) = \sum_{k=0}^{\infty} a_k^{(L)} z^k.$$

These batch sizes are again mutually independent and they are also independent of the batch sizes of the high-priority customers. We call these arrivals primary arrivals. Each individual high-priority customer requires a service time, measured as a number of time slots, which follows the discrete probability distribution $\left\{b_j^{(H)}\right\}_{j=1}^{\infty}$ with p.g.f.

$$\mathcal{B}_H(w) = \sum_{j=1}^{\infty} b_j^{(H)} w^j.$$

Similarly, every low-priority customer requires a generally distributed service time with distribution $\left\{b_j^{(L)}\right\}_{j=1}^{\infty}$ and p.g.f.

$$\mathcal{B}_L(w) = \sum_{j=1}^{\infty} b_j^{(L)} w^j.$$

All service times are mutually independent and they are also independent of the batch sizes of the arriving customers. A service time requires at least one time slot, so $b_0^{(H)} = b_0^{(L)} = 0$. As said before, the high-priority customers are placed in a queue and the high-priority customers are served individually on a first come first served basis [within a batch in random order]. Low-priority customers behave as the customers in the classical retrial queue, with the only difference that all incoming low-priority customers [primary and secondary arrivals] are *also* sent into orbit when high-priority customers are present in the queue or arrive simultaneously, i.e. in the same slot, with the low-priority customers. In each time slot low-priority customers try to reenter the system individually and independently with the so-called retrial probability r $[0 < r < 1]$.

We are interested in the steady-state behavior of the number of high-priority customers in the queue, the number of low-priority customers in orbit and the residual service time of the customer currently in service. To analyze the mixed delay/retrial queueing model, we define a discrete-time Markov chain (DTMC) by observing the system at the epochs $k-$, that is at the start of the time slots k just after, possibly, a service of a (low- or high-priority) customer has started, but before the arrivals during time slot k have occurred. We define the following random variables,

H_k = the residual service time of the [high- or low-priority] customer in service at time $k-$,

L_k = the number of high-priority customers present in the queue at time $k-$,

Q_k = the number of low-priority customers in orbit at time $k-$.

We define $H_k = 0$ when at epoch $k-$ the server is idle. Then, due to the independencies stated in the description of the model, the stochastic process $\{(H_k, L_k, Q_k) : k = 0, 1, 2, \ldots\}$ is an irreducible aperiodic DTMC and under the stability condition that

$$\mathcal{A}'_H(1)[\mathcal{B}'_H(1) + 1] + \mathcal{A}'_L(1)[\mathcal{B}'_L(1) + 1] < 1$$

it is positive recurrent. A formal proof of this stability condition can be given using Foster's criterion [see NOBEL AND MORENO [8] for the details]. Notice the '+1' added to the mean service times $\mathcal{B}'_H(1)$ and $\mathcal{B}'_L(1)$, due to our technical assumption that after *every* departure the server stays idle for at least one time slot.

3 The Joint Distribution of Queue Length and Orbit Size

In this section we will derive the joint generating function of the steady-state distribution of the DTMC $\{(H_k, L_k, Q_k) : k = 0, 1, 2, \ldots\}$. Under the stability condition

we can define the following limiting joint distribution of this DTMC

$$\pi(j,m,n) = \lim_{k\to\infty} \mathrm{P}(H_k = j; L_k = m; Q_k = n), \qquad j,m,n = 0,1,2,\ldots,$$

with its associated three-dimensional generating function

$$\Pi(w,y,z) = \sum_{j=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \pi(j,m,n) w^j y^m z^n.$$

In the following it is convenient to introduce also the partial generating functions,

$$\Pi_{jm}(z) = \sum_{n=0}^{\infty} \pi(j,m,n) z^n \quad \text{and}$$

$$\Pi_j(y,z) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \pi(j,m,n) y^m z^n = \sum_{m=0}^{\infty} \Pi_{jm}(z) y^m.$$

To find the p.g.f. $\Pi(w,y,z)$ we write down the system of balance equations,

$$\begin{aligned}
\pi(0,m,n) &= \mathbf{I}_{\{m=0\}} a_0^{(H)} a_0^{(L)} (1-r)^n \pi(0,0,n) + \\
&\quad \sum_{i=0}^{m} a_i^{(H)} \sum_{k=0}^{n} a_k^{(L)} \pi(1,m-i,n-k), \qquad (1)\\
&\quad m,n = 0,1,2,\ldots,\\
\pi(j,m,n) &= \sum_{i=0}^{m} a_i^{(H)} \sum_{k=0}^{n} a_k^{(L)} \pi(j+1,m-i,n-k)\\
&\quad + b_j^{(H)} \sum_{i=0}^{m+1} a_i^{(H)} \sum_{k=0}^{n} a_k^{(L)} \pi(0,m+1-i,n-k)\\
&\quad + \mathbf{I}_{\{m=0\}} b_j^{(L)} a_0^{(H)} \left\{ \sum_{k=1}^{n+1} a_k^{(L)} \pi(0,0,n+1-k) \right.\\
&\quad \left. + a_0^{(L)} \left(1-(1-r)^{n+1}\right) \pi(0,0,n+1) \right\}. \qquad (2)\\
&\quad j = 1,2,\ldots; \quad m,n = 0,1,2,\ldots.
\end{aligned}$$

Notice how our technical assumption that after *any* departure the server stays idle for at least one time slot plays its role in these balance equations. This assumption enforces more parallelism between the [services of] low-priority customers and high-priority customers. Below we will see that only due to this enforced parallelism our analysis can be pursued successfully.

From equations (1) and (2) we get by multiplying both sides with z^n and summing over $n = 0, 1, \ldots$, and subsequently multiplying both sides of the result by y^m and summing over $m = 0, 1, \ldots$,

$$\Pi_0(y,z) = a_0^{(H)} a_0^{(L)} \Pi_{00}((1-r)z) + \mathcal{A}_H(y)\mathcal{A}_L(z)\Pi_1(y,z), \tag{3}$$

$$\Pi_j(y,z) = \mathcal{A}_H(y)\mathcal{A}_L(z)\Pi_{j+1}(y,z) + \frac{b_j^{(H)}}{y}\mathcal{A}_L(z)\left[\mathcal{A}_H(y)\Pi_0(y,z) - a_0^{(H)}\Pi_{00}(z)\right] + \frac{b_j^{(L)} a_0^{(H)}}{z}\left[\mathcal{A}_L(z)\Pi_{00}(z) - a_0^{(L)}\Pi_{00}((1-r)z)\right]. \tag{4}$$

Next, multiplying equation (4) by w^j and summing over $j = 1, 2, \ldots$ gives after some simple algebra, using equation (3) to get rid of $\Pi_1(y,z)$,

$$\begin{aligned} yz\left[w - \mathcal{A}_H(y)\mathcal{A}_L(z)\right]\Pi(w,y,z) &= \mathcal{A}_H(y)\mathcal{A}_L(z)z\left[w\mathcal{B}_H(w) - y\right]\Pi_0(y,z) \\ &\quad + a_0^{(H)}\mathcal{A}_L(z)w\left[y\mathcal{B}_L(w) - z\mathcal{B}_H(w)\right]\Pi_{00}(z) \\ &\quad + a_0^{(H)}a_0^{(L)}wy\left[z - \mathcal{B}_L(w)\right]\Pi_{00}((1-r)z). \end{aligned} \tag{5}$$

So, the problem is to find the unknown partial generating functions $\Pi_0(y,z)$ and $\Pi_{00}(z)$. Firstly, take $w = \mathcal{A}_H(y)\mathcal{A}_L(z)$ in (5) to make the left-hand side zero. This gives

$$\begin{aligned} \Pi_0(y,z) &= a_0^{(H)}\frac{\mathcal{A}_L(z)\left[y\mathcal{B}_L(\omega(y,z)) - z\mathcal{B}_H(\omega(y,z))\right]\Pi_{00}(z)}{z\left[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))\right]} \\ &\quad + a_0^{(H)}\frac{a_0^{(L)}y\left[z - \mathcal{B}_L(\omega(y,z))\right]\Pi_{00}((1-r)z)}{z\left[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))\right]} \end{aligned} \tag{6}$$

where $\omega(y,z) := \mathcal{A}_H(y)\mathcal{A}_L(z)$. Now for any z with $|z| \le 1$ let $w = \phi(z)$ be a solution of the system of equations

$$\begin{cases} w = \mathcal{A}_H(y)\mathcal{A}_L(z) \\ y = w\mathcal{B}_H(w) \end{cases} \iff \begin{cases} w = \mathcal{A}_H(w\mathcal{B}_H(w))\mathcal{A}_L(z) \\ y = w\mathcal{B}_H(w). \end{cases}$$

For real z with $0 < z < 1$ it is easy to see that there is a unique real solution $w = \phi(z) \in (0,1)$ and further that $\phi(1) = 1$. So we have for z with $|z| \le 1$

$$\phi(z) = \mathcal{A}_H(\phi(z)\mathcal{B}_H(\phi(z)))\mathcal{A}_L(z) \tag{7}$$

from which we can calculate the derivative $\phi'(z)$ by implicit differentiation. For future use we give the result

$$\phi'(z) = \frac{\mathcal{A}_H(\phi(z)\mathcal{B}_H(\phi(z)))\mathcal{A}_L'(z)}{1 - \mathcal{A}_H'(\phi(z)\mathcal{B}_H(\phi(z)))[\mathcal{B}_H(\phi(z)) + \phi(z)\mathcal{B}_H'(\phi(z))]\mathcal{A}_L(z)}. \tag{8}$$

From equation (6) we get [notice that now $y = \phi(z)\mathcal{B}_H(\phi(z))$ and $\omega(y, z) = \phi(z)$]

$$\Pi_{00}(z) = a_0^{(L)} \frac{\phi(z)\,[z - \mathcal{B}_L(\phi(z))]}{\mathcal{A}_L(z)\,[z - \phi(z)\mathcal{B}_L(\phi(z))]} \Pi_{00}((1-r)z). \tag{9}$$

Introduce (see also NOBEL AND MORENO [8]) the *retrial function*

$$\mathcal{R}(z) := a_0^{(L)} \frac{\phi(z)\,[z - \mathcal{B}_L(\phi(z))]}{\mathcal{A}_L(z)\,[z - \phi(z)\mathcal{B}_L(\phi(z))]}.$$

We see that $\mathcal{R}(0) = 1$ and after some calculation, using L'Hôpital and result (8) we find

$$\mathcal{R}(1) = a_0^{(L)} \frac{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1) + 1] - \mathcal{A}'_L(1)\mathcal{B}'_L(1)}{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1) + 1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1) + 1]}.$$

Notice that in the denominator the stability condition shows up. Rewriting equation (9) gives via iteration

$$\begin{aligned}\Pi_{00}(z) = \mathcal{R}(z)\Pi_{00}((1-r)z) = \mathcal{R}(z)\mathcal{R}((1-r)z)\Pi_{00}\left((1-r)^2 z\right) = \cdots \\ = \prod_{i=0}^{n-1} \mathcal{R}((1-r)^i z)\Pi_{00}\left((1-r)^n z\right), \end{aligned} \tag{10}$$

and now, sending n to infinity, we get

$$\Pi_{00}(z) = \prod_{i=0}^{\infty} \mathcal{R}((1-r)^i z)\Pi_{00}(0) \tag{11}$$

For the technique to prove the convergence of the infinite product $\prod_{i=0}^{\infty} \mathcal{R}((1-r)^i z)$ we refer to [8]. So, our next problem is to calculate $\Pi_{00}(0)$. From equation (11) we see that is it sufficient to calculate $\Pi_{00}(1-r)$. We plug the result (9) in equation (6). This gives

$$\begin{aligned}\Pi_0(y, z) = a_0^{(H)} \frac{\mathcal{A}_L(z)\,[y\mathcal{B}_L(\omega(y,z)) - z\mathcal{B}_H(\omega(y,z))]\,\mathcal{R}(z) + a_0^{(L)} y\,[z - \mathcal{B}_L(\omega(y,z))]}{z\,[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))]} \\ \times \Pi_{00}((1-r)z). \end{aligned} \tag{12}$$

Because $\Pi_0(1, 1)$ is the long-run fraction of time slots that the server is idle and we can conclude from Little's Law that

$$\Pi_0(1, 1) = 1 - \mathcal{A}'_H(1)\mathcal{B}'_H(1) - \mathcal{A}'_L(1)\mathcal{B}'_L(1)$$

we can find an expression for $\Pi_{00}(1-r)$ using equation (12). Notice that $\omega(y, 1) = \mathcal{A}_H(y)$.

$$\Pi_0(1,1) = \lim_{y\to 1} a_0^{(H)} \frac{[y\mathcal{B}_L(\mathcal{A}_H(y)) - \mathcal{B}_H(\mathcal{A}_H(y))]\,\mathcal{R}(1) + a_0^{(L)} y\,[1 - \mathcal{B}_L(\mathcal{A}_H(y))]}{y - \mathcal{A}_H(y)\mathcal{B}_H(\mathcal{A}_H(y))}$$

$$\times \Pi_{00}(1-r) =$$

$$a_0^{(H)} a_0^{(L)} \frac{\begin{array}{l}\left(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1) - \mathcal{B}'_L(1)]\right)\left(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)\mathcal{B}'_L(1)\right) + \\ -\mathcal{A}'_H(1)\mathcal{B}'_L(1)\left(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]\right)\end{array}}{\left(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1]\right)\left(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]\right)}$$

$$\times \Pi_{00}(1-r) =$$

$$\text{[after some algebra!]} \quad = a_0^{(H)} a_0^{(L)} \frac{1 - \mathcal{A}'_H(1)\mathcal{B}'_H(1) - \mathcal{A}'_L(1)\mathcal{B}'_L(1)}{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]} \Pi_{00}(1-r),$$

from which we find

$$\Pi_{00}(1-r) = \frac{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]}{a_0^{(H)} a_0^{(L)}}.$$

We remark here that interchanging the limits, i.e. considering $\Pi_0(1,1) = \lim_{z\to 1} \Pi_0(1,z)$, leads to the same result, because $\Pi_0(y,z)$ is continuous at the point $(1,1)$, although at first sight the expression looks very different. To double-check our result we give the details. Notice that $\omega(1,z) = \mathcal{A}_L(z)$ and we get

$$\Pi_0(1,1) = \lim_{z\to 1} a_0^{(H)} \frac{\mathcal{A}_L(z)\,[\mathcal{B}_L(\mathcal{A}_L(z)) - z\mathcal{B}_H(\mathcal{A}_L(z))]\,\mathcal{R}(z) + a_0^{(L)}\,[z - \mathcal{B}_L(\mathcal{A}_L(z))]}{z\,[1 - \mathcal{A}_L(z)\mathcal{B}_H(\mathcal{A}_L z))]}$$

$$\times \Pi_{00}((1-r)z) =$$

$$a_0^{(H)} a_0^{(L)} \frac{\begin{array}{l}\left(1 - \mathcal{A}'_L(1)[\mathcal{B}'_L(1) - \mathcal{B}'_H(1)]\right)\left(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)\mathcal{B}'_L(1)\right) + \\ -[1 - \mathcal{A}'_L(1)\mathcal{B}'_L(1)]\left(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]\right)\end{array}}{\mathcal{A}'_L(1)[\mathcal{B}'_H(1)+1](1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1])}$$

$$\times \Pi_{00}(1-r) =$$

[again after some algebra!]

$$= a_0^{(H)} a_0^{(L)} \frac{1 - \mathcal{A}'_H(1)\mathcal{B}'_H(1) - \mathcal{A}'_L(1)\mathcal{B}'_L(1)}{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]} \Pi_{00}(1-r).$$

So, slightly rewriting equation (11), we get an explicit expression for the partial p.g.f. $\Pi_{00}(z)$,

$$\Pi_{00}(z) = \prod_{i=0}^{\infty} \frac{\mathcal{R}((1-r)^i z)}{\mathcal{R}((1-r)^i)} \mathcal{R}(1)\Pi_{00}(1-r)$$
$$= \frac{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)\mathcal{B}'_L(1)}{a_0^{(H)}} \prod_{i=0}^{\infty} \frac{\mathcal{R}((1-r)^i z)}{\mathcal{R}((1-r)^i)}. \quad (13)$$

Next, using this expression for $\Pi_{00}(z)$ and the expression for $\mathcal{R}(z)$ we also get an expression for $\Pi_0(y, z)$ from equation (12). After canceling out common factors we find

$$\Pi_0(y,z) = a_0^{(H)} \frac{\mathcal{A}_L(z)\,[y\mathcal{B}_L(\omega(y,z)) - z\mathcal{B}_H(\omega(y,z))]\,\mathcal{R}(z) + a_0^{(L)} y\,[z - \mathcal{B}_L(\omega(y,z))]}{z\,[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))]}$$
$$\times \Pi_{00}((1-r)z) =$$
$$(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]) \left(\prod_{i=1}^{\infty} \frac{\mathcal{R}((1-r)^i z)}{\mathcal{R}((1-r)^i)} \right)$$
$$\times \frac{\phi(z)\,[y\mathcal{B}_L(\omega(y,z)) - z\mathcal{B}_H(\omega(y,z))]\,[z - \mathcal{B}_L(\phi(z))] + y\,[z - \mathcal{B}_L(\omega(y,z))]\,[z - \phi(z)\mathcal{B}_L(\phi(z))]}{z\,[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))]\,[z - \phi(z)\mathcal{B}_L(\phi(z))]}. \quad (14)$$

Finally, we approach our main goal, an expression for the three-dimensional p.g.f. $\Pi(w, y, z)$. From equation (5) we have

$$\Pi(w,y,z) = \frac{\begin{array}{c}\mathcal{A}_H(y)\mathcal{A}_L(z)z\,[w\mathcal{B}_H(w) - y]\,\Pi_0(y,z) + a_0^{(H)}\mathcal{A}_L(z)w\,[y\mathcal{B}_L(w) - z\mathcal{B}_H(w)]\,\Pi_{00}(z) \\ + a_0^{(H)} a_0^{(L)} wy\,[z - \mathcal{B}_L(w)]\,\Pi_{00}((1-r)z)\end{array}}{yz\,[w - \mathcal{A}_H(y)\mathcal{A}_L(z)]}. \quad (15)$$

For future use it is worthwhile to factorize out the common factor $\Pi_{00}((1-r)z)$ in the numerator. This gives after some manipulations and writing throughout $\omega(y, z)$ for $\mathcal{A}_H(y)\mathcal{A}_L(z)$,

$$\Pi(w,y,z) = \frac{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]}{a_0^{(L)}} \left(\prod_{i=1}^{\infty} \frac{\mathcal{R}((1-r)^i z)}{\mathcal{R}((1-r)^i)} \right)$$
$$\times \frac{\left[\begin{array}{c}\omega(y,z)\,[w\mathcal{B}_H(w) - y]\left(\begin{array}{c}\mathcal{A}_L(z)\,[y\mathcal{B}_L(\omega(y,z)) - z\mathcal{B}_H(\omega(y,z))]\,\mathcal{R}(z) \\ + a_0^{(L)} y\,[z - \mathcal{B}_L(\omega(y,z))]\end{array}\right) \\ + \mathcal{A}_L(z)w\,[y\mathcal{B}_L(w) - z\mathcal{B}_H(w)]\,\mathcal{R}(z)\,[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))] \\ + a_0^{(L)} wy\,[z - \mathcal{B}_L(w)]\,[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))]\end{array}\right]}{yz\,[w - \omega(y,z)]\,[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))]}. \quad (16)$$

Notice that in the denominator still the factor $a_0^{(L)}$ is present because we did not spell out the retrial function $\mathcal{R}(z)$ in the numerator. Doing that would also cancel out the factor $a_0^{(L)}$.

From expression (16) we find the marginal p.g.f.'s $\mathcal{L}(y) := \Pi(1, y, 1)$ and $\mathcal{Q}(z) := \Pi(1, 1, z)$ of the limiting distribution of the queue length and the orbit size, respectively. To get rid of the factor $a_0^{(L)}$ introduce $\mathcal{R}^*(z) = \mathcal{R}(z)/a_0^{(L)}$. Then we find

$$\mathcal{L}(y) = (1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1) + 1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1) + 1]) \frac{1 - y}{y[1 - \mathcal{A}_H(y)]} \times \frac{\mathcal{A}_H(y) \begin{pmatrix} [y\mathcal{B}_L(\mathcal{A}_H(y)) - \mathcal{B}_H(\mathcal{A}_H(y))]\mathcal{R}^*(1) \\ +y\,[1 - \mathcal{B}_L(\mathcal{A}_H(y))] \end{pmatrix} - \mathcal{R}^*(1)\,[y - \mathcal{A}_H(y)\mathcal{B}_H(\mathcal{A}_H(y))]}{y - \mathcal{A}_H(y)\mathcal{B}_H(\mathcal{A}_H(y))}, \tag{17}$$

and, using the definition of $\mathcal{R}^*(z)$ and some further simplification,

$$\mathcal{Q}(z) = (1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1) + 1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1) + 1]) \left(\prod_{i=1}^{\infty} \frac{\mathcal{R}((1-r)^i z)}{\mathcal{R}((1-r)^i)} \right) \times \left(\frac{1 - z}{1 - \mathcal{A}_L(z)} \right) \left(\frac{\phi(z) - 1}{z - \phi(z)\mathcal{B}_L(\phi(z))} \right). \tag{18}$$

Notice that from the expressions (17) and (18) we can check that $\mathcal{L}(1) = 1$ and $\mathcal{Q}(1) = 1$. Of course we can also write down the two-dimensional p.g.f. $\mathcal{T}(y, z) := \Pi(1, y, z)$ of the joint limiting distribution of the queue length and the orbit size,

$$\mathcal{T}(y, z) = (1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1) + 1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1) + 1]) \left(\prod_{i=1}^{\infty} \frac{\mathcal{R}((1-r)^i z)}{\mathcal{R}((1-r)^i)} \right) \times \frac{\left[\begin{array}{c} \omega(y,z)\,[1-y] \begin{pmatrix} \mathcal{A}_L(z)\,[y\mathcal{B}_L(\omega(y,z)) - z\mathcal{B}_H(\omega(y,z))]\mathcal{R}^*(z) \\ +y\,[z - \mathcal{B}_L(\omega(y,z))] \end{pmatrix} \\ +\mathcal{A}_L(z)\,[y - z]\,\mathcal{R}^*(z)\,[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))] \\ +y\,[z-1]\,[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))] \end{array} \right]}{yz\,[1 - \omega(y,z)]\,[y - \omega(y,z)\mathcal{B}_H(\omega(y,z))]}. \tag{19}$$

Because $\mathcal{T}(y, z) \neq \mathcal{L}(y)\mathcal{Q}(z)$ we see immediately that the queue length and the orbit size are dependent. Our next step is to calculate the mean queue length $\overline{\mathcal{L}}$ and the mean orbit size $\overline{\mathcal{Q}}$. Of course we have

$$\overline{\mathcal{L}} = \mathcal{L}'(1) \quad \text{and} \quad \overline{\mathcal{Q}} = \mathcal{Q}'(1).$$

After tedious calculations we find

$$
\begin{aligned}
\overline{\mathcal{L}} = &- \left(\frac{\mathcal{A}''_H(1)}{2\mathcal{A}'_H(1)} + \mathcal{A}'_H(1)\mathcal{B}'_H(1) + \mathcal{A}'_L(1)\mathcal{B}'_L(1) \right) \\
&+ \frac{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]}{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1]} \\
&\times \left[\frac{\mathcal{A}''_H(1)}{2\mathcal{A}'_H(1)} + \left(\mathcal{A}'_H(1) - \frac{\mathcal{A}''_H(1)}{2\mathcal{A}'_H(1)} - 1 \right) \mathcal{B}'_H(1) + \left(1 + \frac{\mathcal{A}''_H(1)}{2\mathcal{A}'_H(1)} \right) \mathcal{B}'_L(1) + \right. \\
&\left. \frac{1}{2}\mathcal{A}'_H(1)\left(\mathcal{B}''_L(1) - \mathcal{B}''_H(1)\right) \right] \\
&+ \frac{\mathcal{A}''_H(1)[\mathcal{B}'_H(1)+1] + \left[\mathcal{A}'_H(1)\right]^2 [\mathcal{B}''_H(1) + 2\mathcal{B}'_H(1)]}{2(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1])} \\
&+ \frac{\mathcal{A}'_L(1)}{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1]} \left[\frac{\mathcal{A}''_H(1)}{2\mathcal{A}'_H(1)} + \left(1 + \frac{\mathcal{A}''_H(1)}{2\mathcal{A}'_H(1)} \right) \mathcal{B}'_L(1) + \mathcal{A}'_H(1)\mathcal{B}'_H(1) + \right. \\
&\left. \frac{\mathcal{A}''_H(1)}{2\mathcal{A}'_H(1)} \mathcal{B}''_L(1) \right]
\end{aligned}
$$

and

$$
\begin{aligned}
\overline{\mathcal{Q}} = &(1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)[\mathcal{B}'_L(1)+1]) \\
&\left\{ \frac{\phi''(1)[1 - \phi'(1)\mathcal{B}''_L(1)] + \left[\phi'(1)\right]^3 \left[2\mathcal{B}'_L(1) + \mathcal{B}''_L(1)\right]}{2\mathcal{A}'_L(1)\left(1 - \phi'(1)[\mathcal{B}'_L(1)+1]\right)^2} \right. \\
&\left. + \left(\frac{\phi'(1)}{1 - \phi'(1)[\mathcal{B}'_L(1)+1]} \right) \left[\frac{\mathcal{A}''_L(1)}{2\left[\mathcal{A}'_L(1)\right]^2} + \frac{1}{\mathcal{A}'_L(1)} \sum_{i=1}^{\infty} \frac{(1-r)^i \mathcal{R}'((1-r)^i)}{\mathcal{R}((1-r)^i)} \right] \right\}.
\end{aligned}
$$

Using equation (8) we can easily evaluate $\phi'(1)$ and $\phi''(1)$ in terms of the p.g.f.'s $\mathcal{A}_L(\cdot)$, $\mathcal{A}_H(\cdot)$, $\mathcal{B}_L(\cdot)$ and $\mathcal{B}_H(\cdot)$. It is more cumbersome to evaluate the terms of the series because for every argument $(1-r)^i$ the calculation of $\mathcal{R}((1-r)^i)$ and $\mathcal{R}'((1-r)^i)$ requires that the values $\phi((1-r^i)$ and $\phi'((1-r)^i)$ are determined as the solution of the two equations (7) and (8) with $z = (1-r)^i$. This solution must be found numerically. We skip further details.

To find the covariance of the queue length and the orbit size we first calculate $\overline{\mathcal{LQ}} := \sum_{i=1}^{\infty}\sum_{n=1}^{\infty} in\pi(1,i,n)$. Using the two-dimensional p.g.f $\mathcal{T}(y,z)$ we have $\overline{\mathcal{LQ}} = \left[\frac{\partial^2}{\partial y \partial z}\mathcal{T}(y,z)\right]_{y=1,z=1}$ and then the covariance is $\mathrm{Cov}(L,Q) = \overline{\mathcal{LQ}} - \overline{\mathcal{L}} \cdot \overline{\mathcal{Q}}$, where we used L and Q as artifact random variables denoting the steady-state queue length and the orbit size, respectively. We do not spell out the long expression for $\overline{\mathcal{LQ}}$, the evaluation simply requires a lot of tedious algebra. We end this section to announce that numerical results for $\overline{\mathcal{L}}$, $\overline{\mathcal{Q}}$ and $\mathrm{Cov}(L,Q)$ will be presented in an extended version of this paper. This work is in preparation.

4 The Mean Busy Period

The busy period in the delay/retrial model is defined as the time lapse from the epoch that the server starts a first service after the server has been idle due to the fact that the system was empty, i.e. no waiting high-priority customers in the queue and no low-priority customers in the orbit, until the first departure epoch leaving behind an empty system again. Introduce B for this busy period and I for the time lapse that the system is empty between two successive busy periods. It is clear that the idle period is geometrically distributed with parameter $1 - a_0^{(H)} a_0^{(L)}$. So, from the the Renewal Reward Theorem we get

$$\pi(0,0,0) = \frac{1 \Big/ \left(1 - a_0^{(H)} a_0^{(L)}\right)}{1 \Big/ \left(1 - a_0^{(H)} a_0^{(L)}\right) + \mathrm{E}[B]}.$$

From (13) we have

$$\pi(0,0,0) = \Pi_{00}(0) = \frac{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)\mathcal{B}'_L(1)}{a_0^{(H)}} \prod_{i=0}^{\infty} \frac{1}{\mathcal{R}((1-r)^i)}.$$

So we get

$$E[B] = \frac{1}{1 - a_0^{(H)} a_0^{(L)}} \left[\frac{a_0^{(H)}}{1 - \mathcal{A}'_H(1)[\mathcal{B}'_H(1)+1] - \mathcal{A}'_L(1)\mathcal{B}'_L(1)} \prod_{i=0}^{\infty} \mathcal{R}((1-r)^i) - 1 \right].$$

References

1. Artalejo, J.R., Gómez-Corral, A.: Retrial Queueing Systems. Springer-Verlag, Heidelberg (2008)
2. Choi, B.D., Kim, J.W.: Discrete-time $Geo_1, Geo_2/G/1$ retrial queueing systems with two types of calls. Computers and Mathematics with Applications **33**, 79–88 (1997)
3. Falin, G.I., Artalejo, J.R., Martin, M.: On the single server retrial queue with priority customers. Queueing Systems **14**, 439–455 (1993)
4. Falin, G.I., Templeton, J.G.C.: Retrial Queues. Chapman & Hall, London (1997)
5. Li, H., Yang, T.: $Geo/G/1$ discrete time retrial queue with Bernoulli schedule. European Journal of Operational Research **111**, 629–649 (1998)
6. Li, H., Yang, T.: Steady-State Queue Size Distribution of Discrete-Time $PH/Geo/1$ Retrial Queues. Mathematical and Computer Modelling **30**, 51–63 (1999)
7. Nobel, R.D., Moreno, P.: A discrete-time priority loss/retrial queueing model with two types of traffic. In: Choi, B.D. (ed.) Proceedings of the Korea-Netherlands Joint Conference on Queueing Theory and its Applications to Telecommunication Systems, Seoul, pp. 189–207 (2005)
8. Nobel, R.D., Moreno, P.: A discrete-time retrial queueing model with one server. EJOR **189**(3), 1088–1103 (2008)
9. Yang, T., Li, H.: On the steady-state queue size distribution of the discrete-time $Geo/G/1$ queue with repeated customers. Queueing Systems **25**, 199–215 (1995)

Author Index