Computers & Operations Research 73 (2016) 43–55

Contents lists available at ScienceDirect

Computers & Operations Research

journal homepage: www.elsevier.com/locate/caor

Strategic queueing behavior for individual and social optimization in managing discrete time working vacation queue with Bernoulli interruption schedule

Miaomiao Yu [a,b,*], Attahiru Sule Alfa [b,c]

[a] School of Science, Sichuan University of Science and Engineering, 643000 Zigong, Sichuan, China
[b] Department of Electrical & Computer Engineering, University of Manitoba, Winnipeg, Manitoba, Canada R3T 5V6
[c] Department of Electrical, Electronic & Computer Engineering, University of Pretoria, 0002 Pretoria, South Africa

ARTICLE INFO

Article history:
Received 13 May 2014
Received in revised form
21 January 2016
Accepted 22 March 2016
Available online 24 March 2016

Keywords:
Queueing
Working vacation
Bernoulli interruption schedule
Joining strategy
Conditional sojourn time

ABSTRACT

In this paper, we consider a discrete time working vacation queue with a utility function for the reward of receiving the service and the cost of waiting in the system. A more flexible switching mechanism between low and regular service states is introduced to enhance the practical value of the working vacation queue. Under different precision levels of the system information, namely observable, almost unobservable and fully unobservable cases, the utility function is studied from both the individual customer's and the system administrator's points of view. By analyzing the steady-state behavior of the system, the associated optimal joining decisions under different information scenarios are obtained. We find that for the fully observable queue, the joining threshold for individual optimization may be less than the one for social optimization in working vacation period. A similar situation also appears in almost unobservable case. Such phenomenon is not possible for the classic first come first served queue due to the fact that there is no vacation time and thus will not cause large fluctuations in customers' conditional waiting time. Additionally, we also conduct some numerical comparisons to demonstrate the effect of the information levels as well as system parameters on customer joining behavior.

1. Introduction

In a queueing system, customers arrive at the service facility to get a certain benefit from the service, but they often encounter the annoyance from waiting. Under normal circumstances, upon arrival at the system, customers usually observe the existing queue length and the service fee, and then decide whether or not to join the queue based on the perceptions of personal benefits. Thus, during the last decades, there exists a widespread tendency to investigate decision making problems in the waiting line system from an economic viewpoint. Some natural reward-cost structures which incorporate customers' desires for service and their unwillingness to wait are imposed on the system. Such an economic analysis for decision making in the queues can be traced to the pioneering work of Naor [1] who studied the situation where arriving customers are admitted or not based on the observed queue length. By establishing a queueing cost model which envisages the imposition of tolls on newly arriving customers, Naor showed that levying tolls is an effective strategy that might attain social welfare optimization. In the several years following this paper, a number of authors have addressed related issues. Yechiali [2,3] extended Naor's results to $GI/M/1$ and $GI/M/s$ queues with one customer type and linear holding cost. Stidham [4] introduced a fixed reward and a waiting cost for each job passing through the system, and considered the optimal control of admission to a queueing system. Mendelson and Whang [5] studied optimal pricing and capacity decisions for a service facility in a microeconomic framework, and also investigated the effects of queueing delays and customer's related costs on the management of computing systems. These studies invariably assumed that customers can balk if the expected waiting cost of their jobs is too high. At the same time, they also demonstrated the inconsistency between the individually and socially optimal joining rules. Additionally, Edelson and Hildebrand [6] reexamined the work of Naor and introduced a balking model in which customers do not observe the system state before making an unchangeable joining decision. They further revealed that revenue maximization and social optimization occur simultaneously under such a situation.

In recent times, there has been an upsurge of interest towards economic analysis for decision making in the waiting line system, especially those classified as optimal design and control of queues. For different precision levels of system information, Burnetas and

* Corresponding author at: School of Science, Sichuan University of Science and Engineering, 643000 Zigong, Sichuan, China.

http://dx.doi.org/10.1016/j.cor.2016.03.011
0305-0548/

Economou [7] considered a Markovian single-server queueing system with setup times. They derived the equilibrium balking strategies for the customers and analyzed the stationary behavior of the system. Based on the above work, Economou and Kanta [8] further studied the equilibrium joining behavior in an $M/M/1$ repairable queue. Following the idea of Economou and Kanta, by assuming that repair is not provided immediately, Wang and Zhang [9] reconsidered the customer's balking strategy in fully and partially observable queues. Moreover, equilibrium analysis is also conducted on an observable queueing system with setup and closedown times by Sun et al. [10], and on a clearing queueing system in alternating environment by Economou and Manou [11]. Meanwhile, accomplishing the development of vacation queueing theory, the economic analysis for decision making in vacation queues has drawn the attention of numerous researchers. Guo and Hassin [12,13] studied a Markovian vacation queue with N-policy and exhaustive service. They presented individually and socially optimal strategies for unobservable and observable queues. An essential extension to queue with general service and vacation times appeared in the recent paper written by Economou et al. [14]. Mean value approach was employed by them for the derivation of the main performance measures. Liu et al. [15] were the first to study customer's strategic queueing behavior in discrete time vacation queue. Shortly after the publication of this seminal paper, Ma et al. [16] developed this research topic by considering the $Geo/Geo/1$ queue with multiple vacation policy, in which customers' equilibrium balking strategies were discussed under four different information scenarios. More recently, inspired by the working vacation mechanism, equilibrium analysis for the $M/M/1$ queue with multiple working vacations are conducted simultaneously by Sun and Li [17] and Zhang et al. [18]. Both individually and socially optimal joining rules are obtained and compared by them. More interesting extensions about the effect of information on the strategic behavior in queueing systems can be found in papers Economou and Kanta [19], Boudali and Economou [20,21], Wang and Zhang [22] and Li et al. [23].

Based on the above brief literature review, we note that significant progress has been made in the continuous time queues with customers' strategic queueing behaviors. But it still seems that little more than a beginning has been made in their discrete time counterparts. Except a limited number of studies done by several Chinese scholars (see, e.g. [15,16]), no work in this direction has come to our notice. Many economic analyses and decision making problems for discrete time queues have been left unexplored. More importantly, for practical measurement related purpose, time is sometimes considered as a discrete quantity although it is continuous. We often hear people say a system is observed every minute, every second, every half a second, etc. Meanwhile, with the development of digital communication technology, many modern communication systems are operated based on a time-slot basis, which are naturally and appropriately modeled by discrete time queues. Thus, the main purpose of this paper is to develop an analytical model that allows us extensively analyze and explore the strategic queueing behavior arising in $Geo/Geo/1$ working vacation queue with Bernoulli interruption schedule.

For evaluating the performance measures of gateway router in fiber communication networks, the concept of working vacation policy was first introduced by Servi and Finn [24] in 2002. A major difference between working vacation queue and classical vacation queue is that during a vacation period, customers in the former can be served in a lower service rate; however, customers in the latter can impossibly be served and depart the system. Due to the strong application background in optimal design of stochastic service systems, working vacation queues have received considerable attention in the past 10 years. Many fruitful theoretical results and interesting applications are presented in this area (see, e.g. [25–37]). On the other hand, in both single and multiple working vacation policies, server resumes fast service rate only when the system is non-empty at the end of a vacation. Obviously, such assumption might not be tenable in many occasions. For example, when a batch of patients who are injured badly in a car accident need surgery, idle surgeons may temporarily interrupt their vacations and return to the hospital for doing emergency surgery. In addition to the above, it is generally known that virus scan is an important maintenance activity for the terminal server which helps keep its functioning well. This type of maintenance could be performed when the system load is relatively light. Although running a virus scan will consume some system resources and result in a slow processing speed, the terminal server could still provide his service with a lower rate. Further, if there are some tasks that require special treatment, the process of virus scan can be interrupted, and the server can resume his original service speed depending on the time-delay sensitivity of these tasks. Thus, it seems that the introduction of the Bernoulli vacation interruption schedule into working vacation queue is very reasonable and necessary for some practical situations. So, investigating individually and socially optimal joining rules for this system would be an interesting and significant research topic. At the same time, it is worth mentioning here that one of the complexities in the analysis of discrete time queues is the occurrences of simultaneous arrivals and departures at the boundary epochs of a slot. The analysis becomes further complicated in the case of working vacation queue with Bernoulli interruption schedule as there exist many directly accessible states for each system state.

The rest of this paper is organized as follows. The next section describes the mathematical model. Sections 3–5 are devoted to the fully observable, the almost unobservable and the fully unobservable queues, respectively. We demonstrate how to obtain the individually and socially optimal joining strategies for each type of queues. Some numerical results are also presented and discussed in these sections. This paper ends with Section 6 where conclusions and future scope are given.

2. Model description

We consider a discrete time multiple working vacation queue with Bernoulli interruption schedule, whose service will not completely remain inactive during the server vacation period. In our model, the inter-arrival times $\{T_r, r \geq 1\}$ of customers are independent and identically distributed random variables with probability mass function (p.m.f.) $\Pr\{T_r = k\} = \lambda\bar{\lambda}^{k-1}$, $k \geq 1$, where we use symbol $\bar{x} = 1 - x$, for any real number $x \in (0, 1)$. The service time S_b in a regular busy period follows geometric distribution with parameter μ_b, namely, $\Pr\{S_b = k\} = \mu_b\bar{\mu}_b^{k-1}$, $k \geq 1$. The server commences a working vacation of random length at the epoch when the system becomes empty. The working vacation time V is geometrically distributed with p.m.f. $\Pr\{V = k\} = \theta\bar{\theta}^{k-1}$, $k \geq 1$. It is an operating period with a lower service rate, the service time S_v in working vacation period follows a geometric distribution with parameter $\mu_v (0 < \mu_v < \mu_b < 1)$, namely, $\Pr\{S_v = k\} = \mu_v\bar{\mu}_v^{k-1}$, $k \geq 1$. After serving a customer in working vacation period, if the server finds any customer waiting in the queue, the vacation either is interrupted with probability $p (0 \leq p \leq 1)$ or continues with probability $\bar{p}$. If there are no customers in the queue, the working vacation continues. Furthermore, after completing a working vacation, if the system is non-empty at that moment, the server switches its service rate from μ_v to μ_b and starts a regular busy period immediately; otherwise, the server takes another working vacation. The low speed service interrupted at the end of working