Performance Evaluation 101 (2016) 20–33

Contents lists available at ScienceDirect

Performance Evaluation

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

Hysteresis control by the number of active servers in queueing system *MMAP*/*PH*/*N* with priority service

Chesoong Kim [a,*], Alexander Dudin [b], Sergey Dudin [b], Olga Dudina [b]

[a] *Sangji University, Wonju, Kangwon, 220-702, Republic of Korea*

[b] *Belarusian State University, 4, Nezavisimosti Ave., Minsk, 220030, Belarus*

HIGHLIGHTS

- Multi-server queueing system with hysteresis strategy of control by the number of active servers is analyzed.
- Ergodicity condition is derived.
- A procedure for computation of the steady state distribution of the system states is presented.
- The problem of optimal choice of hysteresis parameters is discussed.
- Presented results are numerically illustrated.

ARTICLE INFO

Article history:
Received 25 November 2014
Received in revised form 14 January 2016
Accepted 25 April 2016
Available online 3 May 2016

Keywords:
Hysteresis strategy
Marked Markovian arrival process
Generalized phase-type distribution

ABSTRACT

The problem of choosing the optimal hysteresis strategy of control by the number of active servers in the multi-server queue is considered. Customers of two types arrive to the system according to the marked Markovian arrival process (*MMAP*). Type 1 customers have a non-preemptive priority, but the buffer for these customers is finite. The buffer for type 2 customers is infinite. The service time distribution is of phase-type (*PH*) depending on the type of customers. Some servers are always active. The rest of servers can be switched on or off depending on the number of customers in the system. The strategy of control by the number of active servers is of hysteresis type. Such a strategy is defined by two sets of thresholds. The servers are activated or switched off depending on the relation of the number of customers and the thresholds. The main contribution of the paper is development of a procedure for computation of the stationary distribution of the system states and the value of economical cost criterion under any fixed thresholds. Numerical results show effectiveness of the hysteresis control and importance of account of correlation in the arrival process and variance of service times.

1. Introduction

Queueing theory is addressed to the problems of optimization of operation of various systems where arriving at random epochs customers should obtain service at some pool of devices or servers. One of the oldest problems, which was addressed in works by A.K. Erlang, founder of queueing theory, at early 1900s, is to define the optimal (minimal) number of servers sufficient for providing satisfactory quality of service of customers. This problem was called as operator staffing problem in [1]. As recent paper considering this problem, the paper [2] can be mentioned. This direction of research deals with so

* Corresponding author.
E-mail addresses: dowoo@sangji.ac.kr (C. Kim), dudin@bsu.by (A. Dudin), dudin85@mail.ru (S. Dudin), dudina_olga@email.com (O. Dudina).

http://dx.doi.org/10.1016/j.peva.2016.04.002
0166-5316/

called static optimization. Given known arrival and service rates, the optimal, with respect to some fixed cost criterion and constraints imposed on some performance measures of the system, number of servers should be computed.

More advanced direction of research deals with so called dynamic optimization. This direction assumes that the number of active servers may vary in some range depending on the current load of the system, e.g., current queue length. Importance of this direction is explained as follows. Even if there is a lot of servers available for providing service to customers, it may be not reasonable to use all these servers because keeping the servers in active state may be quite costly. This takes place, e.g., in systems where the servers have a finite capacity of batteries or in cloud computing systems where virtual machines are based on real computers having high energy consumption. It also should be taken into account that the server, which is not currently active for providing service to the considered flow of customers, can be effectively used for providing service to some other flows and types of customers.

Thus, the problem of optimal switching on and off the servers depending on the current load of the system is important even for single server systems. The number of papers in this direction, called as queues with server vacations, is huge. We mention only a few of them, namely the survey paper [3], the books [4,5] and the recent papers [6,7], [8]. In this paper, we consider the multi-server queueing systems. Brief survey of the state of the art in the study of multi-server queueing systems with variable number of servers can be found in [9,10]. As earlier work (in Russian) in this subject we can mention the book [11]. Among the papers published after the paper [9] it is worth to mention the papers [12,13].

The overwhelming majority of papers devoted to the problem of optimal control by the number of active servers suggest that the strategy of control belongs to the class of threshold strategies. The threshold strategy is defined, in general, by the set of integer numbers (so called thresholds). A new server is activated when the number of customers in the system or in the queue exceeds one of thresholds while some of the active servers are switched off at the service completion moment when the number of customers in the system or in the queue drops below the corresponding threshold. The threshold strategies are quite reasonable because they react by activating additional server when the system becomes congested and by switching off one of the active servers when the number of customers decreases. It allows to reach some trade-off between the requirement to provide good quality of service for customers (e.g., in terms of the mean sojourn time, the mean queue length or the probability of customer dropping) and the obvious desire of the system manager to reduce expenditures related to maintenance of the redundant servers. A well-known disadvantage of the threshold strategy is the effect of necessity of frequent switching servers on and off (oscillation). Frequent switching may be not desirable if the switching requires some time (setup and removal times, hiring and releasing times, etc.), during which the server cannot provide service, or if the system manager has to pay for every switching. To smooth such a negative effect of frequent switching, the so called hysteresis strategies can be used. Such strategies are quite popular in optimal dynamic control by the arrival and (or) service rate in different queueing models. Likely, the most old reference to the use of such a strategy is the paper [14] where the model with variable service rate was considered. The difference between the threshold and hysteresis strategy in case of two available servers is the following. Instead of fixing one threshold, j, for activating and deactivating the servers as in case of the threshold strategy, to use the hysteresis strategy we "split" threshold j into two thresholds, say j_1 and j_2, $j_1 < j_2$. Then, the additional server is activated when the queue length becomes larger than the larger threshold j_2 and the server is deactivated when the queue length drops below the smaller threshold j_1. When the queue length takes values in the range $[j_1, j_2)$ no actions related to switching the servers are assumed.

The set of the papers where the multi-server queues with hysteresis strategies of control by the number of active servers, to the best of our knowledge, is quite narrow. We can mention the following works. The chapter 2 of book [11] is devoted to analysis of the multi-server queue with infinite buffer, stationary Poisson arrival process and exponential distribution of service times.

Similar model was considered in [15]. More general model (including possibility to have bulk arrivals for the identical servers and the case of arbitrary finite number of non-identical servers) was considered in [10]. As possible field of application of the results of this paper, dynamic resource management in video-on-demand servers was mentioned. Another possible applications are in the field of dynamic activating and deactivating virtual machines in cloud computing networks and in the field of optimal control by staff involvement in contact centers. In the paper [16], the consideration of the model with variable number of servers is motivated by the needs of optimization parameters of the virtual path based fast reservation protocol in ATM.

Taking in mind the same and many other fields of application, in our paper, we consider the following generalization of the models from [10,11,15]. Instead of the stationary Poisson arrival process of customers, we assume that the arrival flow is described by the much more general Marked Markovian arrival process (*MMAP*). This generalization is very important from the point of view of potential applications because the flows of information in modern communication networks are correlated while the stationary Poisson arrival process fails to take correlation in real arrival process into account. The second generalization inspired by possible application for analysis of contact centers is assumption that the customers are heterogeneous. One type is tolerant to long waiting in a buffer while another one is not tolerant. So, non-tolerant customers (type 1 customers) have a non-preemptive priority over the tolerant customers (type 2 customers). The third generalization consists of the assumption that the service time distribution has so called phase-type (*PH*) distribution depending on the type of a customer while only a very special case of *PH* distribution (exponential distribution) was assumed in [10,11,15].

Results of the numerical experiments, partially presented in this paper, show essential effect of correlation and variance in the arrival process and variance in the service process what confirms the necessity of the consideration of more complicated arrival and service processes than in [10,11,15]. Indeed, the profit gained by means of the optimal hysteresis control is more