Transportation Research Part E 75 (2015) 49–66

Contents lists available at ScienceDirect

Transportation Research Part E

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

A spatially distributed queuing model considering dispatching policies with server reservation

Ana Paula Iannoni [a], Fernando Chiyoshi [b], Reinaldo Morabito [c,*]

[a] Laboratoire Genie Industriel, Ecole Centrale Paris, Chatenay Malabry 92295, France
[b] Programa de Engenharia de Produção, Universidade Federal do Rio de Janeiro, 21945-970 Rio de Janeiro, RJ, Brazil
[c] Departamento de Engenharia de Produção, Universidade Federal de Sao Carlos, 13565-905 São Carlos, SP, Brazil

ARTICLE INFO

Article history:
Received 4 August 2014
Received in revised form 18 December 2014
Accepted 19 December 2014
Available online 23 January 2015

Keywords:
Hypercube queuing model
Server reservation
Emergency medical services
SAMU
Cutoff queuing model

ABSTRACT

In this paper we propose a cutoff hypercube queuing model to analyze server-to-customer emergency services operating with server reservation. We are motivated by certain SAMU's (*Système d'Aide Médicale Urgente*) that give assistance to different classes of emergency requests, including specialized transfer of patients, and use the reservation strategy to improve the probability that ambulances will be available to high priority calls. The aim is to show how this cutoff priority service discipline can be handled by the cutoff hypercube queuing model to evaluate relevant system performance measures, and the main impacts of this policy to the different classes of users.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

The main concern of emergency medical services (EMS) is to provide immediate and suitable response to emergency medical requests, where ambulances should arrive at the call location as rapidly as possible transporting specialized personal and equipment (e.g., doctor, rescuers, and medicines). Some of these emergency systems can also receive non-scheduled requests to transfer patients between hospitals and clinics, requiring doctor assistance during the ambulance travel. These pre-hospital systems typically operate within server dispatching policies considering different priorities to the patient calls according to the urgency/seriousness involved. When preemption dispatching policies (interruption of service of low priority services) may not be a feasible option to deal with these priorities, an alternative is to use server reservation policies to high priority calls by assuming that low priority calls may be held on a waiting line until a predetermined number of servers are available. In this way, this dispatching strategy aims to improve the probability that a high priority call finds an available server upon its arrival, while a low priority call may wait in queue even if there are available servers (idling policy).

This dispatching policy is also known in the queuing literature as cutoff service discipline, meaning a refuse of immediate service to lower priority customers when the number of available server is below a threshold number (cutoff level). Usually, highest priority customers are served immediately unless all servers are occupied, in which case they can wait in queue or be lost to the system, depending on the dispatching strategy (Taylor and Templeton, 1980; Sacks et al., 1993). The server reservation policy has been also referred as "*idle-server-based threshold-priority*" when applied to call center systems, since it allows that low priority customers are queued even if there are servers available (Gans and Zhou, 2003; Gurvich et al., 2008).

* Corresponding author.
E-mail address: morabito@ufscar.br (R. Morabito).

http://dx.doi.org/10.1016/j.tre.2014.12.012
1366-5545/© 2014 Elsevier Ltd. All rights reserved.

EMS are server-to-customer systems designed to cover given areas and service spatially distributed demand. A queuing model capable of handling spatially distributed demand is the hypercube queuing model (Larson, 1974; Larson and Odoni, 1981). The incorporation of the geographic characteristics of the area covered by the system into this queuing model is accomplished by: (i) dividing the area into small atoms of demand and (ii) assigning to each atom a list of servers, generally sorted in ascending order of the distance from the atom to the base where the servers are located. Based on such server preference lists, we can model the dispatch policy that sends the first free closest service unit to every incoming call. The basic idea is to expand the state-space description of a multi-server queuing system (e.g. $M/M/N/\infty$ or $M/M/N/N$, where N is the number of servers) in order to represent each server individually and incorporate the complex dispatching policies involved.

Inspired by the real operation of French SAMU systems, in this study we incorporate the server reservation discipline to the hypercube queuing model in order to develop an effective method to analyze service systems operating with distinct classes of users (classes of patient calls) and this particular non–preemptive dispatching policy (queue and servers cutoff). This cutoff hypercube model considers EMS delivery systems with N service units, limited waiting space and three classes of users. (i) Class a: high priority emergency requests, which require immediate attention and cannot wait. (ii) Class b: other emergency requests, which also requires advanced medical assistance but can be held in the waiting line. (iii) Class c: patient transfer (between hospitals, clinics and home) requests. Class a users that find all service unit busy on arrival are referred to a backup service system. Class b users that find all servers busy are kept in the waiting line, if the waiting room is not fully occupied. Otherwise, they are also referred to a backup service system. Class c customers are serviced according to a rule by which they are denied immediate access to service (cutoff) if the number of busy servers is equal to or greater than a predefined number N_1, so that $N–N_1$ reserve servers are kept free for higher priority customers, intending to improve the service availability for those. The queue/loss alternatives are the same of the class b customers. The queued class c customers are serviced when the number of busy servers falls below N_1 (cutoff level). In the queuing theory, the customers referred to a backup system are called ‘‘lost customers’’. They are ‘‘lost’’ in the sense that the main EMS was unable to service them. In this context, referral implies the use of extra resources to provide service of standard quality to referred customers.

The development of a priority hypercube queuing model capable of handling the cutoff service discipline is the incremental contribution of this work. The main objective is to show how the distinct characteristics of a cutoff policy are incorporated and how the associated output measures are evaluated by this cutoff hypercube model. To that end, we use the smallest and non-trivial structure that incorporates the characteristics of systems under analysis. As detailed in Chiyoshi et al. (2011), this kind of structure, referred to as ‘‘toy-model’’, is commonly used to provide useful insights into problems of interest, but is mostly unable to directly address the complexity of real world emergency services. Thus, we first present the cutoff hypercube model by using an illustrative example and then the model is applied to analyze diverse problem instances to obtain key system performance measures. In particular, we discuss the effects of applying the cutoff queuing model to the probability that calls are serviced immediately on arrival, loss fractions and mean waiting times for each user class, when compared to the use of the corresponding priority queuing model without server reservation. We also compare the model results with the outcomes obtained by a discrete simulation model, used as an independent companion model.

This paper is organized as follows: in Section 2 we briefly review the related literature and previous cutoff queuing models and state the contribution of this study. In Section 3 we present the hypercube queuing model and we shortly describe typical urban EMS with dispatching policies with server reservation, such as some SAMU systems. In Section 4 we discuss how to extend the hypercube model to consider cutoff queuing; and in Section 5 we present and analyze some results obtained by applying the model in several experiments. Finally, in Section 6 we present concluding remarks and perspectives for future studies.

2. Literature review and background

Cutoff queuing models have been driven by several applications, such as emergency medical delivery, police dispatching, allocation of beds and surgeries in a hospital, telecommunication channels allocation, computer and network systems, and call-center systems. Only a few and important contributions in the literature address cutoff queuing discipline in server-to-customer emergency systems such as EMS and police patrol, as the studies in Taylor and Templeton (1980), Schaack and Larson (1986, 1989) and Sacks et al. (1993). For instance, Taylor and Templeton (1980) presented two cutoff priority models motivated by applications in urban ambulance deployment with two classes of requests: emergency calls and patients transfer. Low emergency calls (transfer of patients) enter service only if fewer than N_1 servers are busy and $N–N_1$ servers are reserved to high priority calls. In their first model, high emergency calls can wait in queue in case all ambulances are busy, whereas in the second model, they are lost. Schaack and Larson (1986) extended the model of Taylor and Templeton (1980) proposing an N-server queuing model with different priorities. The model was called $M/M/\{N_r\}$, where a queued customer of priority r only enters service when there are fewer than N_r servers busy. They applied this model to a police department data operating with R different priority classes of users, where only some of them require an immediate response. This model was then extended in Schaack and Larson (1989) considering that a police emergency call can request multiple servers and the number of servers requested has a known priority dependent probability distribution.

Sacks et al. (1993) also extended the multiserver queuing model of Schaack and Larson (1986, 1989) for police patrol dispatching and proposed a heuristic procedure to find the set of cutoffs levels N_r that minimize the ‘‘expected total cost’’ of delays for the entire system. This method was also used to investigate how the optimal set of cutoffs changes in response

Download English Version:

https://daneshyari.com/en/article/1023210

Download Persian Version:

https://daneshyari.com/article/1023210

Daneshyari.com