European Journal of Operational Research 242 (2015) 274–285

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

Incorporating priorities for waiting customers in the hypercube queuing model with application to an emergency medical service system in Brazil

Regiane Máximo de Souza[a], Reinaldo Morabito[b,*], Fernando Y. Chiyoshi[c], Ana Paula Iannoni[d]

[a] *Department of Production Engineering, State University of São Paulo – UNESP, SP, Brazil*
[b] *Department of Production Engineering, Federal University of Sao Carlos, SP, Brazil*
[c] *Department of Production Engineering, Federal University of Rio de Janeiro, RJ, Brazil*
[d] *Laboratoire Genie Industriel, Ecole Centrale Paris, France*

ARTICLE INFO

Article history:
Received 6 July 2012
Accepted 26 September 2014
Available online 5 October 2014

Keywords:
OR in health services
Hypercube model
Queuing priority
Emergency medical services
SAMU

ABSTRACT

Emergency medical services (EMS) assist different classes of patients according to their medical seriousness. In this study, we extended the well-known hypercube model, based on the theory of spatially distributed queues, to analyze systems with multiple priority classes and a queue for waiting customers. Then, we analyzed the computational results obtained when applying this approach to a case study from an urban EMS in the city of Ribeirão Preto, Brazil. We also investigated some scenarios for this system studying different periods of the day and the impact of increasing the demands of the patient classes. The results showed that relevant performance measures can be obtained to analyze such a system by using the analytical model extended to deal with queuing priority. In particular, it can accurately evaluate the average response time for each class of emergency calls individually, paying particular attention to high priority calls.

1. Introduction

In emergency medical services (EMS), the response time, which is the interval between the arrival of the emergency call and the arrival of the medical team at the call location, is of major concern as this delay might be the difference between life and death of the patients involved, depending on their medical seriousness. EMS managers must take operational and tactical decisions in these systems in order to evaluate the trade-off between providing suitable care to the system users (service level) and reducing the costs related to the medical resources and capacity (ambulances, stations, specialists and equipment). In Brazil, some urban EMS in moderate to large cities are called SAMU, modeled after the French SAMU (*Système d'Aide Médicale Urgente*). Comprehensive analysis of these systems should take into account the particularities associated to the configuration and operation of the system, such as multiple types of emergency calls, multiple types of system resources, period of the day or week, particular operational policies in locating and dispatching ambulances, among others. Furthermore, more accurate and effective analysis of these systems should also deal with the probabilistic aspects allied to a system's operation, such as the servers' availability, the spatial (location) distribution of calls and servers and the temporal (time) distribution of call arrivals and services.

The hypercube queuing model was initially proposed by Larson (1974) and then extended and applied in various studies (see Galvão & Morabito, 2008; Swersey, 1994; and the references therein). This model can represent systems' uncertainties and incorporates the identity of the servers, as well as the possible cooperation among them. It has been an effective descriptive approach to analyze and plan emergency service systems. Larson's basic model includes priority disciplines for servers' dispatching only if there are servers available when a call arrives. In case all servers are busy, arriving calls wait in a queue (or are transferred to another system) and the queuing calls are served based on a FCFS (First Come First Served) discipline.

Some modifications/extensions of the basic hypercube model include zero capacity queue meaning whereby arriving calls are lost when all servers are busy, for example, in Chelst and Barlach (1981), Mendonça and Morabito (2001), Iannoni and Morabito (2007), Iannoni, Morabito, and Saydam (2009), Larson and Odoni (2007) and Chiyoshi, Iannoni, and Morabito (2011). Some of these studies also include diverse dispatching policies, such as multiple dispatches of ambulances, partial backup, different types of servers and calls, among others. Only a few studies on hypercube models explore limited capacity queue and layering approaches to deal with different types of calls, for example in Burwell, Jarvis, and Mcknew, (1993), Larson and Odoni (2007) and Takeda, Widmer, and Morabito (2007). Alternatively, several researchers address the computational limitations

[*] Corresponding author at: Via Washington Luiz, km 235, 13565-905 São Carlos, SP, Brazil. Tel.: +55 16 33519516; fax: +55 16 33518240.
E-mail address: morabito@ufscar.br, reinaldomorabito@gmail.com (R. Morabito).

http://dx.doi.org/10.1016/j.ejor.2014.09.056
0377-2217/© 2014 Elsevier B.V. All rights reserved.

of the basic hypercube model by proposing approximated versions, where other assumptions are required, such as in Larson (1975), Jarvis (1985) and Atkinson, Kovalenko, and Kuznetsov (2006, 2008). For example, Larson's hypercube approximation proposes correction factors to deal with the assumption of independent and homogeneous servers (all servers with the same service time distribution) when analyzing systems with infinite queue capacity, whereas Jarvis' model extended the later allowing service rates to depend on the server and call locations and null queue capacity. Exact and approximated hypercube models have also been integrated into optimization approaches, for example in Galvão, Chiyoshi, and Morabito (2005), Rajagopalan, Saydam, and Xiao (2008), Ingolfsson, Budge, and Erkut (2008) and Iannoni, Morabito, and Saydam (2011).

An alternative to approximately model distinct classes of users (types of emergency calls) and their respective priorities in the hypercube model is to subdivide the region under analysis into different geographic sub-regions (called atoms) and then subdivide each of these atoms into sub-atoms corresponding to independent sources of the call types. This procedure is referred to in the literature as *layering* (Larson & Odoni, 2007). For example, in Takeda et al. (2007), this procedure was applied to model two classes of calls in the SAMU system in the city of Campinas, Brazil: high priority calls (involving risk to life) requiring advanced life support ambulances and low priority calls requiring basic life support ambulances. Basically, each atom of the system was subdivided into two sub-atoms with different dispatching lists according to the type of resource (ambulance) required. The study showed that in case the utilization level of the ambulances is relatively low, and even without explicitly modeling priority disciplines in queues (i.e., simply using the FCFS discipline combined with layering), the hypercube model can be an accurate method to analyze systems with different types of calls and priorities.

Another policy also adopted by some EMS with multiple priority users and queues is the server reservation, also called cut-off queuing, as studied, for example in Taylor and Templeton (1980) and Schaack and Larson (1986). According to this strategy, servers are reserved to high priority calls by assuming that low priority calls may wait in queue until a predetermined number of servers become available. For example, Schaack and Larson (1986) proposed an N-server cut-off queuing model with different priorities (which is not a hypercube model family model). Their 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 and there are no higher priority customers waiting for service. Other studies addressing problems with different types of calls and/or different types of servers (ambulances), but not studying specifically priorities in queue, can be found, for example in Larson and Mcknew (1982), Goldberg et al. (1990), Iannoni and Morabito (2007) and Chiyoshi et al. (2011).

In some emergency service systems the assistance required differs among the users according to the type of event, resources requested, number of victims/patients and seriousness involved. Examples are found in EMS, police patrol services, emergency services provided by firefighters and also in emergency services to operate in case of major catastrophes such as earthquakes, floods terrorists attacks, among other systems (see e.g., Green & Kolesar, 2004; Larson, 2004; Swersey, 1994). In case of high congested systems, the response process involving priority policies sometimes results in significant divergent waiting times among the user classes, assuming that low priority users may be held behind in a queue if there are also higher priority users waiting for available servers (e.g., ambulance, police patrol, firefighter vehicles). Consequently, the service availability for higher priority users will increase, resulting in an improved quality of service for them.

Appropriate analysis of such systems should explicitly take into account priority policies in queues in order to evaluate accurate response times for the different classes of users. In view of this, the present work studies extensions to the hypercube queuing model in order to properly deal with these situations. The extended model can be applied to these server-to-customer emergency systems cited above to evaluate the response time of each user class individually. The response time includes the waiting time in a queue, in case all servers are busy, the set-up delay, the travel time between the server and call locations and other possible delays occurring after the assignment of ambulances that are recorded as part of the service time statistics. Therefore, in the present study the response time is evaluated as the sum of the waiting time in queue and the mean travel time between the ambulance and call locations.

Hence, in this study we propose a modified hypercube queuing model to analyze emergency systems involving different priorities in queue, assuming a non-preemptive priority discipline, a finite queue size and distinct servers. To the best of our knowledge, there are no other studies in the literature explicitly exploring a priority policy to distinct classes of users in queues using the hypercube model. This model is then applied to analyze a case study of the SAMU system of Ribeirão Preto, a medium-sized city located in Sao Paulo State, Brazil. Similar to other SAMU systems, this EMS provides assistance to different classes of users, including high priority emergencies that require advanced life support ambulances, as well as moderate to low priority emergencies requiring basic life support ambulances. The computational results obtained by applying the model to this EMS are useful to evaluate performance measures for different periods of the day. We also compared the results obtained with a discrete event simulation model that was used to assess the effects of relaxing some restricting hypothesis assumed by the hypercube model, such as the Poisson arrival process, exponentially distributed service times indistinguishable for users' regions and class, a finite queue size and mean travel time inputs indistinguishable for user class (as listed in Section 4). Furthermore, the model results were also compared with sample data of the system.

This paper is organized as follows: Section 2 presents the extended hypercube model to deal with different priorities in queues and Section 3 presents a general procedure to generate the equilibrium probabilities for the system states of the extended model. Section 4 describes the case study, its statistical analysis and the simulation model. The computational results obtained by applying the model to the case study and to alternative scenarios are discussed in Section 5. Finally, Section 6 presents some concluding remarks and discusses some perspectives for future research.

2. The hypercube model with layers and different priorities in queues

The basic idea of the hypercube model is to expand the state space description of a simple multi-server queuing system ($M/M/m/$ or $M/M/m/m$ system, where m is the number of servers) in order to represent each server individually and account for the spatial nature of the system. Supposing that each server has two possible statuses: idle (0) or busy (1) at any given instant, there are 2^m possible system states in the hypercube (i.e., the possible server configurations), which correspond to the vertices of an m-dimensional hypercube (with 2^m vertices), plus the number of queuing states. The model requires the solution of linear systems, where the variables involved are the equilibrium (steady) state probabilities of the system. With these probabilities, the different steady-state performance measures can be calculated, such as mean user response times, server workloads, number of dispatches per server in each region, among others.

The analysis implies that the region under study is divided into geographic atoms, which corresponds to independent sources of calls. It is assumed that the calls arrive at each atom j according to a Poisson process, regardless of other atoms, with arrival rate λ_j. The m servers (ambulances) of the system are spatially distributed and when available, their dispatching occurs according to a preference dispatching list predefined for each atom. The model assumes negative exponential service times. In general, each server i has a distinct mean service