Transportation Research Part E 103 (2017) 143–157

Contents lists available at ScienceDirect

Transportation Research Part E

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

An approximate hypercube model for public service systems with co-located servers and multiple response

Sardar Ansari [a], Soovin Yoon [b], Laura A. Albert [b,*]

[a] University of Michigan, Ann Arbor, MI 48109, United States
[b] University of Wisconsin-Madison, Department of Industrial Engineering & Systems Engineering, 1513 University Avenue, Madison, WI 53706, United States

ARTICLE INFO

Article history:
Received 19 October 2016
Received in revised form 20 April 2017
Accepted 26 April 2017

Keywords:
Spatial queues
Hypercube model approximation
Emergency response
Emergency medical services

ABSTRACT

Spatial queueing models help to evaluate the design of public safety systems such as fire, emergency medical service, and police departments, where vehicles are sent to geographically dispersed calls for service. We propose a new approximate hypercube spatial queueing model that allows for multiple servers to be located at the same station as well as multiple servers to be dispatched to a single call. We introduce the $M[G]/M/s/s$ queueing model as an extension to the $M/M/s/s$ model which allows for a single customer to request multiple servers with a general discrete probability distribution G. We use the $M[G]/M/s/s$ queueing model to derive approximate formulas for the hypercube spatial queueing outputs. A simulation study validates the accuracy of the queueing approximations. Computational results suggest that the models are effective in evaluating the performance of emergency systems.

1. Introduction

Evaluating the performance of public safety systems is critical to ensure that effective care is provided to those in need. The speed of response is one of the primary performance measures for public safety systems such as fire and emergency medical services. This motivates the development of a model that can accurately and efficiently quantify performance measures such as the availability of each vehicle and the dispatch probabilities. Analytical models are needed to evaluate public safety systems when multiple vehicles are co-located at stations and when multiple vehicles are sent to the same call. However, most models in the literature assume that vehicles are located at distinct stations and that one vehicle is sent to each call. This paper addresses this gap in the literature and proposes a new model that lifts both of these assumptions. The proposed model can be used by public safety leaders to inform decisions such as vehicle location and district design decisions.

For several decades, spatial queueing models have supported the design and analysis of public safety systems. The methods used to evaluate the performance of an emergency system are usually based on exact and approximate hypercube spatial queueing models. Larson (1974) provides an exact hypercube model for capturing the statistical dependence between the vehicles serving an area using queueing methods. Computing the exact values for these queueing factors in spatial systems is computationally intractable due to the so-called "curse of dimensionality." As a result, several approximation hypercube methods have been proposed in the literature that are manageable in terms of size and complexity (Larson, 1975). The hypercube models have been used to inform a range of system design decisions in several real settings, including Boston, New York and Orlando, to analyze and study the travel times (Brandeau and Larson, 1986; Larson and Rich, 1987; Sacks

* Corresponding author.
E-mail addresses: sardara@med.umich.edu (S. Ansari), yoon57@engr.wisc.edu (S. Yoon), laura@engr.wisc.edu (L.A. Albert).

http://dx.doi.org/10.1016/j.tre.2017.04.013
1366-5545/

and Grief, 1994). More recently, Larson (2004) used the hypercube model as a deployment model to respond to emergency situations such as terrorist attacks.

Both the exact and approximate hypercube models assume Poisson arrival rates, exponential service times that are independent of call locations, and with one vehicle assigned to each call. Several papers lift the assumptions made in early hypercube models. Halpern (1977) improves the accuracy of the hypercube model by allowing server dependent service times. Jarvis (1985) further extends the hypercube model approximation to allow for service times that depend on both the vehicle and the call locations. Other extensions of the hypercube model relax the assumptions of the original model or improve its computational complexity and are provided by Chelst and Jarvis (1979), Larson and Mcknew (1982), and Mendonça and Morabito (2001). Geroliminis et al. (2009) develop an exact hypercube model with service times that depend on the responding vehicle, and they embed the hypercube model in a location model that seeks to minimize the mean response time subject to a coverage level target. Later, Boyaci and Geroliminis (2015) formulate an aggregate hypercube queueing model where each server has three states: available, busy with intradistrict calls, and busy with interdistrict calls. However, the size of the state space increases from 2^n as in the Larson (1974) model to 3^n. de Souza et al. (2015) develop an exact hypercube model that considers different priorities in a finite-capacity queue.

Thus far, all models have assumed that exactly one vehicle is assigned to each call for service. A customer who requests i servers can be thought as a batch (bulk) arrival of i customers who each request one server in a $M^{(x)}/M/s/s$ batch arrival queueing system. However, in a batch arrival system customers may enter service one at a time and each server assigned to the batch of customer completes service separately. In this paper, we consider servers who complete service at the same time. Similarly, in a general batch queue where a group of customers arrive and are serviced in groups, the arrival group does not necessarily coincide with service group (Miller, 1959).

A series of papers considers a multi-server queueing system in which customers require a random number of servers. Green (1980) explores a queueing system in which servers begin service simultaneously but become available independently. For the same system, Seila (1984) derives the second moment of time a customer spends in queue. Green (1981) compares an alternative service order disciplines to first-in-first-out (FIFO) under various degree of dependence among servers. Brill and Green (1984) use a system point approach to derive the waiting time distribution for customers who need simultaneous service from a random number of servers. Fletcher et al. (1986) derive performance measures for a closed single node queueing system with multiple classes, where each class requires a different number of servers. Recently, Vinayak et al. (2014) analyze the system in which customers require a random number of servers with the queueing disciplines of retrial and preemptive priority. Unlike in our model, partial dispatch is not allowed in these papers, and therefore a customer cannot enter service until all required servers are available.

Few hypercube spatial queueing models in the literature study the impact of multiple response, where more than one vehicle is assigned to a call. Chelst and Barlach (1981) develop a model based on the hypercube model that sends one or two vehicles to a single call. Daskin and Haghani (1984) study the distribution of the arrival time of the first vehicle that arrives at the scene when multiple vehicles are dispatched. McLay (2009) examines the issue of vehicle location when there is multiple response to prioritized calls. Iannoni and Morabito (2007) extend the hypercube model by proposing a model that dispatches one, two, or three vehicles to a single call. Their model assumes a specific dispatching policy designed for emergency medical vehicles that operate on Brazilian highways. On the contrary, the approximate hypercube model that is proposed in this paper is neither restricted by the number of vehicles dispatched, nor does it assume a particular dispatching policy.

Backup coverage models are developed to optimally locate multiple servers while taking server unavailability into account. Daskin and Stern (1981) formulate a hierarchical multiobjective program that locates emergency medical service vehicles. The primary goal is to find the minimum number of vehicles needed to cover all zones while the secondary objective is to maximize multiple coverage. Daskin (1983) develops a maximum expected covering location problem that maximizes the number of calls covered under the assumption that each server has the same independent busy probability. Hogan and ReVelle (1986) modify the model by Daskin and Stern to maximize the population receiving backup coverage. ReVelle and Hogan (1989) maximize the proportion of demand that can be reached within a time standard with a given level of reliability. This approach is extended by Marianov and ReVelle (1996) to consider the dependencies between servers using a queueing model. Models that can approximate dispatch probabilities, such as the approximate hypercube model formulated in this paper, can be used to facilitate comparison of multiple system design alternatives, for example, within an optimization model that locates servers and considers backup coverage.

Much of the previous work in approximate hypercube spatial queueing models assumes one vehicle is sent to a call and that there is one vehicle per station. There are two notable exceptions that allow for co-located servers. Burwell et al. (1993) and Budge et al. (2009) propose modifications to the hypercube model approximation to accommodate vehicles co-located at a single station through "preference ties," i.e., when multiple vehicles are equally preferred to respond to a call. This model requires less computer time and storage than earlier models that accommodate preference ties.

We extend the model by Budge et al. (2009) to accommodate multiple response, thus lifting both of the assumptions commonly found in the literature. To do so, we extend the $M/M/s/s$ loss model and introduce a $M[G]/M/s/s$ model that allows multiple servers to serve a single call. The distribution of the number of servers that are requested for service per call is determined by the general discrete probability distribution G. We present a procedure to compute the steady-state probabilities for this model by solving the balance equations. We use the $M[G]/M/s/s$ model to derive approximate formulas for the hypercube spatial queueing outputs.