Omega ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Omega

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

Analysis and optimization of an ambulance offload delay and allocation problem ☆

Eman Almehdawe [a,*], Beth Jewkes [b], Qi-Ming He [b]

[a] Faculty of Business Administration, University of Regina, 3737 Wascana Parkway, Regina, Canada S4S 0A2
[b] Department of Management Sciences, University of Waterloo, 200 University Avenue West, Waterloo, Canada N2L 3G1

ARTICLE INFO

Article history:
Received 12 June 2014
Accepted 11 January 2016

Keywords:
EMS
Offload delays
Queueing networks
Optimization

ABSTRACT

Ambulance offload delays have recently become one of the most significant operational challenges for Emergency Medical Services (EMS) providers. Offload delays occur when an ambulance arriving at a hospital Emergency Department (ED) is blocked until a bed becomes available for the patient. To formally investigate the effect of patient routing decisions on EMS offload delays, we introduce a stylized queueing network model with blocking. Following a decomposition approach, we develop an approximation scheme to find explicit solutions that can be used to find proper patient allocation policies to multiple hospitals in a region. We introduce a Markov chain representation for a single ED network and solve for its exact steady state distribution. A comprehensive numerical study is carried out to validate the approximation approaches and to gain insight into ambulance offload delays. By keeping the total offload delays at minimal levels, we observe that it is better to load larger EDs more heavily than smaller ones due to resource pooling.

1. Introduction

Emergency Medical Services (EMS) are responsible for transferring patients to Emergency Departments (ED) within a target response time. Sometimes, upon arrival at a highly congested ED, an ambulance is forced to wait to offload a patient until a bed becomes available. This waiting time is referred to as *offload delay* in North America, or *access block* in Australia. In some countries, such as the United States, an ED can declare *diversion* status if it is overcrowded [1]. For EMS management, diversion means that patients should be routed to other less crowded EDs. Diversion, or reallocating patients to another regional hospital, can be key in alleviating overall offload delays experienced, but entails higher costs to healthcare systems. In addition, offload delays increase actual EMS response times and waste scarce resources. Thus, offload delays are a pressing concern for EMS management.

Ambulance offload delays have attracted the attention of researchers and practitioners in the past decade. Studies on ambulance offload delays can be categorized as either observational or analytical. Observational studies focus on identifying the relationships between offload delays and hospital congestion. A survey of such studies can be found in Ting [2] and Taylor et al. [3]. More details can be found in Silvestri et al. [4] and Silvestri et al. [5]. Analytical studies on offload delays utilize queueing theory. Specifically, such studies are related to queueing networks with blocking and priority. Queueing models are widely used in service systems analysis to improve customer service. We refer the reader to Formundam and Herrmann [6] and Green [7] for a comprehensive review on the use of queueing theory in healthcare systems. We also refer to Almehdawe et al. [8] for a summary of some queueing works related to hospital bed use and allocation. Other related references are Kao and Tung [9], Gorunescu et al. [10], Davies and Davies [11], Masselink et al. [12], Côté and Stein [13], Knight et al. [14] and Gorunescu et al. [15].

The problem of ambulance allocation to regional hospitals was studied recently by Leo et al. [16]. In their Mixed Integer Programming model, they consider allocating ambulance and walk-in patients simultaneously to regional hospitals by minimizing their travel and waiting times. Then they recommend reorganization of the EMS network based on those results. Compared to the model developed in this paper, we assume that only patients arriving by an ambulance can be allocated by the EMS dispatcher, while walk-in patients select by themselves the ED to which they will go.

The queueing network investigated in this paper is introduced for the analysis and design of EMS and is similar in structure to that in Almehdawe et al. [8], although the objectives, model assumptions, and methodologies are different. While the objective of Almehdawe et al. [8] is to conduct a performance analysis of EMS, the objective of this paper is to develop an optimization

☆This manuscript was processed by Associate Editor Ghate.

* Corresponding author. Tel.: +1 306 585 4728.
E-mail address: Eman.Almehdawe@uregina.ca (E. Almehdawe).

http://dx.doi.org/10.1016/j.omega.2016.01.006
0305-0483/

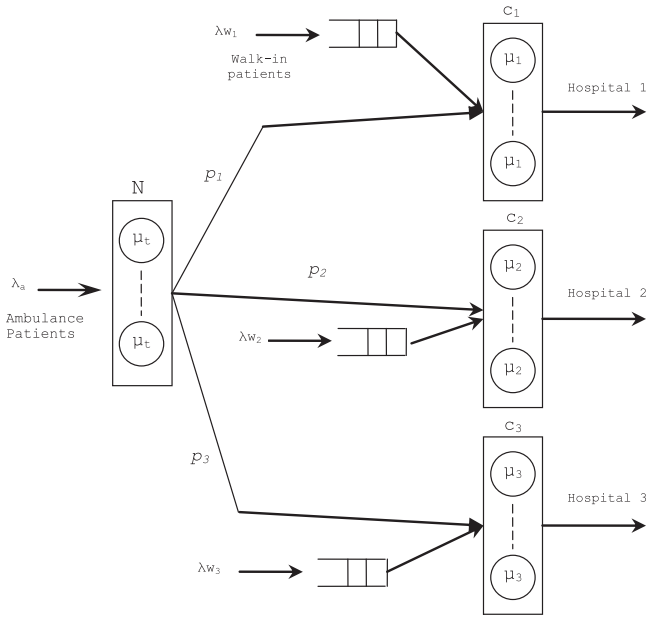

Fig. 1. An EMS-ED queueing network for a region of 3 hospitals.

method for the design of EMS. In Almehdawe et al. [8], the preemption priority is assumed for ambulance patients. In this paper, the non-preemption priority is assumed. In Almehdawe et al. [8], the ambulance transit time is considered to be negligible. In this paper, the ambulance transit time is assumed to have an exponential distribution. As a result, the queueing network becomes far more complicated and the method developed in Almehdawe et al. [8] to derive the steady state probability distributions does not work well due to the curse of dimensionality. In addition, the method used in Almehdawe et al. [8] is not effective in optimization. Thus, a decomposition approach is utilized in this paper, and the matrix-analytic methods and some classical queueing results are applied, at both the hospital level and the region level, to analyze two levels of the queueing network individually.

In the queueing network developed in this paper (see Fig. 1), two types of patients are served: ambulance patients and walk-in patients. We assume that ambulance patients have higher service priority over walk-in patients (see [8] for details on the validity of this assumption). Our approach is to decompose the queueing network into subsystems, each containing one ED. First, we introduce an approximation model for each individual system so that explicit (but approximate) results can be obtained for performance measures such as the mean waiting time of ambulance patients (offload delays). The results are used to select system parameters to minimize the mean offload delays in a given region. Second, we study individual subsystems analytically. Under certain assumptions, we use results from individual subsystems to produce various performance measures of the queueing network. The exact results for individual subsystems are used to check the quality of the parameters selected by using the approximate models. In summary, in this paper, we develop a method to find how to allocate ambulance patients to different hospitals in a given region. We make the following contributions:

- We model the complex problem of offload delays in terms of hospital congestion and EMS system congestion.
- We develop an approximation scheme for individual EDs performance measures and validate the approximation via simulation.
- We construct and solve an optimization problem to find the optimal allocation of ambulance patients to each ED in a region.
- We find explicitly the waiting time distribution for a multi-server queueing system with non-preemptive priorities and blocking.

The rest of the paper is organized as follows. In Section 2, we introduce the EMS system of interest and describe the steps for model approximation and the optimization of the allocation of ambulance patients. In Section 3, we introduce an $M[2]/M/c$ non-preemptive priority queue and an optimization problem for the allocation of ambulance patients. In Section 4, we define a One-ED network and apply matrix-analytic methods to validate the approximation scheme of Section 3. A numerical analysis is carried out in Section 5, where issues such as model validation and optimal allocation of ambulance patients are addressed. Section 6 concludes the paper. Some technical details are collected in Appendices 1–3.

2. The EMS system and solution approaches

We consider an EMS system with N ambulances that serve K hospitals, each with an ED. When the dispatching center receives an emergency call requiring an ambulance, an ambulance is dispatched to the call scene, if one is available. Upon arrival, the paramedic team apply the basic life saving procedure. If the patient requires transport to a hospital, the paramedics load him into the ambulance. Then they transfer the patient to one of the K regional hospitals. The time to reach the patient, load him into the ambulance and then transfer him into the ED is referred to as the patient *transit time*. We refer to such patients as *ambulance patients*. Patients may alternatively go to one of the K EDs by themselves. We refer to such patients as *walk-in patients*. In each ED, both ambulance patients and walk-in patients are served. Ambulance patients have higher service priority. That is: when an ED bed becomes available, it will be assigned to a waiting ambulance patient. If there is no waiting ambulance patient, the bed becomes available to walk-in patients. The service of both types of patients cannot be interrupted. Thus, if an ambulance patient sees all ED beds are occupied upon arrival (i.e., no resource to serve the patient), the patient and its ambulance have to wait, and this waiting time is referred to as *ambulance offload delay*. The ambulance becomes available when the ambulance patient is admitted to the ED. Within each priority group of patients at one ED, patients are served on a first-come-first-served basis. A patient leaves the system immediately after his service is done.

Focusing on the movement of ambulances and patients, the EMS system can be modeled as a queueing network as shown in Fig. 1, which is defined explicitly as follows:

Patient arrival processes: Ambulance patients call the EMS according to a Poisson process with parameter λ_a. Walk-in patients arrive at the k-th ED according to a Poisson process with parameter $\lambda_{w,k}$, for $k=1,2,\ldots,K$. All the Poisson processes are independent. The assumption of Poisson arrivals is supported by empirical studies (see Channouf et al. [17] and the references therein). Justification of the arrival processes can also be found in Almehdawe et al. [8].

Routing probabilities: Upon arrival, an ambulance patient will be transferred to the k-th ED with routing probability p_k, for $k=1,2,\ldots,K$, if an ambulance is available at that moment; otherwise, the patient is lost. Thus, we must have $p_1+p_2+\cdots+p_K=1$. One of the main issues addressed in this paper is how to choose the routing probabilities to minimize the overall offload delays.

Patient transit time: The transit time from dispatching the ambulance to the call scene until it arrives to a hospital is exponentially distributed with parameter $\mu_{T,k}$. Mateo Restrepo and Topaloglu [18] and the references therein use this assumption and

Please cite this article as: Almehdawe E, et al. Analysis and optimization of an ambulance offload delay and allocation problem. Omega (2016), http://dx.doi.org/10.1016/j.omega.2016.01.006