Journal of the Operational Research Society

ISSN: 0160-5682 (Print) 1476-9360 (Online) Journal homepage: https://www.tandfonline.com/loi/tjor20

Optimization in a two-stage multi-server service system with customer priorities

Eman Almehdawe, Beth Jewkes & Qi-Ming He

To cite this article: Eman Almehdawe, Beth Jewkes & Qi-Ming He (2019) Optimization in a two-stage multi-server service system with customer priorities, Journal of the Operational Research Society, 70:2, 326-337, DOI: 10.1080/01605682.2018.1438762

To link to this article: https://doi.org/10.1080/01605682.2018.1438762

Published online: 01 Mar 2018.

JOURNAL OF THE OPERATIONAL RESEARCH SOCIETY
2019, VOL. 70, NO. 2, 326–337
https://doi.org/10.1080/01605682.2018.1438762

THE OPERATIONAL RESEARCH SOCIETY
Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLES

Optimization in a two-stage multi-server service system with customer priorities

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

[a]Faculty of Business Administration, University of Regina, Regina, Canada; [b]Management Sciences Department, University of Waterloo, Waterloo, Canada

ABSTRACT
We consider a two-stage, multi-server queueing network that serves two types of customers, which we refer to as type *a* and type *b*. Type *a* customers require service at both sequential stages and type *b* customers only require service at the second stage. The first stage has one node and the second stage has multiple nodes. Type *a* customers possess a higher non-pre-emptive priority than type *b* customers. Depending on the model application, two goals are explored: the first goal is to allocate type *a* customers to the second- stage nodes in a manner that minimizes the average blocking delay; the second goal is to optimize the service speed of each server in the second stage so that the average blocking delay experienced by type *a* customers is minimized. In this paper, we develop an approximation scheme and an iterative algorithm to find stationary policies, which we then apply to the real-world contexts of Emergency Medical Services planning and airline staffing. Numerical examples show that, compared to some typical heuristic schemes (e.g., proportional allocation based on arrival/service capacity), the suggested allocation policies result in type *a* customers experiencing shorter delays and allow more of them receive service.

ARTICLE HISTORY
Received 28 February 2017
Accepted 5 February 2018

KEYWORDS
Optimization; queueing models; service systems; work load allocation

1. Introduction

The research on workload allocation in queueing networks has been driven by different applications. The objective of workload allocation usually depends on the characteristics of the specific application. In Flexible Manufacturing Systems, for example, the decision is concerned with how to allocate work among groups of flexible machines in order to avoid bottlenecks in the system (Calabrese, 1992). In multi-facility production systems, a decision- maker decides how to assign work to machines in a way that minimizes the total work-in-process inventory (Benjaafar & Gupta, 1999). Likewise, in a Local Area Network, the system administrator allocates data files to multiple servers in order to minimize the overall system response time (Lee & Park, 1995). This work is motivated by an application in Emergency Medical Services (EMS). In this application, a decision-maker decides on how to allocate patients arriving by an ambulance to regional emergency rooms such that patient delays due to blocking are minimized. Additionally, this model is also applied to staffing decisions for special service employees at gates in an airport. In this setting, a decision-maker decides on staff allocation given the demand rate and resources available at each gate.

In this paper, we consider a two-stage, multi-server queueing network that serves two types of customers, which we refer to as type *a* and type *b*. Type *a* customers receive service at stage one before being allocated to one of the K nodes at stage two. If all stage-one servers are busy, then the customer is lost. There is no queue between stage one and stage two for type *a* customers and servers at stage one can be blocked if there is no server available at stage two to serve type *a* customers. Type *b* customers require service at one of the K nodes in the second stage only. In order to reduce the frequency of blocking in the first-stage servers, type *a* customers are assigned a higher non-pre-emptive priority than type *b* customers. In this model, the decision-maker's goal is to minimize the total blocking delays experienced by type *a* customers by efficiently allocating those who finish service at stage one to one of the K nodes at stage two.

The model has applications in different areas such as ambulance dispatching, call centre routing, airport gate service. In the context of ambulance dispatching, EMS ambulances are responsible for transferring patients to Emergency Departments (EDs) on a regional basis. Sometimes, upon arriving at a highly congested ED, an ambulance must wait for a bed to become available before being able to offload the patient. On the other hand, the ED must also serve patients who arrive by themselves (i.e., walk-in patients). Baker, Clayton, and Taylor (1989) developed a non-linear integer optimization model to allocate EMS ambulances to sectors within a county. Their model objective is to meet government-mandated response-time criterion. However, in this work, our objective is to minimize the total

CONTACT Eman Almehdawe ✉ eman.almehdawe@uregina.ca

offload delays experienced by the EMS ambulances. In a call centre setting, there are high-priority calls that are served on a common server and then routed to different severs to continue processing. In an airport gate setting, there are regular passengers and special needs passengers: while regular passengers are able to get to the gates by themselves, special needs passengers require additional service to do so.

There is a considerable amount of literature on the analysis of blocking in queueing networks when one or more nodes have multiple servers (e.g., Andriansyah, Van Woensel, Cruz, & Duczmal, 2010; Cruz & Smith, 2007; Gelenbe, Pujolle, & Nelson, 1998; Han & Smith, 1991; Jain & Smith, 1994; Stidham, 2009). While many existing works consider general topology networks, there are some papers that analyse special queueing networks, such as tandem queues consisting of two or more nodes (e.g., Akyildiz, 1989; Latouche & Neuts, 1980; van Vuuren, Adan, & Resing-Sassen, 2005). Almehdawe, Jewkes, and He (2013) studied a multi-server network with customer service priority. As such, the queueing network in this paper was motivated by the study of the EMS model considered in Almehdawe et al. (2013) and Almehdawe, Jewkes, and He (2016). In Almehdawe et al. (2016), two approximations were developed: the first approximation provides a computation method for performance measures, and the second provides explicit results that can be used to optimally allocate patients arriving by an ambulance (which we refer to thereafter as ambulance patients). These approximation methods work well under the condition that the probability of losing high-priority ambulance patients (i.e., type a customers) is close to zero. In this paper, we present an allocation problem and develop an iterative algorithm that optimizes both the allocation probabilities of high-priority customers in the first stage and resource allocation in the second stage. The iterative algorithm is based on two approximation models of the queueing network; thus, the results obtained in this paper are approximate in nature. This work features three main differences when compared to Almehdawe et al. (2016). First, in the present work, we assume that first-stage service time is dependent on the destination node decision, whereas it was assumed to be independent in Almehdawe et al. (2016). This makes the current model more realistic; for example, in an EMS setting, the time to reach the destination ED is dependent on its location. Second, unlike Almehdawe et al. (2016), our approximation method works significantly better when the loss probability of type a customers is not close to zero. Third, we explore a different optimization problem that is relevant to the airport staffing applications. In addition to optimizing the allocation policy, we also optimize the second-stage service rate which is an issue that was not considered in Almehdawe et al. (2016).

The rest of the paper is organised as follows. In Section 2, we introduce the queueing system of interest and describe the steps for model approximation and the optimization problems. In Sections 3 and 4, we define and analyse two queueing models for the approximation. In Section 5, two optimization problems and an iterative scheme are introduced for optimizing allocation probabilities and service rates. Section 6 presents two applications (an EMS application and an airport staffing application) with some numerical analysis. Finally, Section 7 concludes the paper.

2. The queueing network

We consider a multi-server queueing network that serves two types of customers: type a and type b. The network consists of two stages: stage one, which only serves type a customers and stage two, which serves both type a and type b customers. Stage one consists of a single multi-server node and stage two consists of K multi-server nodes. The structure of the network is illustrated in Figure 1, and the service process and flow of customers can be described as follows.

- When a type a customer arrives to the system, the customer first receives service at stage one provided a server is available; if a server is not available, the customer is lost. Upon arrival at stage one, the type a customer is assigned to one of the K nodes hereafter referred to as *destination nodes* at stage two. The service time for the customer at stage one depends on the destination node. If a server in the destination node is available, the customer moves to that node to receive service at stage two once they have received service at stage one; if all servers are busy, the customer has to wait in stage one until a server at the destination node becomes available. The period of time that a server at stage one is occupied – that is, unavailable to serve other type a customers – is called *blocking delay*. After receiving service at stage two, the type a customer leaves the system.
- Type b customers join the network in one of the K nodes at stage two. A type b customer who arrives to node k starts service if there is a server available at node k. Otherwise, the customer waits in a queue with infinite capacity.
- In order to reduce blocking in the first-stage servers, we assign type a customers higher non-pre-emptive priority over type b customers. Thus, once a server in a node becomes available, the server begins to serve a blocked type a customer who has been allocated to that node, if there is one. Customers of the same type are served on a first-come-first-served basis.

The above described queueing network can be defined explicitly as follows.

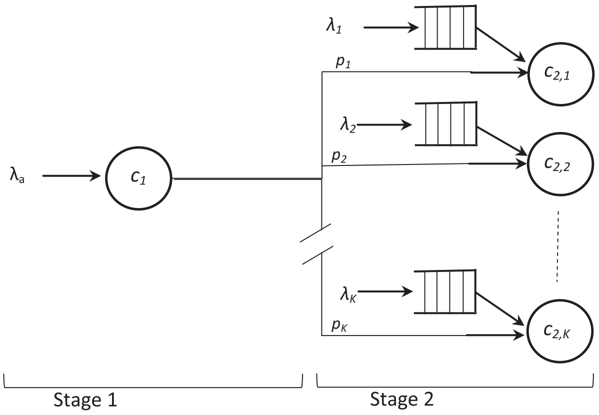

Figure 1. The queueing network of interest.

Customer arrival processes: We assume that customers arrive independently to the network according to Poisson processes with parameters λ_a and $\{\lambda_k, k = 1, 2, \dots, K\}$, for type a and type b customers, respectively.

Customer service time at stage one: Stage-one service times for type a customers are dependent on their destination nodes at stage two. This relationship assumption would be useful when stage-one service time is used to model, for example, the transportation time to Emergency Rooms (See Example 6.2). If more than one customer has the same destination node, then they will have the same service time distribution and will be served on a FCFS basis. We assume that the service time is exponentially distributed with parameter $\mu_{T,k}$, for $k = 1, 2, \dots, K$ and T stands for transportation time.

Allocation probabilities: Upon arrival, a type a customer is allocated to one of the K nodes at stage two with probabilities $\{p_k$, for $k = 1, 2, \dots, K\}$.

Customer service time at stage two: We assume the service time at stage two is exponentially distributed with parameter μ_k, for $k = 1, 2, \dots, K$, regardless of customer type.

Service capacity: We use c_1 to represent the number of servers at stage 1, and $c_{2,k}$ for the number of servers in node k at stage two, for $k = 1, 2, \dots, K$.

Service priority: We assume that type a customers possess higher non-pre-emptive priority over type b customers. That is, if a server in node k becomes available, it will first be assigned to a type a customer who is allocated to node k but is presently blocked. However, service to a type b customer will not be interrupted by the arrival of type a customers.

In this paper, we are concerned with how to most effectively allocate type a customers to the K nodes at stage two. Our objective is to find a method of allocating type a customers to each node of stage two that minimizes total blocking delays at stage one. Due to the complexity of the system (e.g., c_1 and $\{c_{2,k}, k = 1, 2, \dots, K\}$ can be large), an exact analysis of the system is prohibitive (see examples in Almehdawe et al., 2013, 2016). We take an approximation approach and develop an iterative algorithm to find an allocation of type a customers that strikes a balance between blocking delays and loss probability of type a customers. Furthermore, we develop an optimization problem with the objective of minimizing the average blocking delays experienced at stage one. We also consider whether average blocking delays can be minimized by distributing service resources among the K nodes at stage two.

To develop the iterative algorithm, we introduce an approximation scheme for the present queueing system by decomposing the original network depicted in Figure 1 into individual nodes. This leads to $K+1$-isolated nodes each with adjusted effective parameters. We name these nodes as Node 0 (which represents the stage 1 node) and Node 1 (which represents each node at stage two) and they can be described as follows:

- Node 0 considers only type a customers. In total, there are c_1 servers serving type a customers. The service time at each server depends on which stage-two node the customer will be allocated to. A server can either be busy due to serving a customer, or holding the customer if the downstream node at stage two is full (blocked). If a server is available, the customer enters the server and begins to receive service upon arrival; otherwise, the customer is lost. By considering servers not blocked, and assuming there are n of them, we approximately model Node 0 as an $M/M[K]/n/n$ queue, which will be defined and analysed explicitly in Section 3. The service time in the $M/M[K]/n/n$ queue is the service time at stage one plus the blocking delay. This model is used to estimate the effective total arrival rate of type a customers into the nodes at stage two.
- Node 1 represents individual nodes at stage two. At each node, the two types of customers are served based on a non-pre-emptive priority service discipline. Approximately, we model Node 1 as an $M[2]/M/c$ non-pre-emptive priority queue, which will be defined and analysed in Section 4. This model is used to estimate the blocking delays of stage-one servers and the distribution of the number of type a customers experiencing blocking delay (which is also the number of blocked servers at stage one).

Node 0 (i.e., stage one) and node 1 (i.e., stage two) are not independent. In fact, both Node 0 and Node 1 cover the blocking delay part (i.e., the number of occupied servers in the $M/M[K]/n/n$ queue and the number of waiting type a customers in the $M[2]/M/c$ queue). On the other hand, the two nodes are analysed independently. Thus, our analysis in Section 5 gives approximate results. In Section 6, we use two applications related to EMS planning and airline staffing to demonstrate how the queueing networks developed in this paper can be applied in real-world settings.

3. Node 0 and the $M/M[K]/n/n$ queue

According to the description of Node 0 given in Section 2, the $M/M[K]/n/n$ queue is defined as follows.

- There are in total n identical servers.
- Type a customers arrive according to a Poisson process with parameter λ_a.
- An arriving customer will receive type k service (i.e., allocated to the kth node in the queueing network) with probability p_k, for $k = 1, 2, \ldots, K$.
- The service time of a customer who receives type k service has an exponential distribution with parameter $\mu_{T,k}$. (We note that $\mu_{T,k}$ will be modified to $\mu_{T,k} =: 1/(E[W_{a,k}] + 1/\mu_{T,k})$ to include the effect of blocking delays in the iterative algorithm to be introduced in Section 5.)
- If all n servers are busy, an arriving type a customer is lost.

Let $q_{a,k}(t)$ be the number of type a customers who are receiving type k service at time t. Then $\{(q_{a,1}(t), \ldots, q_{a,K}(t)), t \geq 0\}$ is a continuous-time Markov chain (CTMC) with state space $\{(n_1, \ldots, n_K) : n_1 + \cdots + n_K \leq n, n_k \geq 0, k = 1, 2, \ldots, K\}$. If $K = 1$, the queueing model is reduced to the classical Erlang loss model $M/M/n/n$. Similar to the $M/M/n/n$ queue, we show that $\{(q_{a,1}(t), \ldots, q_{a,K}(t)), t \geq 0\}$ is time-reversible and, consequently, we find its limiting probabilities $\{\pi(n_1, \ldots, n_K), n_1 + \cdots + n_K \leq n, n_k \geq 0, k = 1, 2, \ldots, K\}$.

Theorem 1: *The CTMC $\{(q_{a,1}(t), \ldots, q_{a,K}(t)), t \geq 0\}$ is time-reversible and its limiting probabilities are given by:*

$$\pi(n_1, \ldots, n_K) = \pi(0, \ldots, 0) \prod_{k=1}^{K} \frac{\rho_k^{n_k}}{n_k!}, \quad \text{for } n_1 + \cdots + n_K \leq n, \ n_k \geq 0, \ k = 1, 2, \ldots, K, \tag{1}$$

where $\rho_k = \lambda_a p_k / \mu_{T,k}$, for $k = 1, 2, \ldots, K$, and

$$\pi(0, \ldots, 0) = \left(\sum_{(n_1,\ldots,n_K):\, n_1+\cdots+n_K \leq n,\, n_k \geq 0,\, k=1,2,\ldots,K} \prod_{k=1}^{K} \frac{\rho_k^{n_k}}{n_k!} \right)^{-1}. \tag{2}$$

The proof of Theorem 1 is shown in Appendix 1.

Note: The theorem holds if the K services of customers arrive according to K independent Poisson processes.

A type a customer is lost if there is no server at stage one available at his arrival epoch. Let $\pi_{loss,n}$ be the loss probability of an arbitrary arriving type a customer.

Corollary 1: *The type a customer's loss probability $\pi_{loss,n}$ is given by:*

$$\pi_{loss,n} = \pi(0, \ldots, 0) \times \left(\sum_{(n_1,\ldots,n_K):\, n_1+\cdots+n_K = n,\, n_k \geq 0,\, k=1,2,\ldots,K} \prod_{k=1}^{K} \frac{\rho_k^{n_k}}{n_k!} \right). \tag{3}$$

The departure rate of customers at stage one who are allocated to the kth node is $p_k \lambda_a (1 - \pi_{loss,n})$, for $k = 1, 2, \ldots, K$.

The proof of Corollary 1 is shown in Appendix 1.

Computation of $\pi_{loss,n}$ can be done efficiently as follows. Since there are c_1 servers at stage one in the original queueing network, the number of servers n is less than or equal to c_1. Thus, we compute $\pi_{loss,n}$ for $n = 0, 1, \ldots, c_1$. Let $x(k, n) = 1/\pi(0, \ldots, 0)$ and $y(k, n) = \pi_{loss,n}/\pi(0, \ldots, 0)$ for the $M/M[k]/n/n$ queue. Then $x(k, n)$ and $y(k, n)$ can be computed as:

$$\begin{aligned}
&x(1,0) = 1, \ x(1,n) = x(1,n-1) + \frac{\rho_1^n}{n!}, \\
&\quad \text{for } n = 2, 3, \ldots, c_1; \\
&x(k,n) = \sum_{i=0}^{n} x(k-1, n-i) \frac{\rho_k^i}{i!}, \\
&\quad \text{for } 2 \leq k \leq K, \ 0 \leq n \leq c_1; \\
&y(1,0) = 1, \ y(1,n) = \frac{\rho_1^n}{n!}, \\
&\quad \text{for } n = 2, 3, \ldots, c_1; \\
&y(k,n) = \sum_{i=0}^{n} y(k-1, n-i) \frac{\rho_k^i}{i!}, \\
&\quad \text{for } 2 \leq k \leq K, \ 0 \leq n \leq c_1.
\end{aligned} \tag{4}$$

Then, we obtain the following result.

Corollary 2: *The loss probability $\pi_{loss,n}$ can be obtained as $\pi_{loss,n} = y(K, n)/x(K, n)$, for $n = 0, 1, \ldots, c_1$.*

4. Node 1 and the $M[2]/M/c$ non-pre-emptive priority queue

Node 1 represents individual second-stage nodes, which was used in Almehdawe et al. (2016). For notational convenience, we remove the subscript k from each variable and use more generic notation for the $M[2]/M/c$ queue, which is defined explicitly as follows.

- Type a customers arrive according to a Poisson process with parameter $p\lambda_a(1 - \pi_{loss})$, where p is for the allocation probability p_k, and π_{loss} is the total loss probability of type a customers.
- Type b customers arrive according to a Poisson process with parameter λ, which is independent of the type a customers' arrival process.

- The service time of a customer, regardless of its type (a or b), has an exponential distribution with parameter μ.
- There are c identical servers.
- Type a customers have non-pre-emptive service priority over type b customers.

We want to find the waiting times of the two types of customers, especially, the waiting time of type a customers. Denote by W_a the waiting time in the queue of an arbitrary type a customer (i.e., a high-priority customer) and W_b the waiting time in the queue of an arbitrary type b customer (i.e., a low- priority customer). Let $\sigma = p\lambda_a(1-\pi_{loss})/(c\mu)$ and $\rho = (p\lambda_a(1-\pi_{loss})+\lambda)/(c\mu) = \sigma + \lambda/(c\mu)$, where σ represents the node utilisation by the high-priority customers and ρ represents the node overall utilisation. If $\rho < 1$, it is well known that (see Gross, Shortle, Thompson, & Harris, 2008):

$$E[W_a] = \frac{1}{1-\sigma}\left(c!(1-\rho)c\mu\sum_{n=0}^{c-1}\frac{(c\rho)^{n-c}}{n!}+c\mu\right)^{-1};$$
$$E[W_b] = \frac{1}{(1-\rho)}E[W_a]. \tag{5}$$

The mean waiting time of type a customers $E[W_a]$ is used to approximate the average blocking delay. The function $E[W_a]$ will be used in an optimization problem in Section 5, where the following properties play an important role.

Theorem 2 (Lemma 1 in Almehdawe et al. (2016)): *The function* $E[W_a]$ *is increasing convex in* p.

Theorem 3: *Function* $E[W_a]$ *is a decreasing convex function in* μ.

The proof of Theorem 3 is shown in Appendix 1.

The distribution of the number of blocked type a customers can be found explicitly. Although distributions of the queue lengths for the $M[2]/M/c$ non-pre-emptive priority queue have been given in a number of papers (e.g., Gail, Hantler, & Taylor, 1988), simple and explicit results cannot be found. Next, we present an explicit result. Let

- $q_1(t)$ be the number of all customers in service plus all waiting type a customers and
- $q_2(t)$ the number of waiting type b customers.

Then $\{(q_1(t), q_2(t)),\ t \geq 0\}$ is a continuous-time Markov chain (CTMC). If $q_1(t) \leq c-1$, we must have $q_2(t) = 0$. Then we denote the state space of the Markov chain as $\{0, 1, \ldots, c-1\} \cup \{(i,j), c \leq i \leq c+n, j \geq 0\}$. Let $\{\eta_0, \eta_1, \ldots, \eta_{c-1}, \eta_{(i,j)}, c \leq i \leq c+n, j \geq 0\}$ be the limiting probabilities of the Markov chain. It is easy to see that the limiting probabilities exist if and only if $p\lambda_a(1-\pi_{loss}) + \lambda < c\mu$. To find the limiting probabilities, we first establish a set of balance equations for them:

$$\begin{aligned}
&(p\lambda_a(1-\pi_{loss})+\lambda)\eta_0 = \mu\eta_1;\\
&(p\lambda_a(1-\pi_{loss})+\lambda+i\mu)\eta_i\\
&\quad = (p\lambda_a(1-\pi_{loss})+\lambda)\eta_{i-1}\\
&\qquad + (i+1)\mu\eta_{i+1}, \quad 1 \leq i \leq c-2;\\
&(p\lambda_a(1-\pi_{loss})+\lambda+(c-1)\mu)\eta_{c-1}\\
&\quad = (p\lambda_a(1-\pi_{loss})+\lambda)\eta_{c-2} + c\mu\eta_{(c,0)};\\
&(p\lambda_a(1-\pi_{loss})+\lambda+c\mu)\eta_{(c,0)}\\
&\quad = (p\lambda_a(1-\pi_{loss})+\lambda)\eta_{c-1} + c\mu(\eta_{(c+1,0)}+\eta_{(c,1)});\\
&(p\lambda_a(1-\pi_{loss})+\lambda+c\mu)\eta_{(c,j)}\\
&\quad = \lambda\eta_{(c,j-1)} + c\mu(\eta_{(c+1,j)}+\eta_{(c,j+1)}), \quad j \geq c;\\
&(p\lambda_a(1-\pi_{loss})+\lambda+c\mu)\eta_{(i,0)}\\
&\quad = p\lambda_a(1-\pi_{loss})\eta_{(i-1,0)}\\
&\qquad + c\mu\eta_{(c+1,0)}, \quad c+1 \leq i \leq c+n;\\
&(p\lambda_a(1-\pi_{loss})+\lambda+c\mu)\eta_{(i,j)}\\
&\quad = p\lambda_a(1-\pi_{loss})\eta_{(i-1,j)} + \lambda\eta_{(i,j-1)}\\
&\qquad + c\mu\eta_{(c+1,j)}, \quad c+1 \leq i \leq c+n,\ j \geq 1.
\end{aligned} \tag{6}$$

Using the equations for $i = 0, 1, \ldots, c-1$, we obtain

$$\begin{aligned}
\eta_i &= \eta_0\frac{(c\rho)^i}{i!}, \quad \text{for } 0 \leq i \leq c-1;\\
\eta_{(c,0)} &= \eta_0\frac{(c\rho)^c}{c!}.
\end{aligned} \tag{7}$$

Let $\eta_i = \sum_{j=0}^{\infty}\eta_{(i,j)}$, for $i \geq c$. Using the equations in (6), we obtain

$$\begin{aligned}
&(p\lambda_a(1-\pi_{loss})+\lambda)\eta_c = (p\lambda_a(1-\pi_{loss})+\lambda)\eta_{c-1}\\
&\qquad + c\mu\eta_{c+1} - c\mu\eta_{(c,0)};\\
&(p\lambda_a(1-\pi_{loss})+c\mu)\eta_i = p\lambda_a(1-\pi_{loss})\eta_{i-1}\\
&\qquad + c\mu\eta_{i+1}, \quad \text{for } c+1 \leq i \leq c+n.
\end{aligned} \tag{8}$$

Using the relationship between η_{c-1} and $\eta_{(c,0)}$, the above equations lead to

$$\eta_i = \eta_c\sigma^{i-c}, \quad \text{for } i \geq c. \tag{9}$$

Using Equation (1) in Gail et al. (1988), which is established by computing the mean number of working servers at an arbitrary time, and the law of total probability, the following linear system for η_0 and η_c can be established:

$$\begin{aligned}
&\left(\sum_{i=0}^{c-1}\frac{i}{i!}(c\rho)^i\right)\eta_0 + c(1-\sigma)^{-1}\eta_c = c\rho;\\
&\left(\sum_{i=0}^{c-1}\frac{1}{i!}(c\rho)^i\right)\eta_0 + (1-\sigma)^{-1}\eta_c = 1.
\end{aligned} \tag{10}$$

Solving the linear system, we obtain

$$\eta_0 = \frac{c(1-\rho)}{\sum_{j=0}^{c-1} \frac{(c-j)}{j!}(c\rho)^j} \quad \text{and} \quad \eta_c = \frac{(1-\sigma)\frac{1}{(c-1)!}(c\rho)^c}{\sum_{j=0}^{c-1} \frac{(c-j)}{j!}(c\rho)^j}. \tag{11}$$

Since the number of waiting type a customers (blocked) is given by $q_l = \max\{0,\ q_1(t) - c\}$, its distribution can be obtained as follows.

Corollary 3:

$$P\{q_l = 0\} = \sum_{i=1}^{c} \eta_i = 1 - \frac{\frac{\sigma}{(c-1)!}(c\rho)^c}{\sum_{j=0}^{c-1} \frac{(c-j)}{j!}(c\rho)^j};$$
$$P\{q_l = i\} = \frac{(1-\sigma)\frac{1}{(c-1)!}(c\rho)^c\sigma^i}{\sum_{j=0}^{c-1} \frac{(c-j)}{j!}(c\rho)^j}, \quad \textit{for } i = 1, 2, \ldots. \tag{12}$$

The distribution of q_l is used to estimate the number of stage- one servers that are blocked.

5. Two optimization problems

It is well known that server blocking significantly deteriorates the performance of service systems. Instead of utilising the upstream node limited capacity for serving the high-priority type a customers, some of the servers at stage one are used as waiting spots for those customers who finished service but cannot join the destination node, since all servers in that node are in service. In queueing networks, this type of blocking is referred to as Blocking After Service (BAS), see Perros (1994), for more details on BAS and other types of blocking. We use blocking delays as a measure for the network performance which we try to minimize by either allocating the type a customers to stage-two nodes properly, or by deciding on the service speed at the second-stage nodes. In our model, "properly" means that the optimization problem output will result in smaller average blocking delays for the system in the long run.

For both optimization problems considered, our objective is to find the decision variables that will result in minimum average blocking delays. To obtain the average blocking delays of an arbitrary type a customer who received service at stage one, all K nodes have to be considered simultaneously. We assume that the actual total arrival rate of type a customers to the K nodes is $\lambda_a(1 - \pi_{loss})$, where π_{loss} is the loss probability of type a customers, which can be estimated by using Equation (3). Then the actual arrival rate of type a customers to the kth node is $p_k\lambda_a(1 - \pi_{loss})$. For the kth node, recall that (i) the number of servers is $c_{2,k}$; (ii) the service rate is μ_k; and (iii) the arrival rate of type b customers is λ_k. In Equation (5), we add subscript k to quantities W_a, μ, ρ, c and p. Then an expression for the mean waiting time of type a customers (i.e., the average blocking delays) in the kth node can be obtained from Equation (5). Consequently, the average blocking delays of stage-one servers can be obtained as the weighted average: $\sum_{k=1}^{K} p_k E[W_{a,k}]$.

5.1. Optimization problem 1: Load allocation optimization

In this subsection, we focus on developing an optimization problem in which we find proper allocation probabilities to allocate the high-priority customers from stage 1 to stage 2. To minimize the average blocking delays, we first make sure that type a customers blocking delays in front of individual stage-two nodes are finite. This can be achieved only if (i) individual stage-two nodes have enough capacity to serve all type a customers arrived to them; and (ii) the system have enough capacity to serve all type a customers who are not lost. Part (i) is equivalent to assume that $\sigma_k < 1$, where $\sigma_k = (p_k\lambda_a(1 - \pi_{loss}))/(c_{2,k}\mu_k)$, for $k = 1, 2, \ldots, K$. This condition on $\{\sigma_k,\ k = 1, 2, \ldots, K\}$ restricts the choices of allocation probabilities. In fact, *feasible allocation probabilities* (i.e., $p_k \geq 0$, for $k = 1, 2, \ldots, K$, and $\sum_{k=1}^{K} p_k = 1$) exist to ensure $\sigma_k < 1$ for $k = 1, 2, \ldots, K$, if and only if

$$1 \leq \frac{\sum_{k=1}^{K} c_{2,k}\mu_k}{\lambda_a(1 - \pi_{loss})} = \sum_{k=1}^{K} p_k^{(\max)}, \tag{13}$$

where $p_k^{(\max)} = c_{2,k}\mu_k/(\lambda_a(1 - \pi_{loss}))$. If the conditions hold, then the set of feasible allocation probabilities $\{(p_1, \ldots, p_K) : 0 \leq p_k \leq p_k^{(\max)},\ k = 1, \ldots, K,$ $\sum_{k=1}^{K} p_k = 1\}$ is not empty. To ensure condition (13), we must have $\pi_{loss} \geq \pi_{loss}^{(\min)}$, where

$$\pi_{loss}^{(\min)} = \max\left\{0, 1 - \frac{\sum_{k=1}^{K} c_{2,k}\mu_k}{\lambda_a}\right\}. \tag{14}$$

We are ready to propose the following optimization problem to find allocation probabilities $\{p_1, \ldots, p_K\}$ that minimize the average blocking delays:

$$\min_{p_1,\ldots,p_k} \sum_{k=1}^{K} p_k E[W_{a,k}] = \sum_{k=1}^{K} \frac{p_k}{1 - \sigma_k}$$
$$\times \left(c_{2,k}!(1-\rho_k)c_{2,k}\mu_k \sum_{n=0}^{c_{2,k}-1} \frac{(c_{2,k}\rho_k)^{n-c_{2,k}}}{n!} + c_{2,k}\mu_k \right)^{-1} \tag{15}$$

$$s.t. \quad \sum_{k=1}^{K} p_k = 1; \tag{16}$$

$$0 \leq p_k \leq p_k^{(\max)}, \quad \text{for } k = 1, 2, \ldots, K. \tag{17}$$

where $\rho_k = \sigma_k + \lambda_k/(c_{2,k}\mu_k)$, for $k = 1, 2, \ldots, K$. We note that, if we are concerned about the waiting time of type b customers (to be finite), then we must modify $p_k^{(\max)}$ as $p_k^{(\max)} = (c_{2,k}\mu_k - \lambda_k)/(\lambda_a(1 - \pi_{loss}))$, for $k = 1, 2, \ldots, K$. Note that optimization problem (15)–(17) was introduced in Almehdawe et al. (2016).

By Theorem 2, the objective function (15) is convex in $\{\sigma_1, \ldots, \sigma_K\}$ and, consequently, convex in $\{p_1, \ldots, p_K\}$. All constraints (16)–(17) are linear. Thus, the optimization problem is a convex program, which can be solved effectively by existing methods. For the numerical examples presented in Section 6, we solve the above optimization problem using the *fmincon* solver in Matlab where the *interior-point* algorithm is used.

Since π_{loss} depends on the optimal allocation probabilities, we propose an iterative scheme to find a *stationary solution* for the system. The iterative scheme consists of two stages: (i) estimation of the total effective arrival rates of type a customers to the stage-two individual nodes; and (ii) optimization of allocation probabilities and the distribution of the total number of servers blocked at stage one.

(1) Assume that the allocation probabilities are $\{p_1, p_2, \ldots, p_K\}$ and the distribution of the total number of blocked type a customers B is $\boldsymbol{\xi} = (\xi(0), \xi(1), \xi(2), \ldots, \xi(c_1))$. Since there are only c_1 servers at stage one, we can have at most c_1 blocked type a customers. Thus, we must have $B \le c_1$. We use the $M/M[K]/n/n$ queue to find the loss probability of type a customers and the effective type a customer arrival rates to individual stage-two nodes. Specifically, given that $B = b$, we use the $M/M[K]/(c_1 - b)/(c_1 - b)$ queue to find loss probability $\pi_{loss, c_1 - b}$ (see Equation (3)), for $b = 0, 1, 2, \ldots, c_1$. Conditioning on B, the loss probability of type a customers can be found as:

$$\pi_{loss} = \sum_{b=0}^{c_1} \xi(b) \pi_{loss, c_1 - b}. \tag{18}$$

The effective arrival rates of type a customers to individual nodes can be found as $p_k \lambda_a (1 - \pi_{loss})$, for $k = 1, 2, \ldots, K$.

(2) In the $M[2]/M/c$ non-pre-emptive priority queue, π_{loss} affects the determination of the mean blocking delays. Given π_{loss} and other original network parameters, we solve the optimization problem (15–17) to obtain solutions $\{p_k^*,\ k = 1, \ldots, K\}$ and $\{E[W_{a,k}^*],\ k = 1, \ldots, K\}$. We also find the distributions of the servers blocked $q_{b,k}$, which is given by (12), for individual stage-two nodes. Then the total number of blocked stage-one servers can be found as $B = \min\{c_1,\ q_{b,1} + q_{b,2} + \cdots + q_{b,K}\}$.

A solution $\{\pi_{loss},\ p_k,\ k = 1, 2, \ldots, K\}$ is called a *stationary solution* if (i) Given π_{loss}, the solution of optimization problem (15)–(17) is $\{p_k,\ k = 1, 2, \ldots, K\}$; and (ii) Given $\{p_k,\ k = 1, 2, \ldots, K\}$, the distribution of B, and $\{E[W_{a,k}],\ k = 1, 2, \ldots, K\}$, then (18) gives π_{loss}. A solution $\{\pi_{loss},\ p_k,\ k = 1, 2, \ldots, K\}$ is called a *consistent solution* if $\{p_k,\ k = 1, 2, \ldots, K\}$ is applied in the original queueing system, the loss probability of type a customers is π_{loss}.

Intuitively, a stationary solution gives consideration to both customer blocking delays and customer loss. For a (non-stationary) solution, if π_{loss} is big, then the average blocking delays can be small since there are less type a customers to stage- two nodes. On the other hand, if the average blocking delays are small, then π_{loss} is small since there are more type a customers arriving to the network. This observation indicates that, it is possible to find a stationary solution by using the above two steps iteratively. Thus, to find a stationary solution, we introduce an iterative scheme involving the above two steps.

An Iterative Scheme for Computing a Stationary Solution Summarising the above discussion, the iterative scheme can be stated as follows.

(0) Input system parameters: λ_a, $\{\mu_{T,k},\ c_{2,k}, \mu_k, \lambda_k,\ k = 1, 2, \ldots, K\}$. Choose $\epsilon > 0$. Set $\pi_{loss}(1) = \pi_{loss}^{(\min)}$ and $n = 1$.
(1) Given $\pi_{loss}(n)$, compute $\{p_k^{(\max)}(n),\ k = 1, 2, \ldots, K\}$. Solve the optimization problem (15–17) to find $(p_1^*(n),\ \ldots,\ p_K^*(n))$.
(2) Use $(p_1^*(n),\ \ldots, p_K^*(n))$ and Equation (5) to calculate $\{E[W_{a,k}^*](n),\ \text{for}\ k = 1, 2, \ldots, K\}$. Use Corollary 3 to find the distributions of $q_{b,k}$, for $k = 1, 2, \ldots, K$, and the distribution of $B = \min\{c_1,\ q_{b,1} + \cdots + q_{b,K}\}$.
(3) Use $\{p_k^*(n),\ E[W_{a,k}^*](n),\ k = 1, 2, \ldots, K\}$, the distribution of B, the $M/M[K]/c_1/c_1$ queue, and Equation (3) to calculate $\pi_{loss}(n+1)$. Reset $\pi_{loss}(n+1) =: \max\{\pi_{loss}(n+1),\ \pi_{loss}^{(\min)}\}$. Note that the rate of type k service is $1/(1/\mu_{T,k} + E[W_{a,k}^*](n))$.
(4) If $|\pi_{loss}(n+1) - \pi_{loss}(n)| < \epsilon$, stop; otherwise, set $n =: n + 1$ and go to step 1.

5.2. Optimization problem 2: Capacity allocation optimization

Another optimization problem considered in this work is to find the optimal service speed for each server in stage two. Our objective is to minimize the average blocking delays experienced by the high -priority customers. We propose the following optimization problem to find $\{\mu_1, \mu_2, \ldots, \mu_K\}$:

$$\min_{\mu_1,\ldots,\mu_k} \sum_{k=1}^{K} p_k E[W_{a,k}] = \sum_{k=1}^{K} \frac{p_k}{1 - \sigma_k} \times \left(c_{2,k}!(1 - \rho_k) c_{2,k} \mu_k \sum_{n=0}^{c_{2,k}-1} \frac{(c_{2,k}\rho_k)^{n - c_{2,k}}}{n!} + c_{2,k}\mu_k \right)^{-1} \tag{19}$$

$$s.t. \qquad \sum_{k=1}^{K} c_k \mu_k = \mu_{total} \tag{20}$$

Table 1. Parameters for Example 6.1.

k	$\mu_{T,k}$	$c_{2,k}$	λ_k	μ_k
1	1/2	30	3.6	1/6
2	1/2	32	4.5	1/5
3	1/3	34	4.3	1/5

where μ_{total} corresponds to the total service capacity available at stage two. By Theorem 3, the objective function in (19) is convex in $\{\mu_1, \mu_2, \ldots, \mu_K\}$. The constraint (20) is linear. Thus, similar to optimization problem 1, this optimization problem is a convex program, which can be solved effectively by existing methods. To find a stationary solution, we use the iterative algorithm of Section 5.1 by adjusting for the new decision variables. In the next section, we illustrate how the above two optimization problems can be utilized in two different applications and show some numerical examples.

6. Model applications and numerical examples

In this section, We utilize our queueing network to explore two applications: a health care application and an airline staffing application. We validate the approximation scheme through simulation and then apply it to a real case. For all the simulations in the following examples, we use a discrete event simulation package (Simul8) to estimate performance measures. The numbers in parentheses are the 95% confidence interval lengths that were generated by 100 runs each with a simulation time of 100,000 h and a warm-up period of 200 h.

Example 6.1: To demonstrate the applicability of the queueing network model developed in this paper, we consider an application in health care systems. More specifically, we focus on the dispatching decisions of EMS ambulances to regional EDs. In some countries, such as Canada, it is a common practice to have an ambulance wait outside an ED if it is full and cannot accept care of the patient. This type of ambulance delay is referred to as an *offload delay*. It is well known that ambulance offload delays affect EMS performance significantly. Ambulance offload delays can be reduced by adjusting the allocation of ambulance patients to regional EDs based on the prospective congestion and capacity of a given ED. See Almehdawe et al. (2016), Laan, Vanberkel, Boucherie, and Carter (2016) and Carter et al. (2015) for more details on this problem.

In the current queueing model, the first-stage servers represent the EMS ambulances and the second-stage nodes represent regional EDs. So, for a region that is served by three hospitals, $K = 3$. The high- priority customers are the patients arriving by an ambulance who require service in both, the ambulance and later in the ED. Conversely, the low-priority customers are the patients who arrive at the ED by themselves. We refer to this group of customers as walk-in patients. In this context, the objective is to select allocation probabilities (the allocation policy) that minimize the average offload delays experienced by ambulance patients.

Analysis of a real EMS case. For this case, we consider a network that consists of a single EMS provider and three hospitals (i.e., $K = 3$). We set $\lambda_a = 4.5$ and $c_1 = 15$. Table 1 provides the detailed data for the rest of the EMS-ED network parameters. First, we solve for the optimal allocation policy using the iterative algorithm developed earlier in Section 5. We find that the optimal policy for this example is $(22.60\%, 33.73\%, 43.67\%)$. The reported results in Table 2 achieve two objectives: first, to compare the approximation via decomposition approach performance with simulation. Second, to check the efficiency of our optimal solution by comparing the performance of the EMS-ED network under our optimal allocation policy to its performance under a *common allocation policy*. The common allocation policy is based on the contribution of each ED capacity to the total regional capacity. Namely, we set $p_k = \frac{c_{2,k}\mu_k}{\sum_{k=1}^{K} c_{2,k}\mu_k}$, which is $(p_1, p_2, p_3) = (27.47\%, 35.16\%, 37.36\%)$ for this example. The performance measures under the two allocation policies are recorded in Table 2.

Based on Table 2, we are able to make the following observations:

- The results for the approximation via decomposition for the high-priority ambulance patients ($E[\hat{W}_{a,k}]$) are within 10% to the simulation results ($E[W_{a,k}]$). While the walk-in patients results from the approximation scheme ($E[\hat{W}_k]$) are within 17% of the simulation results ($E[W_k]$).
- The common allocation policy only considers the capacity contribution of an ED with respect to the region's total capacity. However, our allocation policy takes into consideration the arrival rates of low-priority patients and the other sources of ED congestion in addition to each ED capacity. For that reason, we notice that total expected offload (blocking) delays are lower using our methodology (0.1058 compared to 0.1141) corresponding to a 7.28% decrease.
- Based on the iterative algorithm allocation policy, the three EDs are loaded such that the ED utilisation is balanced (between 90.12% and 92.41%). However, this is not true for the common allocation policy where the ED utilisation is between 86.41% and 95.13%.
- The low-priority walk-in patients who choose to go to ED1 are expected to experience long waiting time (3.71 h on average) under the common allocation policy. However, under our stationary allocation pol-

Table 2. Simulation and optimization results for Example 6.1.

	Stationary allocation policy			Common allocation policy		
k	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
p_k	22.60%	33.73%	43.67%	27.47%	35.16%	37.36%
$E[W_{a,k}]$	0.1239	0.1101	0.0853	0.1822	0.1234	0.0565
	$(\pm 0.0024)^*$	(± 0.0013)	(± 0.0010)	(± 0.0031)	(± 0.0017)	(± 0.0010)
$E[\hat{W}_{a,k}]$	0.1290	0.1163	0.0944	–	–	–
$E[W_k]$	1.3740	1.4180	0.8472	3.7855	1.9264	0.4028
	(± 0.0636)	(± 0.0500)	(± 0.0277)	(± 0.2499)	(± 0.0937)	(± 0.0147)
$E[\hat{W}_k]$	1.4664	1.5919	0.9930	–	–	–
ρ_k	90.90%	92.41%	90.12%	95.13%	93.85%	86.41%
$\sum p_k E[W_{a,k}]$		0.1024			0.1145	

* For simulation results, the half-length of the 95% confidence interval is presented.

icy, their expected delays are cut by more than 50% (1.36 h). Although the expected waiting time of walk-in patients in ED3 increases, overall, the expected waiting time of walk-in patients decreases significantly overall.

- At the regional level, an arbitrary walk-in patient can expect delays of 1.83 h under the common allocation policy, whereas they would only experience a delay of 1.16 h under the stationary allocation policy. Even though the main objective of this model is to minimize waiting times for high-priority patients, we observed that the mean waiting time for walk-in patients decreased substantially when the optimal allocation policy was used.

Example 6.2: The second application of the queueing model considered in this paper relates to staffing decisions in an airport. In this setting, there are two types of passengers: regular passengers and special needs passengers. Regular passengers are able to get to their gates by themselves (type b customers). Special needs passengers (type a customers) are helped to their gates by special staff. At each gate, several staff members are available to serve the passengers. Special staff can only leave the gate only if the gate personnel can serve the special needs passenger (e.g., help them get seated). The issue of interest here is not to optimize $\{p_1, p_2, \ldots, p_K\}$; rather, it is to optimize $\{\mu_k, k = 1, 2, \ldots, K\}$ for given $\{p_1, p_2, \ldots, p_K\}$.

Analysis of an airport staffing case. For this case, we consider a terminal that consists of five gates ($K = 5$). We set $\lambda_a = 2.5$ and $c_1 = 15$. Table 3 provides detailed information for the rest of the network input parameters. We set $\mu_{total} = 20$ and the routing probabilities to $\{0.09, 0.36, 0.18, 0.27, 0.09\}$. Table 4 summarises the results for the optimal service rate policy based on the optimization problem in (19) and (20). We also report the expected waiting time for high and low-priority customers for the stationary policy. Based on Table 4, we make the following observations:

- Comparing the decomposition approach results for waiting time with the simulation results show the validity of our approach to mimic the original network.

Table 3. Parameters and results for Example 6.2.

k	$\mu_{T,k}$	$c_{2,k}$	λ_k	p_k
1	1/2	5	2.6	1/11
2	1/2	4	2.5	4/11
3	1/3	3	2.0	2/11
4	1/3	5	3.3	3/11
5	1/3	4	2.3	1/11

- Gate 1 and Gate 4 have the same number of servers. However, we notice that, based on our optimization model, the suggested service speed at Gate 4 should be about double that of Gate 1 ($\mu_4 = 1.0193$ compared to $\mu_1 = 0.6457$). This is because the low-priority customer arrival rate at Gate 4 is higher ($\lambda_4 = 3.3$ compared to $\lambda_1 = 2.6$). Additionally, the high-priority customer arrival rate at the same gate is higher ($3/11 * \lambda_a(1 - \pi_{loss})$ compared to $1/11 * \lambda_a(1 - \pi_{loss})$).
- Gate 2 and Gate 5 have the same number of servers, but we notice that Gate 2 suggested service speed ($\mu_2 = 1.2708$) should be much higher than Gate 5 ($\mu_5 = 0.7537$). This is due to the fact that the arrival rate of high-priority customers at Gate 2 is triple that of Gate 5.
- If we compare the resulting utilisations of the optimal service speed policy for each gate, we notice a considerable variation in congestion between the gates. Furthermore, we notice that each gate's utilisation is inversely related to the total load on the corresponding gate (from both high- and low-priority customers). We can explain this as follows: when the load is high at a given gate, the servers at that gate should work much faster than those at other gates to decrease the total blocking delays. One managerial insight from this could be to always assign more workers to gates with higher customer flows. Even though some less-congested gates might experience longer delays, the total average delays could be lower using this rule of thumb. This observation actually motivates our future research, which is touched on in the following section.

Table 4. Simulation and optimization results for Example 6.2.

k	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
μ_k	0.6457	1.2708	1.1921	1.0193	0.7537
$E[W_{a,k}]$	0.2338	0.0918	0.1482	0.1173	0.2342
	(± 0.0102)	(± 0.0041)	(± 0.0051)	(± 0.0040)	(± 0.0046)
$E[\hat{W}_{a,k}]$	0.2361	0.0919	0.1512	0.1173	0.2391
$E[W_k]$	1.7895	0.2835	0.4877	0.5296	1.4320
	(± 0.1836)	(± 0.0214)	(± 0.0268)	(± 0.0217)	(± 0.1082)
$E[\hat{W}_k]$	1.8967	0.2787	0.4816	0.5357	1.4771
ρ_k	87.56%	67.03%	68.61%	78.10%	83.81%

7. Conclusion

This paper investigated a load allocation problem for a two-stage queueing network model that serves two types of customers. The proposed solution is based on the development of approximations for the network and using the explicit but approximate results for Nodes 0 and 1. In the end, an iterative algorithm was developed for computing a stationary allocation policy that strikes a balance between the average blocking delays and the loss probability of the high-priority (type a) customers. Through numerical experiments, we showed that the stationary policy improved the performance of the system in terms of both average blocking delays and loss probability of type a customers.

One major application for the queueing network analysed in this paper relates to EMS dispatching decisions. The optimal allocation policy derived from this model can be used by EMS dispatchers as a guideline in establishing long-term targets. Furthermore, it can also be used in other application areas such as call centres, hospital bed management, airport staffing management.

To find the optimal allocation policy, this work focused only on blocking delays related to high-priority customers. However, from a practical perspective, there is usually a service level constraint for low-priority customers to make sure that they are not severely delayed because of the optimal allocation policy. To that end, we would like to conduct future research that explores how to include a performance metric for low-priority customers into the model. In addition, we are also interested in investigating how the inclusion of this performance metric would affect the optimal allocation policy.

A key step in the iterative solution approach proposed in this paper is to approximate the loss probability of ambulance patients in Node 0. In the current model, an $M/M[K]/n/n$ queue is utilized, in which the mean service time is modelled as the sum of the mean transportation time and the mean waiting time in Node 1. Since the Laplace–Stieltjes transform (LST) of the waiting time of ambulance patients in Node 1 can be found (see Kella & Yechiali, 1985), the service time distributions of customers in Node 0 can be found. Consequently, Node 0 can be modelled as an $M/G[K]/n/n$ queue. Apparently, this approach may yield a better approximation to the loss probability, which is yet to be found. Alternatively, we can use the LST of the waiting time to modify the service rates in the (currently used) $M/M[K]/n/n$ queue so as to improve the approximation to the loss probability. Thus, an interesting future research problem is to find approximations obtained by those methods, and to compare the quality of those approximations under different congestion rates.

Another broader area of future research arising from this work is related to the concept of reserved capacity. Instead of minimizing high-priority customers' delays, we are interested in investigating what effect reserved capacity at the second stage has on the whole network performance, and in finding the optimal reserved capacity decisions at each node required to achieve target performance measures.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Akyildiz, I. F. (1989). Product form approximations for queueing networks with multiple servers and blocking. *IEEE Transactions on Computers, 38*(1), 99–114.

Almehdawe, E., Jewkes, B., & He, Q.-M. (2013). A markovian queueing model for ambulance offload delays. *European Journal of Operational Research, 226*(3), 602–614.

Almehdawe, E., Jewkes, B., & He, Q.-M. (2016). Analysis and optimization of an ambulance offload delay and allocation problem. *Omega, 65*, 148–158.

Andriansyah, R., Van Woensel, T., Cruz, F. R., & Duczmal, L. (2010). Performance optimization of open zero-buffer multi-server queueing networks. *Computers & Operations Research, 37*(8), 1472–1487.

Baker, J. R., Clayton, E. R., & Taylor, B. W., III. (1989). A non-linear multi-criteria programming approach for determining county emergency medical service ambulance allocations. *Journal of the Operational Research Society, 40*(5), 423–432.

Benjaafar, S., & Gupta, D. (1999). Workload allocation in multi-product, multi-facility production systems with setup times. *IIE Transactions, 31*(4), 339–352.

Calabrese, J. M. (1992). Optimal workload allocation in open networks of multiserver queues. *Management Science, 38*(12), 1792–1802.

Carter, A. J., Gould, J. B., Vanberkel, P., Jensen, J. L., Cook, J., Carrigan, S., ... Travers, A. H. (2015). Offload zones to mitigate emergency medical services (EMS) offload delay in the emergency department: A process map and hazard analysis. *CJEM, 17*(06), 670–678.

Cruz, F. R., & Smith, J. M. (2007). Approximate analysis of M/G/c/c state-dependent queueing networks. *Computers & Operations Research, 34*(8), 2332–2344.

Gail, H., Hantler, S., & Taylor, B. (1988). Analysis of a non-preemptive priority multiserver queue. *Advances in Applied Probability, 20*(4), 852–879.

Gelenbe, E., Pujolle, G., & Nelson, J. (1998). *Introduction to queueing networks* (Vol. 2). Chichester: Wiley.

Gross, D., Shortle, J., Thompson, J., & Harris, C. (2008). *Fundamentals of queueing theory*. New York: Wiley.

Han, Y., & Smith, J. M. (1991). Approximate analysis of M/M/C/K queueing networks. In R. O. Onvural & I. F. Akylldiz (Eds.), *Queueing networks with finite capacity* (pp. 113–126). The Netherlands: Elsevier Science.

Jain, S., & Smith, J. M. (1994). Open finite queueing networks with M/M/C/K parallel servers. *Computers & Operations Research, 21*(3), 297–317.

Kella, O., & Yechiali, U. (1985). Waiting times in the non-preemptive priority M/M/c queue. *Stochastic Models, 1*(2), 257–262.

Laan, C. M., Vanberkel, P. T., Boucherie, R. J., & Carter, A. J. (2016). Offload zone patient selection criteria to reduce ambulance offload delay. *Operations Research for Health Care, 11*, 13–19.

Latouche, G., & Neuts, M. F. (1980). Efficient algorithmic solutions to exponential tandem queues with blocking. *SIAM Journal on Algebraic Discrete Methods, 1*(1), 93–106.

Lee, H., & Park, T. (1995). Allocating data and workload among multiple servers in a local area network. *Information Systems, 20*(3), 261–269.

Perros, H. G. (1994). *Queueing networks with blocking*. Oxford: Oxford University Press.

Stidham, S., Jr. (2009). *Optimal design of queueing systems*. Chapel Hill, NC: CRC Press.

van Vuuren, M., Adan, I. J. B. F., & Resing-Sassen, S. A. E. (2005). Performance analysis of multi-server tandem queues with finite buffers and blocking. *OR Spectrum, 27*(2), 315–338.

Appendix 1.

Proof of Theorem 1: Since the Markov chain is irreducible and has a finite number of states, it is ergodic and its limiting probabilities exist. It is well known that the proofs of time-reversibility of a Markov chain and the existence of the limiting probabilities of the Markov chain can be obtained simultaneously. Consider any two states $(n_1, \ldots, n_k, \ldots, n_K)$ and $(n_1, \ldots, n_k + 1, \ldots, n_K)$. It is easy to see that the transition rate from state $(n_1, \ldots, n_k, \ldots, n_K)$ to $(n_1, \ldots, n_k + 1, \ldots, n_K)$ is $\lambda_a p_k$, and $\mu_{T,k}(n_k + 1)$ from $(n_1, \ldots, n_k + 1, \ldots, n_K)$ to $(n_1, \ldots, n_k, \ldots, n_K)$. Then it is easy to verify that

$$\lambda_a p_k \pi(n_1, \ldots, n_k, \ldots, n_K) = (n_k + 1)\mu_{T,k}\pi(n_1, \ldots, n_k + 1, \ldots, n_K), \tag{A1}$$

for all states $(n_1, \ldots, n_K)$ and k. Since $\{\pi(n_1, \ldots, n_K),\ n_1 + \cdots + n_K \leq N,\ n_k \geq 0,\ k = 1, 2, \ldots, K\}$ given in Equation (1) is summed up to one, it is the unique limiting distribution of the Markov chain, and the Markov chain is time-reversible. □

Proof of Corollary 1: The expression for $\pi_{loss,n}$ is obtained by definition. The expression for the departure rate of type a customers who are allocated to the kth node is also intuitive, yet a simple formal proof can be obtained easily. By definition, the departure rate of received type k service is given by:

$$\begin{aligned}
&\sum_{(n_1,\ldots,n_K):\ n_1+\cdots+n_K \leq n,\ n_j \geq 0,\ j=1,2,\ldots,K} \pi(n_1, \ldots, n_K) n_k \mu_{T,k} \\
&= \pi(0, \ldots, 0) \left(\sum_{(n_1,\ldots,n_K):\ n_1+\cdots+n_K \leq n,\ n_j \geq 0,\ j=1,2,\ldots,K} n_k \mu_{T,k} \left(\prod_{j=1}^{K} \frac{\rho_j^{n_j}}{n_j!} \right) \right) \\
&= \pi(0, \ldots, 0) \left(\sum_{(n_1,\ldots,n_K):\ n_1+\cdots+n_K < n,\ n_j \geq 0,\ j=1,2,\ldots,K} \left(\prod_{j=1}^{K} \frac{\rho_j^{n_j}}{n_j!} \right) \right) \lambda_a p_k \\
&= \left(\sum_{(n_1,\ldots,n_K):\ n_1+\cdots+n_K < n,\ n_j \geq 0,\ j=1,2,\ldots,K} \pi(n_1, \ldots, n_K) \right) \lambda_a p_k \\
&= \left(1 - \sum_{(n_1,\ldots,n_K):\ n_1+\cdots+n_K = n,\ n_j \geq 0,\ j=1,2,\ldots,K} \pi(n_1, \ldots, n_K) \right) \lambda_a p_k \\
&= p_k \lambda_a (1 - \pi_{loss,n}).
\end{aligned} \tag{A2}$$

□

Proof of Theorem 3: First, we rewrite $E[W_a]$ as:

$$E[W_a] = \frac{1}{c(\mu - p\lambda_a(1 - \pi_{loss})/c)} \frac{1}{\left(1 + c!(1 - a/c) \sum_{n=0}^{c-1} \frac{a^{n-c}}{n!}\right)} = \frac{1}{c} f_1(\mu) f_2(a), \tag{A3}$$

where $a = c\rho$. Note that $f_2(a)$ is the Erlang C function, which is increasing and convex in a, i.e., $f_2^{'} < 0$ and $f_2^{''} > 0$. Let $f_3(\mu) = f_2((\lambda + p\lambda_a(1 - \pi_{loss}))/\mu)$. It is clear that $f_3(\mu)$ is decreasing in μ, i.e., $\mathrm{d}f_3(\mu)/\mathrm{d}\mu < 0$. Further, we have

$$\frac{\mathrm{d}^2 f_3(\mu)}{\mathrm{d}\mu^2} = \frac{\mathrm{d}^2 f_2(a)}{\mathrm{d}a^2} \left(\frac{-(\lambda + p\lambda_a(1 - \pi_{loss}))}{\mu^2} \right)^2 + \frac{\mathrm{d}f_2(a)}{\mathrm{d}a} \frac{2(\lambda + p\lambda_a(1 - \pi_{loss}))}{\mu^3} \geq 0. \tag{A4}$$

Therefore, $f_3(\mu)$ is a decreasing convex function in μ. It is clear that $f_1(\mu)$ is decreasing convex in μ. Note that $E[W_a] = f_1(\mu) f_3(\mu)/c$. Combining the properties of $f_1(\mu)$ and $f_3(\mu)$, it is easy to see that $E[W_a]$ is decreasing in μ. Since the second derivative of $E[W_a]$ with respect to μ is given by $(f_1^{''} f_3 + f_1 f_3^{''} + 2f_1^{'} f_3^{'})/c$, which is nonnegative, therefore, $E[W_a]$ is decreasing convex in μ. This completes the proof. □