European Journal of Operational Research 196 (2009) 996–1007

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

An analytic finite capacity queueing network model capturing the propagation of congestion and blocking

Carolina Osorio *, Michel Bierlaire

Ecole Polytechnique Fédérale de Lausanne, Transport and Mobility Laboratory, Station 18, CH-1015 Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 8 November 2007
Accepted 23 April 2008
Available online 7 May 2008

Keywords:
Queueing
Queueing networks
Finite capacity
Blocking

ABSTRACT

Analytic queueing network models often assume infinite capacity queues due to the difficulty of grasping the between-queue correlation. This correlation can help to explain the propagation of congestion. We present an analytic queueing network model which preserves the finite capacity of the queues and uses structural parameters to grasp the between-queue correlation. Unlike pre-existing models it maintains the network topology and the queue capacities exogenous. Additionally, congestion is directly modeled via a novel formulation of the state space of the queues which explicitly captures the blocking phase. The model can therefore describe the sources and effects of congestion.

The model is formulated for networks with an arbitrary topology, multiple server queues and blocking-after-service. It is validated by comparison with both pre-existing methods and simulation results. It is then applied to study patient flow in a network of units of the Geneva University Hospital. The model has allowed us to identify three main sources of bed blocking and to quantify their impact upon the different hospital units.

1. Introduction

Detecting the sources and effects of congestion within a network allows us to better understand its behavior and to improve its performance. The study of congestion is relevant in a variety of sectors ranging from the analysis of spillbacks (i.e. the backwards propagation of congestion) in urban traffic or pedestrian traffic (Cheah and Smith, 1994) to that of hospital bed blocking (Koizumi et al., 2005) or prison cell blocking (Korporaal et al., 2000).

The most common approach to analyze network congestion is the development of simulation models that capture the details of the underlying system. They are cumbersome to use within an optimization framework. On the other hand, analytic models naturally fit within such a framework but are rarely developed due to the complexity of modeling the propagation of congestion while preserving a flexible model. We focus on analytic models and more specifically on analytic queueing network models.

When modeling a network using a queueing theory framework it is crucial to capture the interactions between the queues. Consider a network of hospital units (e.g. operative and post-operative units) where each unit is modeled as a specific queue and where it is the patient flow that is of main interest. For such a network understanding the correlation between the occupation of the different units can help to avoid bed blocking and to improve a patients recovery procedure. More generally, the between-queue correlation helps to explain the propagation of congestion as well as its effects (such as spillbacks). Moreover, in networks containing loops spillbacks are of special interest because they may lead to deadlocks (also known as gridlocks) (Daganzo, 1996).

The most researched queueing network model is the Jackson network model (Jackson, 1963, 1957) which assumes infinite capacity for all queues. Infinite capacity is a strong assumption that is often maintained due to the difficulty of grasping the between-queue correlation of finite capacity networks. In order to capture this correlation we resort to models with finite capacity queues. The main challenge of such an approach lies in adequately grasping this correlation while also maintaining a tractable model.

Exact finite capacity queueing network (FCQN) models exist only for networks with two or three queues with specific topologies. For more general networks FCQN models are based on approximation methods. Existing analytic FCQN models based on approximation methods either revise queue capacities or vary the network topologies. If queue capacities are revised then they become endogenous parameters. Moreover, approximations need to be used to ensure their integrality and their positivity is only checked a posteriori. We propose an FCQN model which preserves these parameters as exogenous.

Moreover, in this model congestion is not regarded as an underlying phenomenon but is directly modeled. More specifically, we

* Corresponding author. Tel.: +41 21 693 9327; fax: +41 21 693 8060.
E-mail addresses: carolina.osoriopizano@epfl.ch (C. Osorio), michel.bierlaire@epfl.ch (M. Bierlaire).

0377-2217/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2008.04.035

propose a novel formulation of the state space of the queues that explicitly models the blocking phase. Few analytic models incorporating blocking have been developed and there is a recently recognized need for them: "The next generation of the methodology would include an approximation of the blocking of patients in the queueing model" (Cochran and Bharti, 2006). Our formulation yields performance measures that describe both the sources and the effects of congestion.

This paper is structured as follows. We describe the FCQN framework and then review the existing models. The proposed model is then described, followed by its validation versus both pre-existing methods and simulation results. The model is then applied to the study of patient flow within a network of units of the Geneva University Hospital.

2. General framework

We are interested in evaluating the performance of a network of queues. A job is the generic name for the units of interest that flow through the network, e.g. a pedestrian, a prisoner, a patient. We consider open queueing networks where jobs are allowed to leave the network and where the external arrivals arise from an infinite population of jobs. We now describe the general process that a job goes through upon arrival to a queue. Jobs arriving to a queue are either served immediately or wait until a server becomes available. Once a job is served it is routed to its next queue according to a probabilistic routing model. We call this queue the target queue. If this target queue has finite capacity then it may be full. If it is full then the job is **blocked** at its current location. Once there is a place at the target queue the job is unblocked and proceeds to the target queue. The jobs are unblocked with a first in first out (FIFO) mechanism.

Various blocking mechanisms have been defined in the literature (Balsamo et al., 2001). They differ either in the moment the job is considered to be blocked (e.g. before or after-service) or in the routing mechanism of blocked jobs. The blocking mechanism that we have just described is known as blocking-after-service.

The average arrival rate to queue i is denoted λ_i. Queue i has c_i parallel servers, each one serving with an average rate μ_i. The total number of jobs allowed in the queue is called the capacity of the queue, k_i, the buffer size is $k_i - c_i$. The possible routings among queues are given by the transition probability matrix (p_{ij}), where p_{ij} denotes the probability that a job at queue i is routed to queue j.

3. Literature review

A first survey of FCQN models was made by Perros (1984), who later on also wrote a historical overview of the research motivations and advances in networks with blocking (Perros, 2003). A detailed introductory book was written by Balsamo et al. (2001). Surveys focusing on specific application fields exist for the software architecture sector (Balsamo et al., 2003), the production and manufacturing sector (Papadopoulos and Heavey, 1996) and on retrial queues for the telecommunications sector (Artalejo, 1999).

The joint stationary distribution of the network, which contains the probability of each possible state of the network, allows us to derive the main network performance measures. We distinguish between models that allow the exact evaluation of this joint stationary distribution and those based on approximation methods.

3.1. Exact methods

Exact methods consist of either closed form expressions or numerical evaluation of the joint stationary distribution. For an FCQN the between-queue correlation suggests a non-product form joint stationary distribution. Thus closed form expressions are difficult to obtain and are only available for single server networks with two or three queues in tandem topologies (Grassman and Derkic, 2000; Langaris and Conolly, 1984; Latouche and Neuts, 1980; Konheim and Reiser, 1978; Konheim and Reiser, 1976) or two queues in closed networks (Akyildiz and von Brand, 1994; Balsamo and Donatiello, 1989).

On the other hand, exact numerical evaluation of the joint stationary distribution can be obtained by solving the global balance equations (these are detailed in Section 4.1). A detailed description of these numerical methods can be found in Stewart (2000). These equations require the construction of the transition rate matrix, i.e. the description of the transition rates between all feasible states of the network. This time consuming task is therefore only conceivable for small networks (i.e. small in the number of queues and their capacity). This approach also lacks flexibility because changes in the network topology require redefining the transition rate matrix. If the networks of interest have a more general topology or an arbitrary size then their analysis is done by models based on approximation methods.

3.2. Approximation methods

Models based on approximation methods can be classified into either simulation-based or analytic models. The use of simulation models is the most popular approach to evaluate the performance of finite capacity queueing networks. Surveys of simulation models exist for sectors such as transportation (Nagel, 2002; Ben-Akiva et al., 2001), healthcare (Fone et al., 2003; Jun et al., 1999), computer science (Sadoun, 2000; Obaidat, 1990) and the analysis of call centers (Koole and Mandelbaum, 2002; Mandelbaum, 2001). This approach although more realistic and detailed, is cumbersome to optimize, and its accuracy is strongly dependent on the quality of the calibration data (Korporaal et al., 2000). Analytic models are simpler, less data expensive and more flexible.

The main motivation of analytic models based on approximation methods is to reduce the dimensionality of the system under study. Decomposition methods achieve this by decomposing the network into subnetworks and modeling each subnetwork independently. The structural parameters of each subnetwork (e.g. average arrival and service rates) depend on the state of other subnetworks and thus capture the correlation with other subnetworks. The main difficulty lies in obtaining good approximations for these parameters so that the stationary distribution of the subnetwork is a good estimate of its marginal stationary distribution. Given a subnetwork its stationary distribution can be obtained by either establishing a behavioral analogy with a network whose distribution has a closed (and often product) form, or by exact numerical evaluation of the global balance equations which now have a smaller dimension but are often nonlinear.

Existing models based on decomposition methods have defined subnetworks consisting of single queues, pairs of queues or triplets. We call these methods single, two queue and three queue decomposition methods, respectively. If not stated otherwise the models concern open finite capacity networks with exponentially distributed service times.

The most commonly used decomposition method is single queue decomposition. The first model based on this method dates back to the work of Hillier and Boling (1967) who considered tandem single server networks. One of the most used models based on single queue decomposition concerns single server feed-forward networks where each finite capacity queue is transformed into an M/M/1 queue, and the blocking is taken into account by revising the arrival and service rates of the queues (Takahashi et al., 1980). An extension of this model to queues with multiple servers is given

by Koizumi et al. (2005). Each queue is treated as an M/M/c queue for which closed form expressions of the performance measures are used. The buffers are considered infinite for each isolated queue. This approximation holds if the capacity of adjacent predecessor queues can accommodate the average queue length of the downstream queues. This constraint is checked only a posteriori.

A model applicable to networks with an arbitrary topology is given by Korporaal et al. (2000). The individual queues are modeled as M/M/c/K queues for which closed form performance measures are used. As for the method of Koizumi et al. (2005) the capacity of the queues are revised. Here the average queue length updates the capacity of predecessor queues. They use linear interpolation in order to ensure the integrality of the capacities, and their positivity is verified a posteriori.

The Expansion method (Kerbache and Smith, 1988, 1987, 2000), was developed for networks of M/M/1/K queues. Here a network reconfiguration expands all finite capacity queues to artificial infinite capacity holding queues, which register the blocked jobs. This model was later extended to multiple servers and applied to pedestrian traffic flows by Cheah and Smith (1994). Gupta and Kavusturucu (2000) applied this model to production feed-forward systems, where service interruptions are allowed. Singh and Smith (1997) used it to evaluate network performance measures within a buffer allocation problem. A similar transformation where all GE/GE/c/K queues are transformed into GE/GE/c queues, and thus the joint distribution is approximated by a product form joint distribution, was proposed by Tahilramani et al. (1999).

Models based on single queue decomposition have also been proposed for single server networks with phase-type service distributions for both tandem (Altiok, 1982) and feed-forward topologies (Altiok and Perros, 1987). Jun and Perros (1988) have extended this work to an arbitrary topology and have also considered general service times for an open tandem network in Jun and Perros (1990). The use of a phase-type service distribution accounts for all possible blockings but, as stated in Altiok and Perros (1987), it requires the construction of very detailed phase-type service mechanisms, which is a cumbersome and CPU intensive task for large networks. In these models queue capacity is also augmented in order to allow for storage of all predecessor queue capacities.

Few authors have considered subnetworks larger than single queues. Models based on two queue decomposition methods have been proposed for open tandem networks (Alfa and Liu, 2004; Brandwajn and Jow, 1988; Brandwajn and Jow, 1985) and for an arbitrary topology (Lee et al., 1998). Two queue decomposition was used by van Vuuren et al. (2005) to study multiple server tandem queues with generally distributed service times. As an extension of the work by Brandwajn and Jow (1988), Schmidt and Jackman (2000) proposed a model based on a three queue decomposition method for a single server arbitrary topology network. Subnetworks consisting of more than one queue can theoretically provide more accurate results than single queue decomposition, but are computationally more intensive (Perros, 1994).

In order to acknowledge the finite capacity property of the networks the existing models modify either the network topologies or the queue capacities. In both cases a posteriori validations are required. Additionally, if queue capacities are revised then approximations are needed in order to guarantee their integrality. We believe that a flexible and optimization friendly model is one that maintains the network topology and its configuration (number of queues and their capacities) as exogenous parameters. We propose such a method. We are also interested in explicitly modeling the blocking phase within our analytical approach. The outputs of this model therefore provide a description of both the causes and the effects of congestion.

4. Model

In this section, we describe a model that allows the analysis of a network of finite capacity queues. The model accounts for multiple server queues with an arbitrary topology and blocking-after-service. The model is based on a decomposition of the network into single queues. Let $\pi(i)$ denote the stationary distribution of the isolated queue i. The main aim of our method is to make $\pi(i)$ a good estimate of the marginal stationary distribution of queue i.

4.1. Global balance equations

The distribution $\pi(i)$ can be obtained via the global balance equations along with the use of a normalizing constraint:

$$\begin{cases} \pi(i)Q(i) = 0, \\ \sum_{s\in\mathscr{S}(i)} \pi(i)_s = 1, \end{cases} \tag{1}$$

where $\pi(i)_s$ denotes element number s of $\pi(i)$. The global balance equations involve the state space of queue i, $\mathscr{S}(i)$, as well as the transition rate matrix, $Q(i)$, which is a square matrix. We now define these two elements.

4.1.1. State space, $\mathscr{S}(i)$

Since, we are interested in explicitly modeling the blocking phase that a job may go through we define the processing of a job as follows. A job

1. arrives to a queue,
2. waits if all the servers are occupied,
3. is served (this is called the active phase),
4. is blocked if its target queue is full (this is called the blocking phase),
5. leaves the queue.

The state of queue i at any point in time is thus described by the number of active jobs A_i, blocked jobs B_i and waiting jobs W_i. The sample space of this triplet of random variables (A_i, B_i, W_i) is called the state space and is defined as $\mathscr{S}(i) = \{(a,b,w) \in \mathbb{N}^3,\ a+b \leqslant c_i,\ a+b+w \leqslant k_i\}$, where c_i is the number of servers and k_i is the capacity.

4.1.2. Transition rate matrix, $Q(i)$

The matrix $Q(i)$ contains the transition rates between all pairs of states in $\mathscr{S}(i)$. Hereafter all rates are rates averaged over time. The non-diagonal elements, $Q(i)_{sj} s\neq j$, represent the rate at which the transition between state s and j takes place. The diagonal elements are defined as $Q(i)_{ss} = -\sum_{j\neq s} Q(i)_{sj}$. Thus $-Q(i)_{ss}$ represents the rate of departure from state s. Each equation of the system of global balance equations can be written as

$$\sum_{j\neq s} \pi(i)_j Q(i)_{js} = -\pi(i)_s Q(i)_{ss},$$

it therefore balances the inflow and the outflow for a given state s. We define $Q(i)$ as a function of the following structural parameters:

- λ_i: the arrival rate to queue i;
- μ_i: the service rate of a server at queue i;
- $\mathscr{P}_i$: the probability of being blocked at queue i;
- $\tilde{\mu}_{ib}$: the unblocking rate at queue i given that there are b blocked jobs. The vector that considers all possible values of b is denoted $\tilde{\mu}_{i\bullet}$.

These four parameters allow us to describe the transition rates between the different states of queue i. We write $Q(i) = f(\lambda_i, \mu_i, \tilde{\mu}_{i\bullet}, \mathscr{P}_i)$.

As emphasized by Korporaal et al. (2000), the main challenge of models based on decomposition methods is to appropriately approximate these structural parameters so that $\pi(i)$ is a good estimate of the marginal stationary distribution of queue i. We now describe how this is done.

4.2. Transition rates

Assume that queue i is in a given feasible state s such that $s = (a, b, w)$. The possible transitions with their corresponding rates are displayed in Table 1. The set of possible states to where a transition can take place are tabulated in the second column, the corresponding transition rate is in the third column and the conditions under which such a transition can take place are in the last column. The first two lines of the table distinguish between an arrival that can be served immediately and an arrival that must queue before being served. The next two lines concern the completion of an active phase that is not followed by a blocking phase. In the first case, the freed server remains available. In the second case, the freed server immediately starts serving a job that was waiting. The fifth line concerns jobs that have completed their service and become blocked. The last two lines relate to the completion of the blocking phase. They differ in whether the server that was blocked stays available or immediately starts serving a job that was waiting. This table describes how we approximate the transition rates using structural parameters. We now describe the approximations used for these structural parameters.

4.2.1. Arrival rate, λ_i

We model each queue as an M/M/c/K queue (the distributional assumptions are detailed further on). For these models, known as loss models, all the arrivals that arise while the queue is full are considered to be lost. In our model we assume that only external arrivals may be lost, whereas arrivals that arise from within the network are blocked if the target queue is full. We therefore approximate the arrival rates by combining flow conservation with loss model information. We denote by

- λ_i: the total arrival rate to queue i (includes potentially lost arrivals);
- λ_i^{eff}: the effective arrival rate to queue i (accounts only for the arrivals that are actually processed, i.e. excludes all lost arrivals);
- γ_i: the external arrival rate to queue i.

Accounting for the lost arrivals we have

$$\lambda_i = \lambda_i^{\text{eff}}/(1 - P(N_i = k_i)), \tag{2}$$

where N_i denotes the total number of jobs at queue i, and $P(N_i = k_i)$ is known as the blocking probability.

In most existing decomposition methods the arrival rate is obtained via the flow conservation equations. In the loss model context, the flow conservation laws hold for the effective arrival rates and are approximated as follows:

$$\lambda_i^{\text{eff}} = \gamma_i(1 - P(N_i = k_i)) + \sum_j p_{ji}\lambda_j^{\text{eff}}. \tag{3}$$

Inter-arrival times to queue i are assumed to be independent and identically distributed exponential variables with parameter λ_i.

4.2.2. Probability of being blocked, $\mathscr{P}_i$

The probability of being blocked at queue i, $\mathscr{P}_i$, helps us to describe the rate at which a job gets blocked after-service. It is approximated by the weighted average of the blocking probabilities of all target queues

$$\mathscr{P}_i = \sum_j p_{ij}P(N_j = k_j). \tag{4}$$

4.2.3. Service and unblocking rates, μ_i and $\tilde{\mu}_{ib}$

The average service rate of a server at queue i is μ_i. It accounts for the active phase. It is an exogenous parameter.

We now describe how we approximate $\tilde{\mu}_{ib}$, the average unblocking rate at queue i given that there are b blocked jobs. Suppose that queue i is in the state (a, b, w). Then the service rate of the queue is $a\mu_i$, i.e. the active jobs are being processed by a **parallel** servers. In the state (a, b, w) there are b blocked servers, but they do not all work in parallel, as we now describe. Let $D(i, b)$ denote the number of distinct target queues that are blocking the b jobs at queue i. Each target queue unblocks jobs at queue i at its own rate, which we call the acceptance rate of blocked jobs. We approximate the acceptance rate of a target queue by the average acceptance rate (the average is taken across the different target queues), denoted $\tilde{\mu}_i^{\text{a}}$. Thus if all b jobs are blocked by the same target queue, then they can be seen as forming a virtual queue in front of the blocking queue with a FIFO unblocking mechanism. The average unblocking rate at queue i is then $\tilde{\mu}_i^{\text{a}}$. If the jobs are blocked by $D(i, b)$ distinct target queues then they can be seen as forming $D(i, b)$ virtual **parallel** queues, each with a FIFO unblocking mechanism. The average unblocking rate at queue i is then $D(i, b)\tilde{\mu}_i^{\text{a}}$. More specifically we have

$$\frac{1}{\tilde{\mu}_{ib}} = \sum_{d=1}^{min(b,\text{card}(\mathscr{I}^+))} P(D(i, b) = d)\frac{1}{d\tilde{\mu}_i^{\text{a}}}, \tag{5}$$

where $\mathscr{I}^+$ represents the set of target queues of queue i, and $\text{card}(\mathscr{I}^+)$ is its cardinality. Eq. (5) holds because we approximate the acceptance rate of the different target queues by a common acceptance rate, $\tilde{\mu}_i^{\text{a}}$. The approximation for $P(D(i, b) = d)$ is described in the Appendix and involves only exogenous parameters. Thus we write $\tilde{\mu}_{ib}$ in the form

$$\tilde{\mu}_{ib} = \tilde{\mu}_i^{\text{a}}\ \phi(i, b), \tag{6}$$

where $\phi(i, b)$ is exogenous and can be interpreted as the average number of distinct target queues that are blocking the b jobs at queue i ($\phi(i, b)$ is defined in the Appendix by Eq. (12)). We now describe how we approximate $\tilde{\mu}_i^{\text{a}}$.

Table 1
Transition rates of queue i

Initial state s	New state j	Rate $Q(i)_{sj}$	Condition
(a, b, w)	$(a+1, b, w)$	λ_i	$a + b + 1 \leqslant c_i$
(a, b, w)	$(a, b, w+1)$	λ_i	$(a + b == c_i)$ & $(w + 1 \leqslant k_i - c_i)$
(a, b, w)	$(a-1, b, w)$	$a\mu_i(1 - \mathscr{P}_i)$	$w == 0$
(a, b, w)	$(a, b, w-1)$	$a\mu_i(1 - \mathscr{P}_i)$	$w \geqslant 1$
(a, b, w)	$(a-1, b+1, w)$	$a\mu_i\mathscr{P}_i$	Always possible
(a, b, w)	$(a, b-1, w)$	$\tilde{\mu}_{ib}$	$w == 0$
(a, b, w)	$(a+1, b-1, w-1)$	$\tilde{\mu}_{ib}$	$w \geqslant 1$

4.2.3.1. The acceptance rate of blocked jobs, $\tilde{\mu}_i^a$. The scalar $\tilde{\mu}_i^a$ denotes the rate at which a target queue of queue i accepts (i.e. unblocks) jobs that are blocked at queue i. We denote by

- μ_i^{eff}: the effective service rate of a server at queue i (it includes service and blocking). We describe its approximation further on.
- $\tilde{p}_{ij}$: the transition probabilities conditional on a job being blocked at queue i, i.e. $\tilde{p}_{ij} = p_{ij}P(N_j = k_j)/\mathscr{P}_i$.
- r_{ij}: the proportion of arrivals to queue j that arise from blocked jobs at queue i, i.e. $r_{ij} = \tilde{p}_{ij}\lambda_i^{\text{eff}}/\lambda_j^{\text{eff}}$.

Suppose queue j is blocking jobs at predecessor queues. It is therefore full and is serving at rate $\mu_j^{\text{eff}}c_j$. It accepts jobs that are blocked at queue i at the rate $r_{ij}\mu_j^{\text{eff}}c_j$. By averaging over the possible target queues of queue i we obtain an approximation for $\tilde{\mu}_i^a$:

$$\frac{1}{\tilde{\mu}_i^a} = \sum_j \tilde{p}_{ij}\frac{1}{r_{ij}\mu_j^{\text{eff}}c_j} = \sum_{j\in\mathscr{F}^+}\frac{\lambda_j^{\text{eff}}}{\lambda_i^{\text{eff}}\mu_j^{\text{eff}}c_j}. \tag{7}$$

Eq. (7) requires an approximation for the effective service rate of a server, μ_i^{eff}.

4.2.3.2. The effective service rate, μ_i^{eff}. The total time spent by a job in front of a server, called the effective service time $1/\mu_i^{\text{eff}}$, is composed of the service time (active phase) and for some jobs of the blocked time (blocking phase). Let T_i^B denote the blocked time of a job conditional on it being blocked. A given job has an average service time of $1/\mu_i$, is blocked with probability $\mathscr{P}_i$ and once it is blocked the average time it spends blocked is $E[T_i^B]$. We therefore obtain the following approximation for the effective service rate:

$$\frac{1}{\mu_i^{\text{eff}}} = \frac{1}{\mu_i} + \mathscr{P}_i E[T_i^B]. \tag{8}$$

In this equation, μ_i is an exogenous parameter, the approximation of $\mathscr{P}_i$ is given in Eq. (4), and that of $E[T_i^B]$ is detailed in the Appendix.

4.2.3.3. Distributional assumptions. Service time and the time between successive unblockings are each assumed to follow an exponential distribution with parameters μ_i and $\tilde{\mu}_{ib}$, respectively. For a given queue all service times are assumed to be independent and identically distributed, as are all blocked times. By explicitly modeling both of these exponential phases, the number of jobs in front of the servers becomes a two-dimensional system (a, b) composed of the active and the blocked jobs. We are thus in the presence of an M/M/c/K model with a three-dimensional state space (a, b, w). By working in this space we avoid constructing the CPU intensive phase-type service mechanisms defined in some of the pre-existing methods.

4.3. System of equations

The main aim is to obtain the stationary distributions of each queue, $\pi(i)$. The main equations consist of the global balance equations (Eq. (1)), which require the definition of the transition rate matrix (Table 1). We have directly implemented these equations as a single set

$$\pi(i)g(\lambda_i, \mu_i, \tilde{\mu}_{i\bullet}, \mathscr{P}_i) = 0. \tag{9}$$

The system of nonlinear Eqs. (2)–(4), (6)–(9) is solved simultaneously for all queues. For each queue the exogenous parameters are $c_i, k_i, p_{ij}, \mu_i, \gamma_i, \phi(i, b)$. The system of equations has been implemented in terms of six endogenous parameters: $\lambda_i, \tilde{\mu}_i^a, \mu_i^{\text{eff}}, \mathscr{P}_i, P(N_i = k_i), P(B_i > 0)$. For a given queue the dimension of its distribution is equal to $\text{card}(\mathscr{S}_i) = (c_i + 1)(k_i + 1 - \frac{c_i}{2})$. Thus the total size of the system of equations is $\sum_i((c_i + 1)(k_i + 1 - \frac{c_i}{2}) + 6)$.

Pre-existing methods that require a posteriori validations (e.g. to ensure the integrality of endogenous queue capacities) resort to iterative methods. For a given iteration the system of equations for each queue is solved sequentially. Since our method requires no a posteriori validations we are able to solve the set of equations associated to all queues simultaneously.

The system is solved by using the Matlab routine *fsolve*, which implements a trust-region dogleg algorithm based on the method described by Powell (1970). The jacobian of the system has been calculated analytically and implemented. In order to ensure the positivity of the distributions the system of equations has been implemented in terms of an auxiliary variable $y(i)$ such that $y(i)^2 = \pi(i)$.

The endogenous parameters are initialized as follows. The arrival rates, λ, are initialized using the arrival rates that satisfy the classical flow conservation laws. The distributions, π, are initialized using uniform distributions, thus no a priori information concerning the stationary behavior of the queues is required, but such information could be used if available. The other endogenous parameters are deduced from these initializations.

5. Validation

5.1. Validation versus pre-existing methods

5.1.1. Triangular topology

We first compare our method with that of Altiok and Perros (1987) and that of Takahashi et al. (1980). The latter considered a single server network with triangular topology (depicted in Fig. 1), and the following configuration: $p_{12} = \frac{1}{2}, \gamma_1 = 1$. They considered two cases according to the buffer size of the queues: a null buffer and a buffer of size two. For each case they considered a set of scenarios with increasing service rates for queues two and three. These scenarios are displayed in Table 2. The chosen performance measure was the blocking probability of queue one, $P(N_1 = k_1)$. They then compared their estimates with either simulation results or with exact results derived by using the global balance equations of the entire network. The relative error of the estimates of the different methods are displayed in Fig. 2. For both cases all methods yield good estimates, the relative error remaining under 7% for the first case and 4% for the second case. We yield similar estimates to those of Takahashi et al. (1980). For the first case Altiok and Perros (1987) yields the most accurate estimates.

5.1.2. Two queues in a tandem topology

Bell (1982) derived a theoretical upper bound on the mean throughput rate of M/M/c/K networks. By considering two queues in a tandem topology under a set of scenarios he showed that

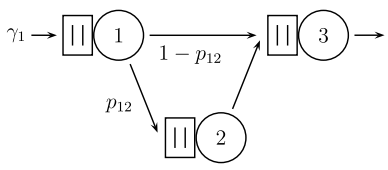

Fig. 1. Triangular topology.

Table 2
Increasing service rate scenarios

Scenario	1	2	3	4	5	6	7	8	9	10
μ_1	1	1	1	1	1	1	1	1	1	1
μ_2	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
μ_3	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3

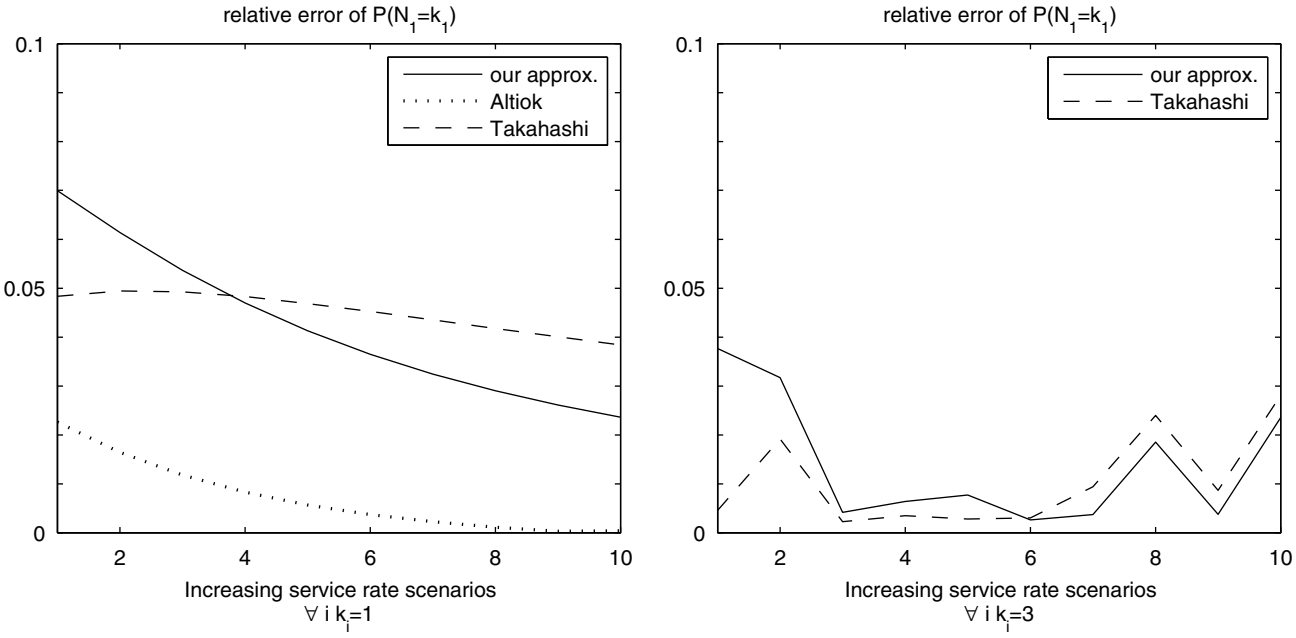

Fig. 2. Comparison with the methods of Altiok and Perros (1987) and of Takahashi et al. (1980) under two capacity configurations.

Table 3
Increasing buffer size scenarios

Scenario	1	2	3	4	5	6	7	8	9
$k_1 - c_1$	1	1	2	2	2	3	4	5	10
$k_2 - c_2$	1	2	1	2	3	3	4	5	10

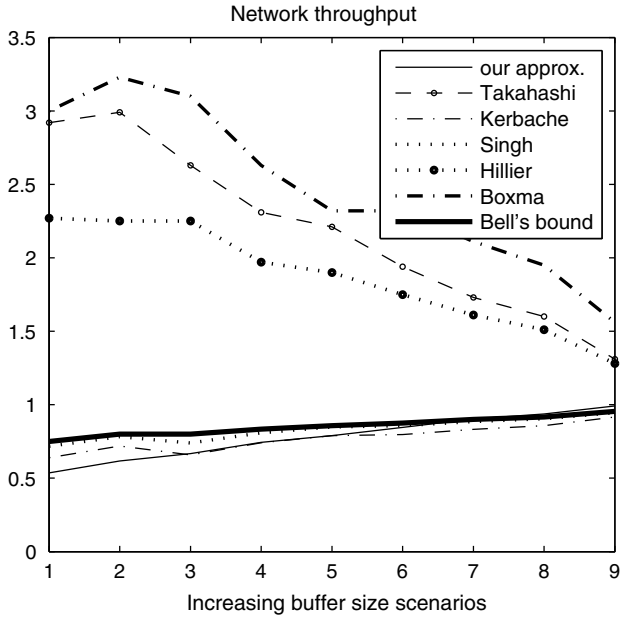

Fig. 3. Comparison of the mean throughput estimate of various decomposition methods with the theoretical upper bound derived by Bell (1982).

several models based on decomposition methods "lead to impossible mean throughput rates". We compare the mean throughput estimates of our method with the methods of Singh and Smith (1997), Kerbache and Smith (1988), Boxma and Konheim (1981), Takahashi et al. (1980) and Hillier and Boling (1967). The configuration of the network is $\mu_1 = 3, \mu_2 = 1, c_1 = c_2 = 1$, and $\gamma_1 = 1$. The different scenarios are given in Table 3 and the mean throughput estimates of the various methods are depicted in Fig. 3. Our mean throughput is estimated by using the effective departure rate at queue two, λ_2^{eff}. Fig. 3 shows that our mean throughput estimate remains near the upper bound, and is similar to that of the Expansion method of Singh and Smith (1997) and Kerbache and Smith (1988). For the last three scenarios it violates the bound by 0.3%, 2.2% and 3.8%, respectively. Our method therefore yields consistent throughputs unlike the methods of Boxma and Konheim (1981), Takahashi et al. (1980), Hillier and Boling (1967).

5.2. Validation versus simulation results

Of main interest in our method are the distributional estimates, which allow us to derive the performance measures that describe congestion. These could not be compared with pre-existing meth-

Table 4
Configuration and scenario definitions for networks A, B and C

Network A	i:	1	2	3	4	5	6	7	8	9
	γ_i	–	0.2	0.2	0.0	0.0	0.0	0.0	0.0	0.0
	μ_i	0.3	0.3	0.3	0.1	0.01	0.014	0.1	0.4	0.5
	Scenario	1	2	3	4					
	γ_1	0.1	0.2	0.3	0.4					
Network B	i:	1	2	3	4	5	6	7	8	9
	γ_i	–	0	0	0	0	0	–	0	0
	μ_i	0.3	0.3	0.3	0.6	0.6	0.6	0.3	0.3	0.3
	Scenario	1	2	3	4	5				
	γ_1	0.1	0.3	0.5	0.7	0.9				
	γ_7	0.1	0.3	0.5	0.7	0.9				
Network C	i:	1	2	3	4	5	6	7	8	9
	γ_i	–	0	0	0	0	0	0	0	0
	μ_i	0.3	0.1	0.1	0.1	0.3	0.1	0.1	0.1	0.3
	Scenario	1	2	3	4	5				
	γ_1	0.1	0.3	0.5	0.7	0.9				

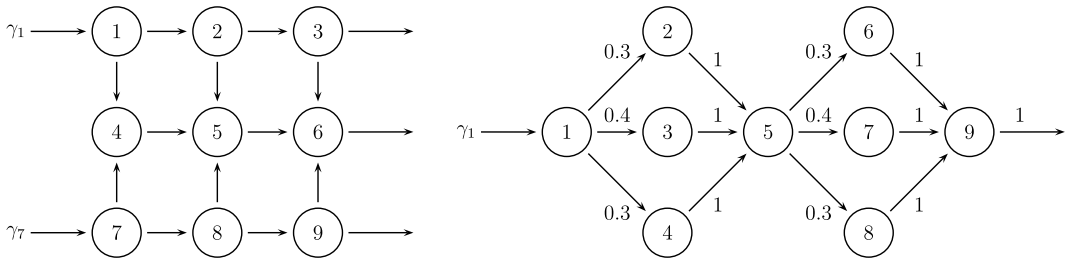

Fig. 4. Topologies of networks B and C (left and right hand side, respectively).

ods because we know of no method that defines the state space in such a way. We resort to simulation results in order to validate our method on a larger set of scenarios and topologies.

We consider three different topologies. Each network consists of nine queues, all of which are bufferless with three servers. For each network we consider a set of scenarios with increasing external arrival rates. The network configurations and scenario definitions of networks A, B and C are displayed in Table 4. Network A is a simplified version of the case study network presented in Section 6. Its topology and transition probabilities are the same as that of the case study. They are displayed in Fig. 8 and Table 7, respectively. The simplifications with regards to the case study concern the number of servers per queue and the external arrival rates. The topologies of networks B and C are displayed in Fig. 4. For a given queue of network B the transition probabilities are uniformly distributed among the possible target queues. For network C the transition probabilities are displayed in Fig. 4. In order to validate our results we developed the corresponding simulation models using a discrete event simulator, ProModel version 4.1. Let t_o denote the temporal unit of the transition rates (e.g. minutes, hours). The simulation runs consisted of 20 replications with a warm-up time of 10000 t_o and further run time of 40000 t_o.

For all three networks, all scenarios, queues and states we consider the errors of the distributional estimates: $\pi(i)_{(a,b)} - \pi^*(i)_{(a,b)}$, where $\pi(i)_{(a,b)}$ denotes our estimate of the probability that queue i is in state (a,b) and π^* is the simulation estimate. Fig. 5 displays a histogram of the errors of the distributional estimates. There are a total of 1200 estimates. 70% of the absolute errors are smaller than 0.0065, 80% smaller than 0.0129 and 90% smaller than 0.0245. Our method therefore yields good distributional estimates.

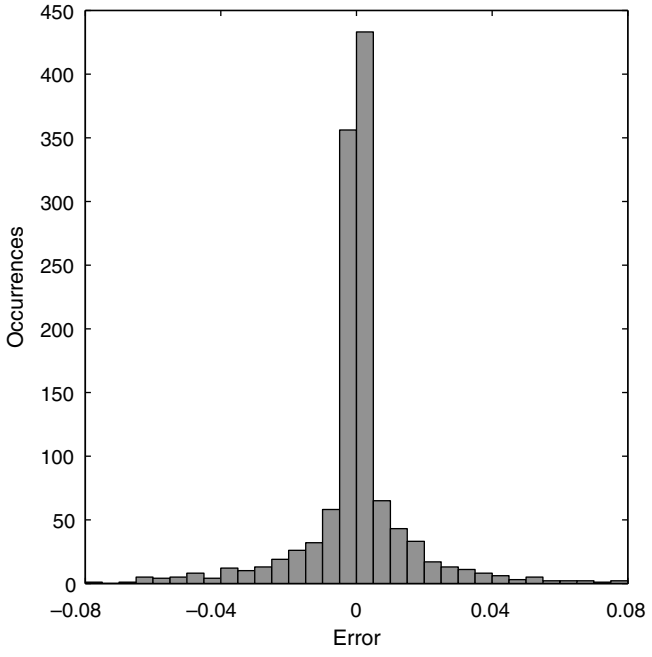

Fig. 5. Histogram of the errors of the distributional estimates for all scenarios of networks A, B and C.

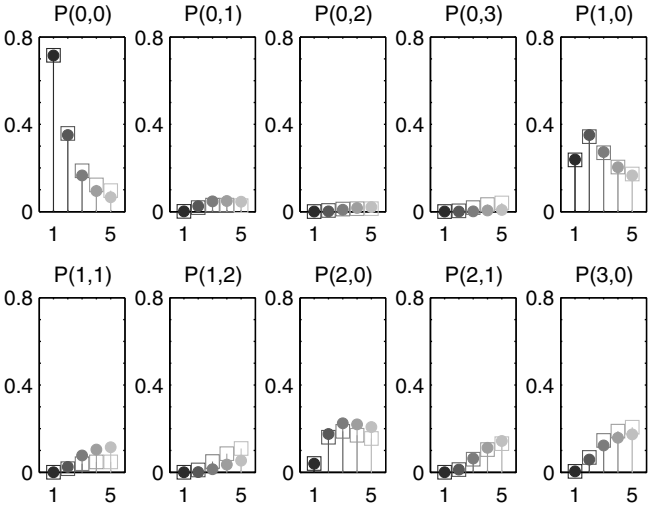

Fig. 6. Distribution of queue 5 for network C across all scenarios.

In order to illustrate the blocking information derived by our method we consider the scenarios of network C (Table 4). Fig. 6 displays the estimates of the distribution of queue five given by our method and those obtained via simulation. Each plot considers a given state (a,b) and plots $\pi(5)_{(a,b)}$ and $\pi^*(5)_{(a,b)}$ for all scenarios. The simulated distribution is depicted as empty squares, whereas our estimates are represented by filled circles. The scenarios are in a lighter color as the external arrival rate of queue one increases. The figure shows that as the external arrival rate increases the states with blocked jobs become more likely, e.g. states (a,b) in $\{(1,1),(1,2),(2,1)\}$. Take for example state $(2,1)$ where there are two active jobs and one blocked job. The probability $P(2,1)$ gradually increases from zero at scenario one to 0.14 at scenario five. For all states our estimates follow the trend of the simulated probabilities. Overall the estimates are very accurate.

5.3. Convergence of the validation runs

For a given tolerance, *tol*, convergence was attained when either the first-order optimality condition was smaller than *tol* or when both the sum of squares of the system of equations was smaller than $\sqrt{tol}$ and the change of its relative value was smaller than $max(tol^2, eps)$, where *eps* is the machine precision which is of magnitude 10^{-16}. The tolerance was chosen as $tol = 10^{-6}$. This choice is based on the criteria given in Dennis and Schnabel (1996). If after 150 iterations there was no convergence the run was stopped and

Table 5
Convergence of validation runs

Case		Number of				Time (seconds)		Total number of scenarios
		Initializations		Iterations				
Triangular	Bufferless	1	(0)	7	(1)	0.08	(0.02)	10
	Buffer of size 2	7	(4)	65	(13)	0.47	(0.1)	10
Two queues in tandem		3	(7)	37	(51)	0.2	(0.2)	9
Networks A, B and C		10	(11)	57	(46)	1.53	(1.1)	14

initialized again with a new starting point. A description of the convergence of the algorithm under the different validation runs is tabulated in Table 5. Columns two and three contain the average number of initializations required until convergence and their standard errors, respectively. Columns four to seven concern the converged run. They give the average number of iterations, their standard errors, the average execution time and their standard errors, respectively.

5.4. Tests on larger networks

In order to further evaluate the speed of our method we have applied it to a set of larger networks. We use network C as a building block. We construct the full networks by putting a set of C networks in a tandem configuration. We evaluated 70 networks, where the nth network has n instances of network C in tandem. This corresponds to networks with 9 to 630 queues. Only the first queue has external arrivals, $\gamma_1 = 0.3$. Recall that the distributions π are initialized with the uniform distribution (Section 4.3). Note that in practice a priori information would be used to initialize π. The average number of iterations required until convergence was 275 with a standard deviation of 125. Fig. 7 displays the time until converge across the networks in minutes.

Additional tests to examine the robustness of this methodology to the distributional assumptions are desirable. For applications where these assumptions do not hold the methods with phase-type distributions are adequate. This is because the phase-type distributions are dense within the class of continuous distributions (Inman, 1999; Altiok, 1989).

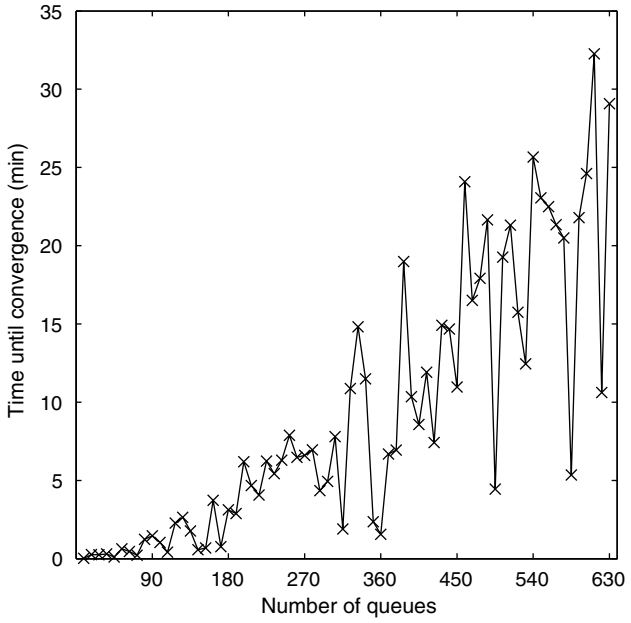

Fig. 7. Time until convergence.

6. Case study

We apply our model to the study of patient flow in a network of hospital operative and post-operative units. Clinically, bed blocking may occur for example when a recovered intensive care patient cannot proceed to the intermediate care facility due to unavailable beds. The patient is said to be blocked until his placement is possible. Studies have acknowledged that bed unavailability renders the emergency and surgical admissions procedure less flexible and less responsive (Mackay, 2001).

Modeling bed blocking and estimating its effects would bring both patient care and budgetary improvements (Cochran and Bharti, 2006; Koizumi et al., 2005). This shows the importance of modeling the bed blocking phase within a patients recovery procedure. Although few analytic models incorporating blocking have been developed, there is a recently recognized need for them (Cochran and Bharti, 2006). The existing analytic models that account for blocking in the healthcare sector have limited their study to feed-forward networks with at most three finite capacity queues (Koizumi et al., 2005; Weiss and McClain, 1987; Hershey et al., 1981).

6.1. HUG network

The hospital of interest is the Hôpitaux Universitaires de Genève (HUG, Geneva University Hospital). The considered units with their corresponding queue index in parenthesis are the emergency operating suite (indexed as queue 1), elective operating suite (2), otorhinolaryngology operating suite (3), surgical intensive care (4), medical intensive care (5), medical intermediate care (6), neuro-surgical intermediate care (7), elective recovery (8), and otorhinolaryngology recovery (9). Hereafter we refer to the units by using either their full name or their queue index.

The patients are modeled as jobs and the beds as servers. Since there is no waiting space each unit is modeled as a bufferless queue. The blocking-after-service mechanism of our model accurately mimics in-patient bed blocking.

The capacities of the different units were estimated according to the evaluations of HUG members. HUG members also extracted patient flow data which we used to estimate the exogenous parameters γ, μ and p_{ij}. Maximum likelihood estimates were used for γ and μ, whereas the transition probabilities were estimated by the transition frequencies. The data consisted of 25336 patient records ranging over a year.

The configuration of the network is presented in Table 6 and its topology is given in Fig. 8. In this figure, the dotted lines correspond to two-way arrows. The network consists of nine operative

Table 6
Configuration of the HUG network

i	1	2	3	4	5	6	7	8	9
c_i	4	8	5	18	18	4	4	10	6
γ_i	0.39	0.5	0.25	0.06	0.18	0.03	0.13	0.16	0
μ_i	0.32	0.26	0.34	0.01	0.02	0.01	0.02	0.22	0.52
card($\mathscr{S}_i$)	15	45	21	190	190	15	15	66	28

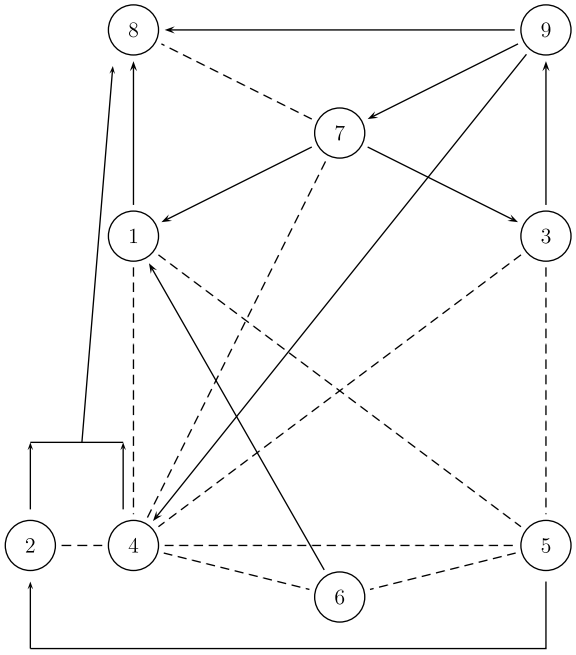

Fig. 8. HUG network topology.

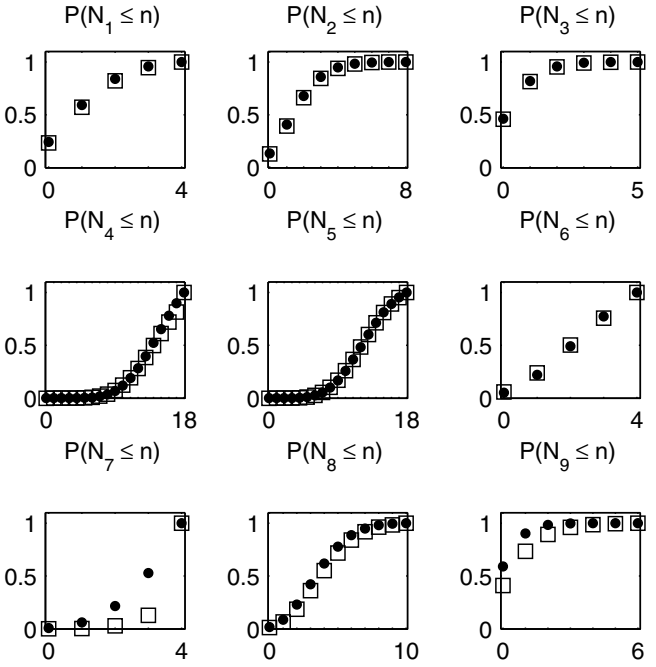

Fig. 9. Comparison of the cumulative distribution function, $P(N_i \leqslant n)$ for all queues.

Table 7
Transition probability matrix of the HUG network, p_{ij}

	1	2	3	4	5	6	7	8	9
1	–	–	–	.16	.02	–	–	.71	–
2	–	–	–	.07	–	–	–	.84	–
3	–	–	–	.03	.01	–	–	–	.95
4	.18	.01	.03	–	.03	.01	.11	.03	–
5	.05	.01	.01	.01	–	.07	–	–	–
6	.02	–	–	.01	.1	–	–	–	–
7	.05	–	.05	.04	–	–	–	.01	–
8	–	–	–	–	–	–	.01	–	–
9	–	–	–	.05	–	–	.05	.02	–

and post-operative units, with 31 possible transitions, containing numerous cycles. This makes the network prone to blocking. Table 7 contains the transition probability matrix. In this table the null probabilities are denoted by dashed lines. Note that the sum of the transition probabilities for a given unit (i.e. a given line) may not sum to 1, in this case $1 - \sum_j p_{ij}$ represents the probability of exiting the network given that the job is at queue i.

6.2. Comparison with simulation results

We have also carried out this case study using the simulator. This allowed us to compare our distributional estimates with those obtained via simulation. The simulation setup was the same as that of Section 5.2. The threshold for the stopping criteria of the algorithm was chosen as 10^{-6}. Convergence was attained after 325 iterations and 84 seconds whereas the time required to complete the simulation was 25 minutes.

We consider once again the absolute errors of the distributional estimates, their 90th, 95th, and 99th percentiles are 0.008, 0.02 and 0.07, respectively. We have four estimates that have an absolute error larger than 0.1. Overall the distributional estimates are very good. The cumulative distribution function for the total number of jobs at each queue are depicted in Fig. 9. The estimates of our method are represented by filled circles, whereas the simulation estimates are denoted by empty squares. All queues except queues seven and nine have excellent estimates.

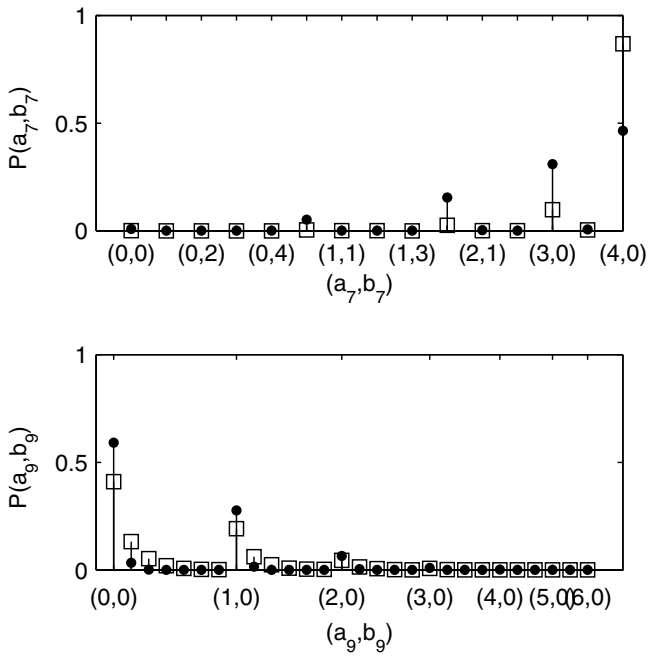

Fig. 10. Distributions of queues seven and nine.

Three of the four previously mentioned estimates with large errors concern queue seven, the fourth error concerns queue nine. Explaining the cause of these large errors is not a straightforward task given the correlation between the endogenous parameters of our system of equations. The detailed distributions of queues seven and nine are displayed in Fig. 10. The estimates of our method are represented by filled circles, whereas the simulation estimates are denoted by empty squares. The states (a, b) are ordered by increasing number of active jobs and then increasing number of blocked

Table 8
Transition probabilities conditional on a patient being blocked, $\tilde{p}_{ij}$

	1	2	3	4	5	6	7	8	9
1	–	–	–	.76	.04	–	–	.19	–
2	–	–	–	.59	–	–	–	.41	–
3	–	–	–	.87	.13	–	–	–	.01
4	.12	–	–	–	.02	.04	.82	–	–
5	.11	–	–	.05	–	.83	–	–	–
6	.13	–	–	.16	.71	–	–	–	–
7	.34	–	.01	.65	–	–	–	.01	–
8	–	–	–	–	–	–	1	–	–
9	–	-	–	.18	–	–	.82	–	–

jobs. This figure shows that for queue seven the state (4,0) is underestimated and for queue nine it is the blocked states (0,1) and (0,2) that are underestimated. These misestimations may be correlated since $\tilde{p}_{97} = 0.82$ (displayed in Table 8 and discussed later on), i.e. given that a job is blocked at queue nine the probability that is has been blocked by queue seven is 0.82. Thus the underestimation of the occupation of queue seven may lead to an underestimation of the blocking at queue nine.

6.3. Congestion analysis

6.3.1. The sources of congestion

The outputs of our model help us to quantify the blocking and also investigate its causes. The transition probabilities conditional on a patient being blocked, $\tilde{p}_{ij}$, are displayed in Table 8. These probabilities can help us to determine the source of blocking. The probabilities have been rounded to 10^{-2}, those smaller than 0.005 are denoted by a dashed line. For a given unit (i.e. a given line in the table) we can identify the target units that are more likely to block patients.

This table helps us to detect three main sources of blocking. The medical intensive care and the medical intermediate care units mutually block each others patients ($\tilde{p}_{56} = 0.83, \tilde{p}_{65} = 0.71$). The same holds for the surgical intensive care and the neuro-surgical intermediate care units ($\tilde{p}_{47} = 0.82, \tilde{p}_{74} = 0.65$). This first type of blocking (mutual blocking) may be irrelevant in practice given that the swapping of patients can be identified and carried out easily. The second source of blocking which may be more difficult to solve is the blocking at the operating suites due to the surgical intensive care unit ($\tilde{p}_{14} = 0.76, \tilde{p}_{24} = 0.59, \tilde{p}_{34} = 0.87$). Moreover, the performance of the emergency operating suite is strongly linked to its responsiveness, which is deteriorated by blocking. The third source of blocking occurs at the recovery units and is due to the neuro-surgical intermediate care unit ($\tilde{p}_{87} = 1, \tilde{p}_{97} = 0.82$).

6.3.2. The frequency and effects of congestion

By explicitly modeling the blocking phase our model yields novel performance measures that quantify the occurrence as well as the impact of congestion. Table 9 displays several performance measures of the different units. It also recalls the capacity, k_i, and the average service time, $1/\mu_i$, of the units which are exogenous parameters. $1/\mu_i$ is given in hours. It is important to notice that although $\mathscr{P}_i$ quantifies the occurrence of blocking at a given unit, it does not capture the impact that a given blocking event may have on the unit or the patient which is blocked. Take for example the otorhinolaryngology (ORL) recovery unit where $\mathscr{P}_9 = 0.03$, that is the probability of a patient getting blocked at that unit is 0.03. In this unit the average service time is 1.9 hours ($1/\mu_9$) and blocking is mainly due to the neuro-surgical intermediate care unit ($\tilde{p}_{97} = 0.82$) where the average service time is 66.7 hours ($1/\mu_7$). Thus the average blocked time at the ORL recovery due to the neuro-surgical intermediate care unit has a strong impact on the ORL recovery unit. This can also be seen when comparing $E[B_i]/E[N_i]$ with $\mathscr{P}_i$. The fact that $E[B_i]/E[N_i]$ is larger than $\mathscr{P}_i$ also indicates that although blocking may be rare the impact that it may have on the unit or on the patient is not to be ignored. In the case of the ORL recovery unit $E[B_i]/E[N_i]$ and $\mathscr{P}_i$ are equal to 0.11 and 0.03, respectively.

Table 9
Performance measures for the HUG network

i	1	2	3	4	5	6	7	8	9
k_i	4	8	5	18	18	4	4	10	6
$\frac{1}{\mu_i}$	3.1	3.9	3.0	76.9	66.7	71.4	66.7	4.6	1.9
$\mathscr{P}_i$	0.02	0.01	0.00	0.06	0.02	0.01	0.01	0.00	0.03
$E[B_i]$	0.04	0.01	0.01	0.22	0.04	0.01	0.01	0.00	0.06
$E[N_i]$	1.37	2.00	0.77	14.03	12.56	2.46	3.19	4.04	0.53

7. Conclusions and future work

We have presented an analytic queueing network model that preserves the finite capacity property of the real system. The model is formulated for multiple server finite capacity queueing networks with an arbitrary topology and blocking-after-service. The model is based on a decomposition of the network into single queues. The structural parameters of the queues are approximated so that they can account for the between-queue correlation. Unlike pre-existing methods the network topology and its configuration (number of queues and their capacity) are preserved throughout the analysis thus no constraints need to be checked a posteriori.

The originality of this method also lies in its ability to explicitly model the blocking phase that jobs may go through under congested traffic conditions. The model yields performance measures that describe congestion in terms of its sources, its frequency and its impact.

Performance measures have been validated by comparison with pre-existing methods on networks with varying buffer size or service rates. The distributional approximations have been compared with those obtained via simulation on a set of networks under a set of scenarios with varying arrival rates, namely under high intensity traffic. In both types of validations the results illustrate the good accuracy of our model. The comparisons versus a simulation-based approach also highlight the important gain in computation time since the time to estimate the parameters of our model is negligible compared to that of running a simulation.

The model has been applied to study patient flow in a network of operative and post-operative units of the Geneva University Hospital. We identified three main sources of bed blocking and quantified their impact upon the different hospital units. The performance measures of the model also revealed that although bed blocking may be a rare event its impact upon the performance of a given unit is not to be ignored.

Additional validation runs to test the sensitivity of the approximations would be desirable. Further work will focus on combining this model with a simulation model within an optimization framework, while ensuring consistency between the two models. The aim of this framework is to allow us to benefit from an optimization friendly analytic model, while accounting for fine details that can be reproduced by the simulation tool.

Like pre-exiting methods that allow for feedback topologies we have assumed that no deadlock occurs or that it is solved instantaneously (e.g. by swapping). Nevertheless we believe that it is of interest to investigate analytic deadlock detection methods.

Acknowledgements

The authors thank Philippe Garnerin and Pau Perez from the Division of Anesthesiology at the Geneva University Hospitals. This

research was supported by the Swiss National Science Foundation Grant 205321-107838.

Appendix

Approximation of $P(D(i,b)=d)$

$P(D(i,b)=d)$ represents the probability that d distinct queues are blocking the b blocked jobs at queue i. Consider $R(i,b,d)$ the random vector containing the b target queues of the blocked jobs, d of which are distinct, and let $\mathscr{R}(i,b,d)$ be its sample space. In order to approximate $P(D(i,b)=d)$ we sum over all possible realizations of $R(i,b,d)$.

$$P(D(i,b)=d)=\sum_{r\in\mathscr{R}(i,b,d)}P(R(i,b,d)=r)=\sum_{r\in\mathscr{R}(i,b,d)}\tilde{p}_{ir_1}\tilde{p}_{ir_2}\cdots\tilde{p}_{ir_b}$$
$$=\sum_{r\in\mathscr{R}(i,b,d)}\prod_{j\in\mathscr{I}^+}\tilde{p}_{ij}^{\ell(i,b,d)_j},$$

where $\ell(i,b,d)_j$ is the number of jobs blocked by queue j at queue i (given that there are a total of b blocked jobs that are blocked by d distinct target queues). This last equation shows that for a given realization of $R(i,b,d)$, what is of interest in determining $P(D(i,b)=d)$ is the occurrence of each target queue (i.e. the vector $\ell(i,b,d)$), the ordering of the target queues is not important. Thus instead of summing over $\mathscr{R}(i,b,d)$, we sum over the set of $\ell(i,b,d)$ vectors. This reduces the size of the space over which we sum. The set of such vectors is noted $\mathscr{L}(i,b,d)$ and is defined by

$$\ell(i,b,d)\in\mathscr{L}(i,b,d)\iff\begin{cases}\sum_{j\in\mathscr{I}^+}\ell(i,b,d)_j=b,\\ \sum_{j\in\mathscr{I}^+}\mathbf{1}(\ell(i,b,d)_j>0)=d,\\ \ell(i,b,d)_j\geqslant 0\quad\forall j\in\mathscr{I}^+,\end{cases}\tag{10}$$

where $\mathbf{1}(x)$ is the indicator function. The first equation of the system of Eqs. (10) means that there are a total of b jobs blocked at queue i. The second means that these jobs are blocked by d different target queues. For a given vector $\ell(i,b,d)$ that satisfies the system of Eqs. (10) there are $b!/(\prod_{j\in\mathscr{I}^+}\ell(i,b,d)_j!)$ different realizations of $R(i,b,d)$ that are associated with it. This corresponds to the number of permutations of a vector of b elements where element j is repeated $\ell(i,b,d)_j$ times. Therefore, we obtain

$$P(D(i,b)=d)=\sum_{\ell(i,b,d)\in\mathscr{L}(i,b,d)}\frac{b!}{\prod_{j\in\mathscr{I}^+}\ell(i,b,d)_j!}\prod_{j\in\mathscr{I}^+}\tilde{p}_{ij}^{\ell(i,b,d)_j}.$$

Coming back to Eq. (5) and replacing $P(D(i,b)=d)$ by the approximation that we have just derived we obtain:

$$\frac{1}{\tilde{\mu}_{ib}}=\frac{1}{\tilde{\mu}_i^{\mathrm{a}}}\sum_{d=1}^{\min(b,\mathrm{card}(\mathscr{I}^+))}\frac{1}{d}\sum_{\ell(i,b,d)\in\mathscr{L}(i,b,d)}\frac{b!}{\prod_{j\in\mathscr{I}^+}\ell(i,b,d)_j!}\prod_{j\in\mathscr{I}^+}\tilde{p}_{ij}^{\ell(i,b,d)_j}.\tag{11}$$

The size of the space $\mathscr{L}(i,b,d)$ is still considerably large therefore when approximating $\tilde{\mu}_{ib}$ we use an exogenous approximation of $\tilde{p}_{ij}$:

$$\tilde{p}_{ij}=\frac{p_{ij}P(N_j=k_j)}{\mathscr{P}_i}=\frac{p_{ij}P(N_j=k_j)}{\sum_l p_{il}P(N_l=k_l)}\approx\frac{p_{ij}}{\sum_l p_{il}}.$$

This approximation makes both summations of Eq. (11) exogenous. These two summations are therefore evaluated only once when solving the entire system of equations. This approximation is appropriate if the blocking probabilities of the target queues have the same magnitude, otherwise it is inadequate. The only endogenous parameter remaining in Eq. (11) is $\tilde{\mu}_i^{\mathrm{a}}$. Thus we have written $\tilde{\mu}_{ib}$ in the form $\tilde{\mu}_{ib}=\tilde{\mu}_i^{\mathrm{a}}\ \phi(i,b)$, where

$$\frac{1}{\phi(i,b)}=\sum_{d=1}^{\min(b,\mathrm{card}(\mathscr{I}^+))}\frac{1}{d}\sum_{\ell(i,b,d)\in\mathscr{L}(i,b,d)}\frac{b!}{\prod_{j\in\mathscr{I}^+}\ell(i,b,d)_j!}\times\prod_{j\in\mathscr{I}^+}\left(\frac{p_{ij}}{\sum_k p_{ik}}\right)^{\ell(i,b,d)_j}.\tag{12}$$

Approximation of $E[T_i^{\mathrm{B}}]$

Given a blocked job at queue i, $E[T_i^{\mathrm{B}}]$ represents its expected blocked time. Recall that B_i denotes the number of blocked jobs at queue i. We approximate $E[T_i^{\mathrm{B}}]$ by conditioning on the length of the blocked queue:

$$E[T_i^{\mathrm{B}}]=E[E[T_i^{\mathrm{B}}|B_i]]=\sum_{b\geqslant 0}P(B_i=b|B_i>0)\ E[T_i^{\mathrm{B}}|B_i=b]$$
$$=\sum_{b\geqslant 1}\frac{P(B_i=b)}{P(B_i>0)}\ E[T_i^{\mathrm{B}}|B_i=b].$$

Let $T(i,b)_j$ denote the blocked time of the job that was unblocked in j^{th} position given that there were b blocked jobs. We have

$$E[T_i^{\mathrm{B}}|B_i=b]=\frac{1}{b}\sum_{j=1}^{b}E[T(i,b)_j].$$

We know that the average time between successive departures given that there are b blocked jobs at queue i is represented by $1/\tilde{\mu}_{ib}$, thus we approximate the average blocked time of the first job to be unblocked by $1/\tilde{\mu}_{ib}$, that of the second job to be unblocked by $1/\tilde{\mu}_{ib}+1/\tilde{\mu}_{i(b-1)}$ and that of the jth by

$$E[T(i,b)_j]=\sum_{k=b-j+1}^{b}\frac{1}{\tilde{\mu}_{ik}}.$$

Putting the last two equations together and then interchanging the summations we obtain:

$$E[T_i^{\mathrm{B}}|B_i=b]=\frac{1}{b}\sum_{j=1}^{b}\sum_{k=b-j+1}^{b}\frac{1}{\tilde{\mu}_{ik}}=\frac{1}{b}\sum_{k=1}^{b}\frac{1}{\tilde{\mu}_{ik}}\sum_{j=b-k+1}^{b}1=\frac{1}{b}\sum_{k=1}^{b}\frac{k}{\tilde{\mu}_{ik}}.$$

Therefore, our approximation of $E[T_i^{\mathrm{B}}]$ is given by

$$E[T_i^{\mathrm{B}}]=\sum_{b\geqslant 1}\frac{P(B_i=b)}{P(B_i>0)}\sum_{k=1}^{b}\frac{k}{b}\frac{1}{\tilde{\mu}_{ik}}.\tag{13}$$

References

Akyildiz, I.F., von Brand, H., 1994. Exact solutions to networks of queues with blocking-after-service. Theoretical Computer Science 125 (1), 111–130.

Alfa, A.S., Liu, B., 2004. Performance analysis of a mobile communication network: The tandem case. Computer Communications 27 (3), 208–221.

Altiok, T., 1982. Approximate analysis of exponential tandem queues with blocking. European Journal of Operational Research 11 (4), 390–398.

Altiok, T., 1989. Approximate analysis of queues in series with phase-type service times and blocking. Operations Research 37 (4), 601–610.

Altiok, T., Perros, H.G., 1987. Approximate analysis of arbitrary configurations of open queuing networks with blocking. Annals of Operations Research 9 (1), 481–509.

Artalejo, J.R., 1999. Accessible bibliography on retrial queues. Mathematical and Computer Modelling 30 (3–4), 1–6.

Balsamo, S., Donatiello, L., 1989. On the cycle time distribution in a two-stage cyclic network with blocking. IEEE Transactions Software Engineering 15 (10), 1206–1216.

Balsamo, S., De Nitto Persone, V., Onvural, R., 2001. Analysis of Queueing Networks with Blocking. International Series in Operations Research and Management Science, vol. 31. Kluwer Academic Publishers, Boston.

Balsamo, S., De Nitto Persone, V., Inverardi, P., 2003. A review on queueing network models with finite capacity queues for software architectures performance prediction. Performance Evaluation 51 (2–4), 269–288.

Bell, P.C., 1982. Use of decomposition techniques for the analysis of open restricted queuing networks. Operations Research Letters 1 (6), 230–235.

Ben-Akiva, M., Bierlaire, M., Burton, M., Koutsopoulos, H., Mishalani, R., 2001. Network state estimation and prediction for real-time transportation management applications. Networks and Spatial Economics 1 (3–4), 293–318.

Boxma, O.J., Konheim, A.J., 1981. Approximate analysis of exponential queueing systems with blocking. Acta Informatica 15 (1), 19–66.

Brandwajn, A., Jow, Y., 1985. Tandem exponential queues with finite buffers. In: Hasegawa, T., Takagi, H., Takahashi, Y. (Eds.), Computer Networking and Performance Evaluation. Amsterdam, The Netherlands, North Holland, pp. 245–258.

Brandwajn, A., Jow, Y., 1988. An approximation method for tandem queues with blocking. Operations Research Letters 36 (1), 73–83.

Cheah, J.Y., Smith, J.M., 1994. Generalized M/G/C/C state dependent queueing models and pedestrian traffic flows. Queueing Systems 15 (1–4), 365–386.

Cochran, J., Bharti, A., 2006. Stochastic bed balancing of an obstetrics hospital. Health Care Management Science 9 (1), 31–45.

Daganzo, C.F., 1996. The nature of freeway gridlock and how to prevent it. In: Lesort, J.B. (Ed.), Proceedings of the 13th International Symposium on Transportation and Traffic Theory. Pergamon Press, pp. 629–646.

Dennis, J.E., Schnabel, R.B., 1996. Numerical methods for unconstrained optimization and nonlinear equations. Classics in Applied Mathematics, vol. 16. SIAM, Philadelphia.

Fone, D., Hollinghurst, S., Temple, M., Round, A., Lester, N., Weightman, A., Roberts, K., Coyle, E., Bevan, G., Palmer, S., 2003. Systematic review of the use and value of computer simulation modelling in population health and health care delivery. Journal of Public Health Medicine 25 (4), 325–335.

Grassman, W., Derkic, S., 2000. An analytical solution for a tandem queue with blocking. Queueing Systems 36 (1–3), 221–235.

Gupta, S.M., Kavusturucu, A., 2000. Production systems with interruptions, arbitrary topology and finite buffers. Annals of Operations Research 93 (1–4), 145–176.

Hershey, J.C., Weiss, E.N., Cohen, M.A., 1981. A stochastic service network model with application to hospital facilities. Operations Research 29 (1), 1–22.

Hillier, F.S., Boling, R.W., 1967. Finite queues in series with exponential or Erlang service times-a numerical approach. Operations Research 15 (2), 286–303.

Inman, R., 1999. Empirical evaluation of exponential and independence assumptions in queuing models of manufacturing systems. Production and Operations Management 8 (4), 409–432.

Jackson, J.R., 1957. Networks of waiting lines. Operations Research 5 (4), 518–521.

Jackson, J.R., 1963. Jobshop-like queuing systems. Management Science 10 (1), 131–142.

Jun, K.P., Perros, H.G., 1988. Approximate analysis of arbitrary configurations of queueing networks with blocking and deadlock. In: Perros, H.G., Altiok, T. (Eds.), Queueing Networks with Blocking: Proceedings of the First international workshop. North-Holland, Amsterdam, pp. 259–279.

Jun, K.P., Perros, H.G., 1990. An approximate analysis of open tandem queueing networks with blocking and general service times. European Journal of Operational Research 46 (1), 123–135.

Jun, J.B., Jacobson, S.H., Swisher, J.R., 1999. Application of discrete-event simulation in health care clinics: A survey. Journal of the Operational Research Society 50 (2), 109–123.

Kerbache, L., Smith, J.M., 1987. The generalized expansion method for open finite queueing networks. European Journal of Operational Research 32 (3), 448–461.

Kerbache, L., Smith, J.M., 1988. Asymptotic behaviour of the expansion method for open finite queueing networks. Computers and Operations Research 15 (2), 157–169.

Kerbache, L., Smith, J.M., 2000. Multi-objective routing within large scale facilities using open finite queueing networks. European Journal of Operational Research 121 (1), 105–123.

Koizumi, N., Kuno, E., Smith, T.E., 2005. Modeling patient flows using a queueing network with blocking. Health Care Management Science 8 (1), 49–60.

Konheim, A.G., Reiser, M., 1976. A queueing model with finite waiting room and blocking. Journal of the Association for Computing Machinery 23 (2), 328–341.

Konheim, A.G., Reiser, M., 1978. Finite capacity queueing systems with applications in computer modeling. SIAM Journal on Computing 7 (2), 210–229.

Koole, G., Mandelbaum, A., 2002. Queueing models of call centers: An introduction. Annals of Operations Research 113 (1–4), 41–59.

Korporaal, R., Ridder, A., Kloprogge, P., Dekker, R., 2000. An analytic model for capacity planning of prisons in the Netherlands. Journal of the Operational Research Society 51 (11), 1228–1237.

Langaris, C., Conolly, B., 1984. On the waiting time of a two-stage queueing system with blocking. Journal of Applied Probability 21 (3), 628–638.

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

Lee, H.S., Bouhchouch, A., Dallery, Y., Frein, Y., 1998. Performance evaluation of open queueing networks with arbitrary configuration and finite buffers. Annals of Operations Research 79 (0), 181–206.

Mackay, M., 2001. Practical experience with bed occupancy management and planning systems: An Australian view. Health Care Management Science 4 (1), 47–56.

Mandelbaum, A., 2001. Call centers (centres): Research bibliography with abstracts. Electronically available: <http://iew3.technion.ac.il/serveng/References/ccbib.pdf>.

Nagel, K., 2002. Traffic networks. In: Bornholdt, S., Schuster, H.G. (Eds.), Handbook of Graphs and Networks. Wiley VCH, LinkWeinheim, pp. 248–272.

Obaidat, M.S., 1990. Simulation of queueing models in computer systems. In: Oezekici, S. (Ed.), Queueing Theory and Applications. Taylor & Francis/Hemisphere, New York, pp. 111–151.

Papadopoulos, H.T., Heavey, C., 1996. Queueing theory in manufacturing systems analysis and design: A classification of models for production and transfer lines. European Journal of Operational Research 92 (1), 1–27.

Perros, H., 1984. Queueing networks with blocking: A bibliography. ACM SIGMETRICS Performance Evaluation Review 12 (2), 8–12.

Perros, H., 1994. Queueing networks with blocking: Exact and Approximate Solutions. Oxford University Press, New York, NY, USA.

Perros, H., 2003. Open queueing networks with blocking – A personal log. In: Kotsis, G. (Ed.), Performance Evaluation – Stories and Perspectives. Austrian Computer Society, Vienna, Austria, pp. 105–115.

Powell, M.J.D., 1970. A fortran subroutine for solving systems of nonlinear algebraic equations. In: Rabinowitz, P. (Ed.), Numerical Methods for Nonlinear Algebraic Equations. Gordon & Breach, London (Chapter 7).

Sadoun, B., 2000. Applied system simulation: A review study. Information Sciences 124 (1–4), 173–192.

Schmidt, L.C., Jackman, J., 2000. Modeling recirculating conveyors with blocking. European Journal of Operational Research 124 (2), 422–436.

Singh, A., Smith, J.M., 1997. Buffer allocation for an integer nonlinear network design problem. Computers and Operations Research 24 (5), 453–472.

Stewart, W.J., 2000. Numerical methods for computing stationary distributions of finite irreducible Markov chains. In: Grassmann, W. (Ed.), Computational Probability. Kluwer Academic Publishers, Boston (Chapter 4).

Tahilramani, H., Manjunath, D., Bose, S.K. 1999. Approximate analysis of open network of GE/GE/m/N queues with transfer blocking, MASCOTS 0, pp. 164–172.

Takahashi, Y., Miyahara, H., Hasegawa, T., 1980. An approximation method for open restricted queuing networks. Operations Research 28 (3), 594–602.

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 (2–3), 315–338.

Weiss, E.N., McClain, J.O., 1987. Administrative days in acute care facilities: A queuing-analytic approach. Operations Research 35 (1), 35–44.