A finite capacity queueing network model capturing blocking, congestion and spillbacks

Carolina Osorio and Michel Bierlaire

Abstract

Analytic queueing network models constitute a flexible tool for the study of network flow. These aggregate models are simple to manipulate and their analytic aspect renders them suitable for use within an optimization framework. Analytic queueing network models often assume infinite capacity for all queues. For real systems this infinite capacity assumption does not hold, but is often maintained due to the difficulty of grasping the between-queue correlation structure present in finite capacity networks. This correlation structure helps explain bottleneck effects and spillbacks, the latter being of special interest in networks containing loops because they are a source of potential deadlock (i.e. gridlock). We present an analytic queueing network model which acknowledges the finite capacity of the different queues. The model is adapted for multiple server finite capacity queueing networks with an arbitrary topology and blocking-after-service. By explicitly modelling the blocking phase the model yields a description of the congestion effects. A decomposition method allowing the evaluation of the model is also described. The methods validation, by comparison to both pre-existing methods and simulation results, is presented, as well as its application to the study of patient flow in a network of operative and post-operative units of the Geneva University Hospital.

1 Introduction

Modelling complex systems using queueing network models allows us to better understand their behaviour, to estimate and ultimately to improve their performance. Consider a network of operative and post-operative hospital units where each unit is modelled 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 (e.g. surgical intensive care, surgical intermediate care) can help avoid bed blocking and improve a patients recovery procedure. The most researched queueing network model is the Jackson network model (Jackson (1957), Jackson (1963)) which assumes infinite capacity for all queues. For real systems this infinite capacity assumption does not hold, but is often maintained due to the difficulty of grasping the between-queue correlation structure present in finite capacity networks; e.g. acknowledging the links between the behaviour of adjacent queues where chained events can take place. This correlation structure helps explaining bottleneck effects and spillbacks, the latter being of special interest in networks containing loops because they are a source of potential deadlocks (i.e. gridlocks) (Daganzo (1996)). In order to capture this correlation and to estimate these congestion effects we resort to models with finite capacities. As we shall detail in the literature review exact analytic results are only available for small networks with specific topologies. When wanting to model larger finite capacity networks with arbitrary topologies the main complexity lies in appropriately acknowledging the between-queue correlation while maintaining a tractable model. Finite capacity queueing network (FCQN) models have been used for a variety of applications such as the study of software architectures performance (Balsamo et al. (2003)), hospital patient flow (Koizumi et al. (2005), Cochran and Bharti (2006)), criminal flow through a network of prisons (Korporaal et al. (2000)), pedestrian flow through circulation systems (e.g. corridors) (Cheah and Smith (1994)), as well as numerous applications in the manufacturing sector (Papadopoulos and Heavey (1996)). This paper is structured as follows. We describe the FCQN framework and then review the existing analysis methods. The proposed model and approximation method are then described, followed by their validation and their application on a real scale case study.

2 General Framework

A queueing network is composed of a set of linked queues, hereafter called stations. Of interest is the study of the flow of "jobs" throughout the network. A job is the generic name for the units of interest, e.g. a pedestrian, a prisoner. 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 first describe the general process that a job goes through upon arrival to a station. Jobs arriving to a station are either served immediately or queue until a server becomes available. Once a job is served it is routed to its next station, which is chosen according to a probability distribution. If this destination station has finite capacity then it may be full. If it is full then the job will be **blocked** at its current station until a server becomes available at the destination station. Various blocking mechanisms, which are at the heart of spillbacks, 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 (BAS). The jobs are unblocked with a First In First Out (FIFO) mechanism. The average arrival rate to station i is denoted λ_i. Station i has c_i parallel servers, each one serves with an average rate μ_i. The total number of jobs allowed in the station is called the capacity of the station, K_i, the buffer size is $K_i - c_i$. The possible routings among stations are given by the transition probability matrix (p_{ij}), where p_{ij} denotes the probability that a job at station i is routed to station j. We now describe the existing methods allowing the analysis of FCQN models.

3 Litterature review

A first survey of FCQN models was made by Perros (1984), who later on also wrote a great 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 are given for the production and manufacturing sector (Papadopoulos and Heavey (1996)), for software architecture performance (Balsamo et al. (2003)), and on retrial queues for the telecommunications sector (Artalejo (1999)).

Exact methods

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. Exact analysis of FCQN models, that is exact evaluation of this joint distribution, can be obtained either in analytic closed form or numerically. For open Jackson networks the joint stationary distribution has a product form, thus the stations are independent. For FCQN the between-station correlation suggests a non-product form stationary distribution, thus exact analysis of FCQN models are limited to very small networks. Closed form analytic expressions for the joint distribution are difficult to obtain and are only available

for specific topologies such as single server two or three station tandem topologies. (Grassman and Derkic (2000), Konheim and Reiser (1978), Konheim and Reiser (1976), Latouche and Neuts (1980), Langaris and Conolly (1984)), or two station 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. A detailed description of these numerical methods can be found in Stewart (1999). 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 stations and their capacity). If the networks of interest have a more general topology or an arbitrary size then their analysis is done by approximation methods, the proposed method falls into this category.

Approximation methods

The most popular approach to evaluate the performance of a finite capacity queueing networks is the use of disaggregate models based on simulation. Simulation models and surveys include (Jun et al. (1999), Fone et al. (2003), Ben-Akiva et al. (1998), Ferrer and Barceló (1994), Messmer and Papageorgiou (1990)). This approach although more realistic and detailed, can be 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, more flexible and more suited for an optimization framework (Cochran and Bharti (2006)). Since the intended use of this model is within an optimization framework we choose an analytical approach.

The main motivation of analytic approximation methods is to reduce the dimensionality of the system under study. Decomposition methods achieve this by decomposing the network into subnetworks and analyzing each subnetwork in isolation. The structural (also called behavioural) parameters of each subnetwork (e.g. average arrival and service rates) depend on the state of other subnetworks and thus acknowledge the correlation with other subnetworks. The main difficulty lies in obtaining good approximations for these parameters so that the stationary distribution of the isolated subnetwork is a good estimate of its marginal stationary distribution.
Given a subnetwork estimates of the marginal distributions and of the main performance measures can be obtained by either establishing a behavioural 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 non-linear. Existing decomposition methods have analysed simple subnetworks consisting of single stations, pairs of stations and triplets. If not stated otherwise the methods concern open finite capacity networks with exponentially distributed service times. The most commonly used decomposition method is single station decomposition, which dates back to the work of Hillier and Boling (1967) who considered tandem single server networks. One of the most used approaches concerns single server feed-forward networks where each station is modelled as an M/M/1 station (Takahashi et al. (1980)). An extension of this method to multiple servers (i.e. M/M/c stations) is given by Koizumi et al. (2005). Here the buffers are considered infinite for each isolated station and their average queue length updates the capacity of the predecessor stations. This approximation holds if the capacity of adjacent predecessor stations can accommodate this average queue length. This constraint is checked only a posteriori. Each station is an M/M/c queue for which closed form expressions of the performance measures exist.
A method applicable to networks with an arbitrary topology is given by Korporaal et al. (2000). The individual stations are modelled as M/M/c/K stations for which closed form performance measures are used. As for the method of Koizumi et al. (2005) the capacity of the stations are revised and the

validity of these capacity adjustments are verified a posteriori.
The Expansion method, (Kerbache and Smith (1987), Kerbache and Smith (1988)), was developed for networks of M/M/1/K stations. Here a network reconfiguration expands all finite capacity stations to artificial infinite capacity holding stations, which register the blocked jobs. This method was later extended to multiple servers and applied to pedestrian traffic flows by Cheah and Smith (1994). Gupta and Kavusturucu (2000) applied this method to production feed-forward systems, where service interruptions are allowed. Singh and Smith (1997) used it to estimate network performance measures within a buffer allocation problem. A similar transformation where all GE/GE/c/K stations are transformed into GE/GE/c stations, and thus the joint distribution is approximated by a product form joint distribution, was proposed by Tahilramani et al. (1999). Single server networks with phase-type service distributions have been proposed for tandem (Altiok (1982)) and feed-forward topologies (Altiok and Perros (1987)), with phase-type service distributions. 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 time consuming task for large networks. In these methods queue capacity is also augmented in order to allow for storage of all predecessor station capacities. Estimates of the marginal distributions are calculated by numerically solving the global balance equations.
Few authors have considered subnetworks larger than single stations. Two-station 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)). van Vuuren et al. (2005) used pairwise decomposition to study multi-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 three-station decomposition method for a single server arbitrary topology network. Subnetworks consisting of more than one station can theoretically provide more accurate results than single station decomposition, but are computationally more intensive (Perros (1994)).

Recent methods, such as Koizumi et al. (2005) and Korporaal et al. (2000), have extended the use of decomposition algorithms mainly to multiple server networks with an arbitrary topology. Nevertheless in order to acknowledge the finite capacity property of these networks the existing methods either revise station capacities or vary the network topologies. The revision of the station capacities renders them dynamic parameters. Additionally approximations need to be used to ensure their integrality and their positivity is only checked a posteriori. We believe that a flexible and optimization-friendly model is one that preserves the network topology and its configuration (number of stations and their capacities) as static parameters. We are also interested in explicitly modelling the blocking phase within our analytical approach, yielding performance measures such as the probability distribution of the number of blocked jobs in a station. Since we have not found methods with these characteristics we have developed the method that we shall now describe.

4 Method

We now describe the decomposition method that allows the analysis of multiple server FCQN with an arbitrary topology and blocking-after-service. The method is based on a decomposition of the network into single stations whose structural parameters are approximated so that they can account for the between-station correlation. The general process that a job goes through upon arrival to a station has been described in section 2. In this context we are interested in explicitly modelling the blocking that a job may go through in a finite capacity network. In this context upon arrival to a station a job:

1. waits if all the servers are occupied
2. is served. This is called the active phase.
3. is blocked if its destination station is full. This is called the blocked phase.
4. leaves the station

We decompose the network into single stations. Let $\pi(i)$ denote the stationary distribution of the isolated station i. $\pi(i)$ can be obtained via the global balance equations along with the use of a normalizing constraint (equations (1)):

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

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

4.1 State space, $\mathcal{S}(i)$

The state of station i is described by the number of active jobs A_i, blocked jobs B_i and waiting jobs W_i:

$$\mathcal{S}(i) = \{(A_i, B_i, W_i) \in \mathbb{N}^3,\ A_i + B_i \leq c_i,\ A_i + B_i + W_i \leq K_i\}$$

Of interest in the validation runs that will be presented in section 5 are bufferless stations, ($K_i = c_i$), where the state space reduces to: $\mathcal{S}(i) = \{(A_i, B_i) \in \mathbb{N}^2,\ A_i + B_i \leq c_i\}$. We denote $card(\mathcal{S}(i))$ the cardinal or dimension of the state space.

4.2 Transition rate matrix, $Q(i)$

$Q(i)$ contains the transition rates between all pairs of states in $\mathcal{S}(i)$. The non diagonal elements, $Q(i)_{sk}$ $s \neq k$, represent the average rate at which the transition between state s and k takes place. The diagonal elements are defined as: $Q(i)_{ss} = -\sum_{k \neq s} Q(i)_{sk}$. 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:

$$\sum_{k \in \mathcal{S}(i)} \pi(i)_k Q(i)_{ks} = -\pi(i)_s Q(i)_{ss}$$

it therefore equates the inflow to the outflow for a given state s. Since we are in the stationary regime this holds for all states.

In this context $Q(i)$ is a function of the following structural parameters:

- the average arrival rate to station i, λ_i.
- the average service rate of a server at station i, μ_i.
- the average unblocking rate at station i given that there are b blocked jobs at station i, $\tilde{\mu}(i, b)$.
- the average probability of being blocked at station i, $\mathcal{P}_i^f$.

Consider a state s such that $(A_i, B_i, W_i) = (a, b, w)$, the possible transitions with their corresponding rates are tabulated 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. We now describe

initial state s	new state k	rate $Q(i)_{sk}$	condition
(a, b, w)	$(a+1, b, w)$	λ_i	$a + b + 1 \leq c_i$
(a, b, w)	$(a, b, w+1)$	λ_i	$a + b == c_i$ & $w + 1 \leq K_i - c_i$
(a, b, w)	$(a-1, b, w)$	$a\mu_i(1 - P_i^f)$	$w == 0$
(a, b, w)	$(a, b, w-1)$	$a\mu_i(1 - P_i^f)$	$w \geq 1$
(a, b, w)	$(a-1, b+1, w)$	$a\mu_i P_i^f$	always possible
(a, b, w)	$(a, b-1, w)$	$\tilde{\mu}(i, b)$	$w == 0$
(a, b, w)	$(a+1, b-1, w-1)$	$\tilde{\mu}(i, b)$	$w \geq 1$

Table 1: Transition rates of station i.

the contents of this table. 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 a service (the active phase) that is not followed by a blocking phase, in the first case the freed server remains available whereas in the second case the freed server immediately starts serving a job that was in the queue. 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 and differ in wether the server that was blocked stays available or immediately starts serving a queued job.

As Korporaal et al. (2000) confirm the main challenge of decomposition methods is to appropriately approximate these structural parameters so that $\pi(i)$ is a good estimate of the marginal stationary distribution of station i. The main complexity lies in appropriately capturing the correlation between the stations via these structural parameters. We now describe how our method revises the structural parameters in order to capture the between-station correlation. Hereafter all rates are average rates.

4.2.1 Arrival pattern

We estimate the arrival rates by combining flow conservation with loss model information. We denote by

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

We model each station as a two-dimensional M/M/c/K station (the distributional assumptions will be detailed further on). For these models, known as loss models, the external arrivals that arise while the station is full are considered to be lost. Thus we have:

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

where N_i denotes the total number of jobs at station i ($N_i = A_i + B_i + W_i$). $P(N_i = K_i)$ 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 station i are assumed to be independent and identically distributed exponential variables with parameter λ_i.

4.2.2 Average probability of being blocked, P_i^f

The average probability of being blocked at station i, P_i^f, is estimated as the average of the blocking probabilities of all downstream stations:

$$P_i^f = \sum_j p_{ij} P(N_j = K_j) \tag{4}$$

4.2.3 Service and unblocking rates, μ_i and $\tilde{\mu}(i, b)$

The average service rate of a server at station i is μ_i. It accounts for the active phase. It is an exogenous parameter thus requires no estimation.
Suppose that station i is in the state $(A_i, B_i, W_i) = (a, b, w)$. Then the service rate of the station 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. We define:

- $\tilde{\mu}_i^o$: the average unblocking rate of a destination station of station i.
- $D(i, b)$: the number of distinct destination stations that are blocking the b jobs at station i.
- $\mathcal{I}^+$: the set of destination stations of station i.
- $card(\mathcal{I}^+)$: the cardinal of $\mathcal{I}^+$

For each destination station that is blocking a job at station i, we approximate the rate at which it unblocks jobs at station i by $\tilde{\mu}_i^o$. Thus if all b jobs are blocked by the same destination station, then they can be seen as forming a virtual queue in front of the blocking station with a FIFO unblocking mechanism. The average unblocking rate at station i is then $\tilde{\mu}_i^o$. If the jobs are blocked by $D(i, b)$ distinct destination stations then they can be seen as forming $D(i, b)$ virtual **parallel** queues, each with a FIFO unblocking mechanism. The average unblocking rate at station i is then $D(i, b)\tilde{\mu}_i^o$. More specifically we have:

$$\frac{1}{\tilde{\mu}(i, b)} = E[\frac{1}{\tilde{\mu}(i, b)} \mid D(i, b)] = \sum_{d=1}^{min(b, card(\mathcal{I}^+))} P(D(i, b) = d)\frac{1}{d\ \tilde{\mu}_i^o} \tag{5}$$

The last equation holds because we assume that each destination station unblocks at rate $\tilde{\mu}_i^o$. We now describe how we estimate both $\tilde{\mu}_i^o$ and $P(D(i, b))$.

The average unblocking rate of a destination station, $\tilde{\mu}_i^o$
We denote by:

- $\hat{\mu}_i$: the effective service rate of a server at station i (it includes service and blocking). We will describe its estimation further on.
- $\tilde{p}_{ij}$: the transition probabilities conditional on a job being blocked at station i: $\tilde{p}_{ij} = \frac{p_{ij} P(N_j = K_j)}{\mathcal{P}_i^f}$
- r_{ij}: the proportion of arrivals to station j that arise from blocked jobs at station i: $r_{ij} = \frac{\tilde{p}_{ij} \lambda_i^{\text{eff}}}{\lambda_j^{\text{eff}}}$

Suppose station j is blocking jobs at predecessor stations. It is therefore full and is serving at rate $\hat{\mu}_j c_j$. It unblocks jobs at station i at the rate $r_{ij} \hat{\mu}_j c_j$. Thus the average time between successive unblockings is:

$$\frac{1}{\tilde{\mu}_i^o} = \sum_j \tilde{p}_{ij} \frac{1}{r_{ij} \hat{\mu}_j c_j}$$

$$\frac{1}{\tilde{\mu}_i^o} = \sum_{j \in \mathcal{I}^+} \frac{\lambda_j^{\text{eff}}}{\lambda_i^{\text{eff}} \hat{\mu}_j c_j} \tag{6}$$

Let us now calculate $P(D(i,b) = d)$

Probability that d distinct stations are blocking the b blocked jobs, $P(D(i,b) = d)$
We denote by:

- $\delta(i,b,d)$: the random vector containing the b destination stations of the blocked jobs, d of which are distinct, i.e. $\delta(i,b,d)_k$ denotes the destination station of the k^{th} blocked job.
- $\Delta(i,b,d)$: the sample space of $\delta(i,b,d)$.
- d: a realization of $\delta(i,b,d)$.

$$\begin{aligned} P(D(i,b) = d) &= \sum_{\mathrm{d} \in \Delta(i,b,d)} P(\delta(i,b,d) = \mathrm{d}) \\ &= \sum_{\mathrm{d} \in \Delta(i,b,d)} P(\delta(i,b,d)_1 = \mathrm{d}_1,\ \delta(i,b,d)_2 = \mathrm{d}_2, ..., \delta(i,b,d)_b = \mathrm{d}_b) \\ &= \sum_{\mathrm{d} \in \Delta(i,b,d)} \tilde{p}_{i\mathrm{d}_1} \tilde{p}_{i\mathrm{d}_2} ... \tilde{p}_{i\mathrm{d}_b} \end{aligned}$$

We define $l(i,b,d)_j$ as the number of jobs blocked by station j at station i (given that there are a total of b blocked jobs that are blocked by d distinct destination stations). We thus have:

$$P(D(i,b) = d) = \sum_{\mathrm{d} \in \Delta(i,b,d)} \prod_{j \in \mathcal{I}^+} \tilde{p}_{ij}^{l(i,b,d)_j}$$

This last equation shows that for a given realization d of $\delta(i,b,d)$, what is of interest in determining $P(D(i,b) = d)$ is the occurrence of each destination station (i.e. the vector $l(i,b,d)$), the ordering of the destination stations is not important. Thus instead of summing over $\Delta(i,b,d)$, we will some over the set of $l(i,b,d)$ vectors. This reduces the size of the space over which we sum. The set of such vectors is noted $L(i,b,d)$ and is defined by:

$$l(i,b,d) \in L(i,b,d) \Leftrightarrow \begin{cases} \sum_{j \in \mathcal{I}^+} l(i,b,d)_j = b \\ \sum_{j \in \mathcal{I}^+} \mathbb{1}(\ l(i,b,d)_j > 0\ ) = d \\ l(i,b,d)_j \geq 0 \quad \forall j \in \mathcal{I}^+ \end{cases} \tag{7}$$

where $\mathbb{1}(x)$ is the indicator function. The first equation of the system of equations 7 means that there are a total of b jobs blocked at station i, and the second means that these jobs are blocked by d different destination stations. For a given vector $l(i,b,d)$ that satisfies the system of equations (7) there are $\frac{b!}{\prod_{j\in\mathcal{I}^+} l(i,b,d)_j!}$ different realizations of $\delta(i,b,d)$ that are associated to it. This corresponds to the number of permutations of a vector of b destination stations, where destination station j is repeated $l(i,b,d)_j$ times. Therefore we obtain:

$$P(D(i,b)=d) = \sum_{\mathrm{d}\in\Delta(i,b,d)} P(\delta(i,b,d)=\mathrm{d}) = \sum_{l(i,b,d)\in L(i,b,d)} \frac{b!}{\prod_{j\in\mathcal{I}^+} l(i,b,d)_j!} \prod_{j\in\mathcal{I}^+} \tilde{p}_{ij}^{\,l(i,b,d)_j}$$

Coming back to equation 5 and using the equation that we have just derived for $P(D(i,b)=d)$ we obtain:

$$\frac{1}{\tilde{\mu}(i,b)} = \sum_{d=1}^{min(b,card(\mathcal{I}^+))} P(D(i,b)=d)\frac{1}{d\ \tilde{\mu}_i^o} = \frac{1}{\tilde{\mu}_i^o} \sum_{d=1}^{min(b,card(\mathcal{I}^+))} \frac{1}{d} \sum_{l(i,b,d)\in L(i,b,d)} \frac{b!}{\prod_{j\in\mathcal{I}^+} l(i,b,d)_j!} \prod_{j\in\mathcal{I}^+} \tilde{p}_{ij}^{\,l(i,b,d)_j} \tag{8}$$

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

$$\tilde{p}_{ij} = \frac{p_{ij}P(N_j=K_j)}{\mathcal{P}_i^f} = \frac{p_{ij}P(N_j=K_j)}{\sum_k p_{ik}P(N_k=K_k)} \approx \frac{p_{ij}}{\sum_k p_{ik}}$$

This approximation makes both summations of equation 8 exogenous and are therefore only evaluated once when solving the entire system of equations. This is a good approximation if the blocking probabilities of the destination stations have the same magnitude, but is poor otherwise. The only endogenous parameter remaining in equation 8 is $\tilde{\mu}_i^o$. Thus we have written $\tilde{\mu}(i,b)$ in the form:

$$\tilde{\mu}(i,b) = \tilde{\mu}_i^o\ \phi(i,b) \tag{9}$$

where $\phi(i,b)$ is estimated exogenously and can be seen as the average number of distinct destination stations that are blocking the b jobs at station i.

When describing the estimation $\tilde{\mu}_i^o$ we came across the effective service rate of a server, $\hat{\mu}_i$. We now describe how we estimate this parameter.

The effective service rate, $\hat{\mu}_i$

The total time spent by a job in front of a server, called the effective service time $\frac{1}{\hat{\mu}_i}$, is composed of the service time (active phase) and for some jobs of the blocked time (blocked phase). We denote by T_i^B the random variable representing the blocked time of a job conditional on it being blocked. For a given station i, all servers serve on average at rate μ_i (active phase). Thus the average time that a job spends in the active phase is $\frac{1}{\mu_i}$. A given job is blocked on average with probability $\mathcal{P}_i^f$ and once he is blocked the average time he spends blocked is $E[T_i^B]$. Accounting for both the service and the possible blocking we obtain the average effective service time $\frac{1}{\hat{\mu}_i}$, which is approximated by:

$$\frac{1}{\hat{\mu}_i} = \frac{1}{\mu_i} + P_i^f E[T_i^B] \tag{10}$$

In this equation μ_i is an exogenous parameter, the estimation of P_i^f was given in equation 4. We can estimte $E[T_i^B]$ by conditioning on the length of the blocked queue.

$$E[T_i^B] = E[E[T_i^B \mid B_i]] = \sum_{b\geq 0} P(B_i = b \mid B_i > 0)\ E[T_i^B \mid B_i = b] = \sum_{b\geq 1} \frac{P(B_i = b)}{P(B_i > 0)}\ E[T_i^B \mid B_i = b]$$

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:

$$E[T_i^B \mid B_i = b] = \frac{1}{b}\sum_{j=1}^{b} E[t(i,b)_j]$$

We know that the average time between successive departures given that there are b blocked jobs at station i is represented by $\frac{1}{\tilde{\mu}(i,b)}$, thus we can approximate the average blocked time of the first job to be unblocked by $\frac{1}{\tilde{\mu}(i,b)}$, that of the second job to be unblocked by $\frac{1}{\tilde{\mu}(i,b)} + \frac{1}{\tilde{\mu}(i,b-1)}$ and that of the j^{th} by:

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

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

$$\begin{aligned} E[T_i^B \mid B_i = b] &= \tfrac{1}{b}\sum_{j=1}^{b}\sum_{k=b-j+1}^{b}\tfrac{1}{\tilde{\mu}(i,k)} \\ &= \tfrac{1}{b}\sum_{k=1}^{b}\tfrac{1}{\tilde{\mu}(i,k)}\sum_{j=b-k+1}^{b} 1 \\ &= \tfrac{1}{b}\sum_{k=1}^{b}\tfrac{k}{\tilde{\mu}(i,k)} \end{aligned}$$

Therefore our estimation of $E[T_i^B]$ is given by:

$$E[T_i^B] = \sum_{b\geq 1} \frac{P(B_i = b)}{P(B_i > 0)} \sum_{k=1}^{b} \frac{k}{b}\frac{1}{\tilde{\mu}(i,k)} \tag{11}$$

Distributional assumptions

Service time and the time between successive unblockings are each assumed to follow an exponential distribution with parameters μ_i and $\tilde{\mu}_i^o$ respectively. For a given station all service times are assumed to be independent and identically distributed, as are all blocked times. By explicitly modelling both of these exponential phases, the number of jobs in front of the servers becomes a two dimensional system (A_i, B_i) composed of the active and the blocked jobs. We are thus in the presence of an M/M/c/K model with a two-dimensional state space. By working in this two-dimensional space we need not construct the CPU intensive phase-type service mechanisms defined in some of the pre-existing methods.

4.3 System of equations

In this method p_{ij}, μ_i, γ_i are exogenous parameters. The set of equations (1-4,6,9-11) are solved simultaneously for all stations in order to obtain the distributions $\pi(i)$, which allow us to derive the performance measures of interest. We use the Matlab solver for systems of nonlinear equations, *fsolve*, which implements a trust-region dogleg algorithm.

5 Validation

We now present validation results by comparing our method to both pre-existing methods and to simulation results on a set of small networks.

5.1 Validation versus pre-exsiting methods

Triangular topology

We first compare our method to that of Altiok and Perros (1987) and that of Takahashi et al. (1980). Takahashi et al. (1980) considered a single server network with triangular topology (depicted in figure 1) and two cases according to the buffer size of the stations: a null buffer and a buffer of size two. For each case they considered a set of scenarios with increasing service rates for stations two and three. These scenarios are displayed in table 2. The chosen performance measure was the blocking probability of station one, $P(N_1 = K_1)$. They then compared their estimates to either simulation results or to 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 figure 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. For both cases we yield similar estimates to those of Takahashi et al. (1980). For the first case Altiok and Perros (1987) yields the most accurate estimates.

$\forall i\ c_i = 1,\ p_{12} = \frac{1}{2}$

$\gamma_1 = 1, \gamma_2 = \gamma_3 = 0$

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

Table 2: Increasing service rate scenarios, corresponding to the triangular topology.

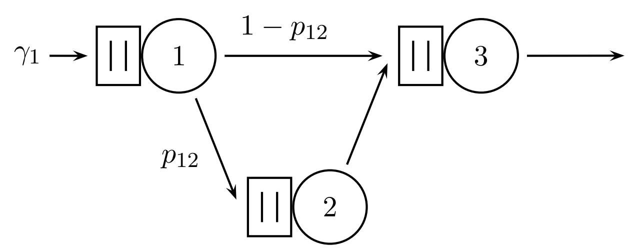


Figure 1: Triangular topology.

Tandem two station topology

Bell (1982) derived a theoretical upper bound on the mean throughput rate of M/M/c/K networks. By considering a tandem two station topology network under a set of scenarios he showed that several decomposition methods "lead to impossible mean throughput rates". We compare the mean throughput estimates of our method with the methods of Hillier and Boling (1967), Takahashi et al. (1980), Boxma and Konheim (1981), Kerbache and Smith (1988) and Singh and Smith (1997). The different scenarios and the topology are displayed in table 3 and the mean throughput estimates of the various methods are depicted in figure 3. Our mean throughput is estimated by using the effective departure rate at station two, λ_2^{eff}. Figure 3 shows that our mean throughput estimate remains near the upper

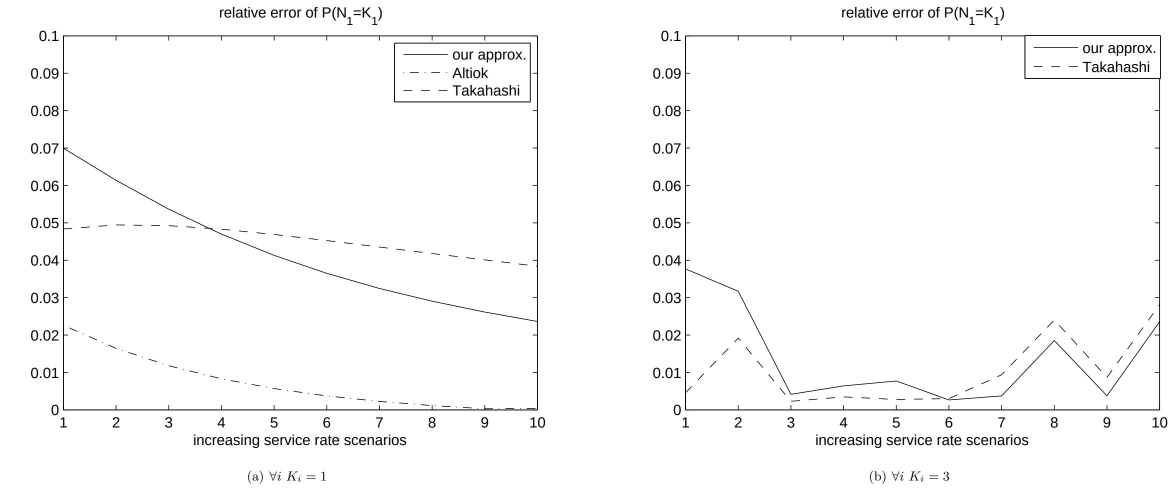


Figure 2: Comparison with the methods of Altiok and Perros (1987) and of Takahashi et al. (1980) under two buffer size configurations.

bound, and is similar to that of the Expansion method of Kerbache and Smith (1988) and of Singh and Smith (1997). It slightly violates the bound for the last three scenarios. The relative violations are: 0.3%, 2.2% and 3.8%. Therefore our method yields consistent throughputs unlike the methods of Takahashi et al. (1980), Hillier and Boling (1967) and Boxma and Konheim (1981).

5.2 Validation versus simulation results

Of main interest in our method are the distributional estimates, which allow us to derive the main performance measures. These could not be compared to pre-existing methods 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 stations, 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, topologies and scenario definitions of networks A, B and C are displayed in tables 4, 5 and 6 respectively. 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, the simplifications concern the number of servers per station and the external arrival rates. We use a discrete event simulator software, 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.

Figure 4 displays a histogram of the errors of the distributional estimates for all states of all scenarios of all three networks. There are a total of 1200 estimates. 70% of the absolute errors are smaller than 0.0065, 80% smaller than 0.0125 and 90% smaller than 0.0241. Our method therefore yields good distributional estimates.

In order to illustrate the blocking information derived by our method we consider the scenarios of network C (Table 6). Figure 5 displays both the estimated and simulated distribution of station five across all five scenarios. The figure shows that as the external arrival rates to increase the states with blocked jobs become more likely, e.g. states (a, b) in $\{(1, 1), (1, 2), (2, 1)\}$. For all states our estimates follow the trend of the simulated probabilities. Overall the estimates are very accurate.

$\mu_1 = 3, \mu_2 = 1,\ c_1 = c_2 = 1$

$\gamma_1 = 1, \gamma_2 = 0$

scenario	$K_1 - c_1$	$K_2 - c_2$
1	1	1
2	1	2
3	2	1
4	2	2
5	2	3
6	3	3
7	4	4
8	5	5
9	10	10

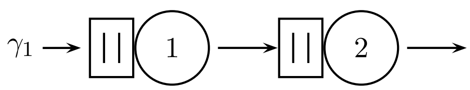


Table 3: Increasing buffer size scenarios that are applied to the tandem two station topology depicted under the table.

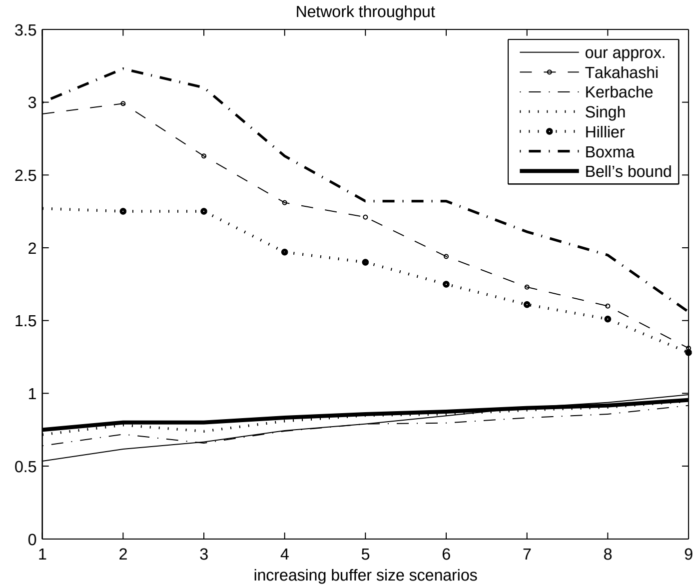
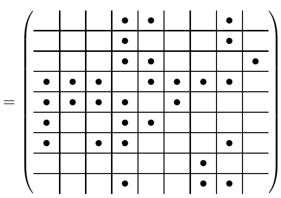


Figure 3: Comparison of the mean throughput estimate of various decomposition methods with the theoretical upper bound derived by Bell (1982).

station index 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

$\forall i \ \ c_i = K_i = 3, \ \ card(\mathcal{S}_i) = 10$

$(p_{ij}) =$



scenario	γ_1
1	0.1
2	0.2
3	0.3
4	0.4

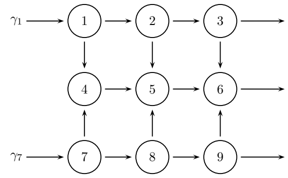
Table 4: **Network A**: configuration, topology and scenario definitions. The possible transitions are depicted as black dots in the transition probability matrix, the full matrix is given in table 7.

6 Case study

We now apply our method to a real scale case study. We consider the patient flow in a network of hospital operating 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, he 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)). Modelling bed blocking and estimating its effects would bring both patient care and budgetary improvements (Cochran and Bharti (2006), Koizumi et al. (2005)). Thus the importance of modelling 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. This is a recent aim defined by Cochran and Bharti (2006): "The next generation of the methodology would include an approximation of the blocking of patients in the queueing model". (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)).

station index 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

$\forall i \;\; c_i = K_i = 3, \;\; card(\mathcal{S}_i) = 10$

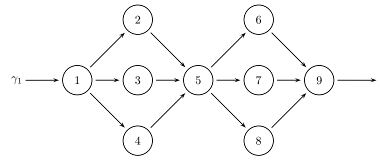


scenario	γ_1	γ_7
1	0.1	0.1
2	0.3	0.3
3	0.5	0.5
4	0.7	0.7
5	0.9	0.9

Table 5: **Network B**: configuration, topology and scenario definitions. For a given station the transition probabilities are uniformly distributed among the possible destinations.

station index 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

$\forall i \;\; c_i = K_i = 3, \;\; card(\mathcal{S}_i) = 10$



scenario	γ_1
1	0.1
2	0.3
3	0.5
4	0.7
5	0.9

Table 6: **Network C**: configuration, topology and scenario definitions. For a given station the transition probabilities are uniformly distributed among the possible destinations.

The hospital of interest is the Geneva University Hospital, HUG. The considered units are the emergency operating suite (denoted BO U), elective operating suite (BO OPERA), otorhinolaryngology operating suite (BO ORL), surgical intensive care (IC CHIR), medical intensive care (IC MED), medical intermediate care (IM MED), neuro-surgical intermediate care (IM NEURO), elective recovery (REV OPERA) and otorhinolaryngology recovery (REV ORL). Here the patients are modelled as jobs. Since there is no waiting space each unit is modelled as a bufferless station. The servers of interest are the beds. The blocking-after-service (BAS) mechanism of our model accurately mimics in-patient bed blocking. HUG members extracted patient flow data which we used to estimate the exogenous parameters of our model. The data consisted of 25246 patient records ranging over a year. The configuration of the network is presented in table 7. The network consists of 9 operating and post-operative units, with 49 possible transitions, containing numerous cycles. This makes the network prone to blocking. We have also carried out this case study using the simulator; we can thus compare our distributional estimates to those obtained via simulation. The simulation setup is the same as that of section 5.2. Figure 6 displays the histogram of the errors of the distributional estimates. The 90^{th}, 95^{th} and 99^{th} percentiles of the absolute errors are 0.008, 0.02 and 0.0733 respectively. We have four estimates that have an absolute error larger than 0.1. Figure 7 displays a more detailed error distribution by omitting the four estimates with absoulte errors beyond 0.1. These figures show that overall the distributional estimates are very good. The cumulative distribution function for the total number of jobs at each station are depicted in figure 8. All stations except stations seven and nine have excellent estimates. Three of the four previously mentioned estimates with large errors concern station seven, the fourth error concerns station nine. The detailed distributions of stations seven and

	BO U	BO OPERA	BO ORL	IC CHIR	IC MED	IM MED	IM NEURO	REV OPERA	REV ORL
c_i	4	8	5	18	18	4	4	10	6
γ_i	0.392	0.502	0.246	0.059	0.176	0.025	0.013	0.155	0
μ_i	0.317	0.255	0.335	0.013	0.015	0.014	0.015	0.22	0.518
$card(\mathcal{S}_i)$	15	45	21	190	190	15	15	66	28
$\forall i \; K_i = c_i$									

$$(p_{ij}) = \begin{pmatrix} 0 & 0 & 0 & 0.16 & 0.02 & 0 & 0 & 0.71 & 0 \\ 0 & 0 & 0 & 0.07 & 0 & 0 & 0 & 0.84 & 0 \\ 0 & 0 & 0 & 0.03 & 0.01 & 0 & 0 & 0 & 0.95 \\ 0.18 & 0.01 & 0.03 & 0 & 0.03 & 0.01 & 0.11 & 0.03 & 0 \\ 0.05 & 0.01 & 0.01 & 0.01 & 0 & 0.07 & 0 & 0 & 0 \\ 0.02 & 0 & 0 & 0.01 & 0.1 & 0 & 0 & 0 & 0 \\ 0.05 & 0 & 0.05 & 0.04 & 0 & 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0 & 0.05 & 0 & 0 & 0.05 & 0.02 & 0 \end{pmatrix}$$

Table 7: Configuration of the Geneva University Hospital network of operative and post-operative rooms.

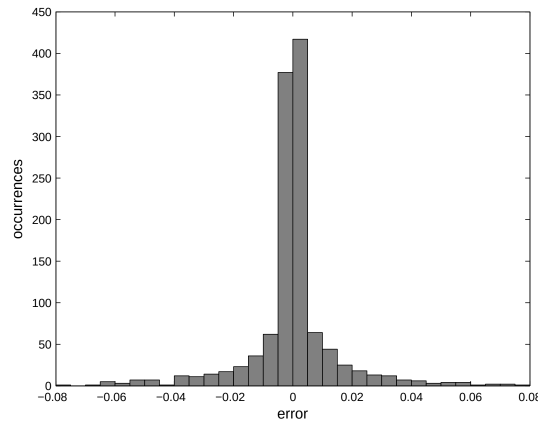


Figure 4: **Histogram of the errors of the distributional estimates.** For networks A, B and C, all scenarios, stations and states we consider: $\hat{\pi}_i(a,b) - \pi_i^*(a,b)$, where $\hat{\pi}$ is our estimate and π^* is the simulation estimate.

nine are displayed in figure 9. Explaining the cause of these large errors is not a straightforward task given that our system of equations consists of a fixed point system, thus parameters are strongly correlated. Nevertheless we believe that the approximation of the transition probabilities conditional on a job being blocked carried out in section 4.2.3 play a role in these erroneous esimations. This is currently being investigated.

7 Conclusion

We have presented a method allowing the analysis of network flows via the use of analytic queueing networks that acknowledge the finite capacity property of the real system. The model is adapted for multiple server finite capacity queueing networks with an arbitrary topology and blocking-after-service. The analysis method is based on a decomposition of the network into single queues whose structural parameters are approximated so that they can account for the between-queue correlation. Unlike pre-existing methods the network topology and its configuration are preserved throughout the analysis thus no constraints need to be checked a posteriori, this renders the method suitable for use within an optimization framework. The originality of this method also lies in its capacity to explicitly model the blocking phase that jobs may go through under congested traffic conditions. Certain performance measures have been validated by comparison to both pre-existing methods and to a theoretical upper bound on the average throughput, on networks with varying buffer size or service rates The distributional estimates have been compared to those obtained via simulation on a set of networks under a set of scenarios with varying arrival rates, namely under high intensity traffic. This has allowed us to validate distributional information concerning blocked jobs, which will be used in the description of congestion effects. In both types of validations the results are very encouraging.

References

Akyildiz, I. F. and von Brand, H. (1994). Exact solutions to networks of queues with blocking-after-service, *Theoret. Comput. Sci.* **125**(1): 111–130.

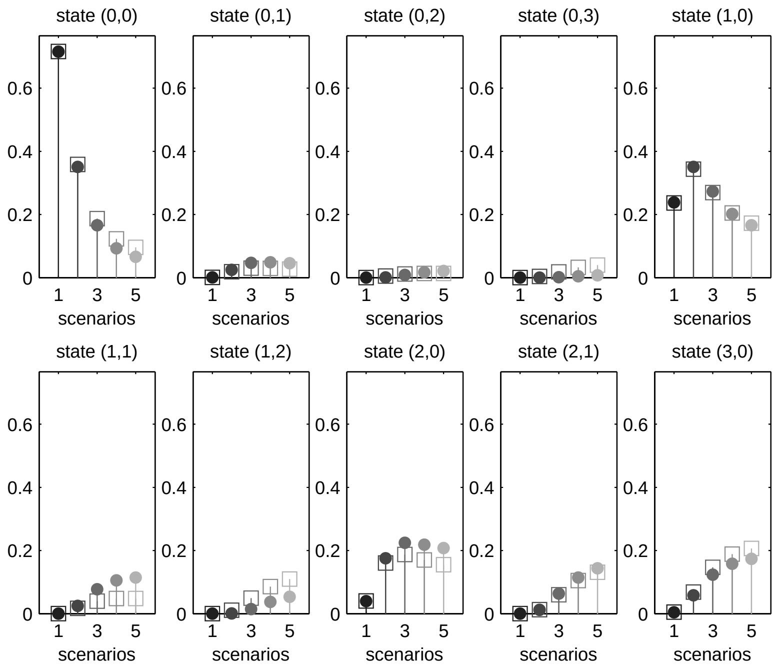


Figure 5: Distribution of station 5 for network C across scenarios 1-5 (these are defined in table 6). Each plot considers a given state $s = (a, b)$, $\pi(5)_s$ is plotted for all scenarios. The scenarios are in a lighter colour as the external arrival rate of station 1 increases. The simulated distribution is depicted as empty squares, whereas our estimates are filled circles.

Alfa, A. S. and Liu, B. (2004). Performance analysis of a mobile communication network: the tandem case, *Comp. Comm.* **27**(3): 208–221.

Altiok, T. (1982). Approximate analysis of exponential tandem queues with blocking, *Europ. J. Operational Res.* **11**(4): 390–398.

Altiok, T. and Perros, H. G. (1987). Approximate analysis of arbitrary configurations of open queuing networks with blocking, *Ann. Oper. Res.* **9**(1): 481–509.

Artalejo, J. R. (1999). Accesible bibliography on retrial queues, *Math. Comput. Modelling* **30**: 1–6.

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

Balsamo, S., De Nitto Persone, V. and Onvural, R. (2001). *Analysis of Queueing Networks with Blocking*, Vol. 31 of *International Series in Operations Res. and Management Sci.*, Kluwer Academic Publishers.

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

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

Ben-Akiva, M., Bierlaire, M., Koutsopoulos, H. and Mishalani, R. (1998). Dynamit: A simulation-based system for traffic prediction, *Paper presented at the DACCORD Short Term Forecasting Workshop, Delft, The Netherlands* .

Boxma, O. J. and Konheim, A. J. (1981). Approximate analysis of exponential queueing systems with

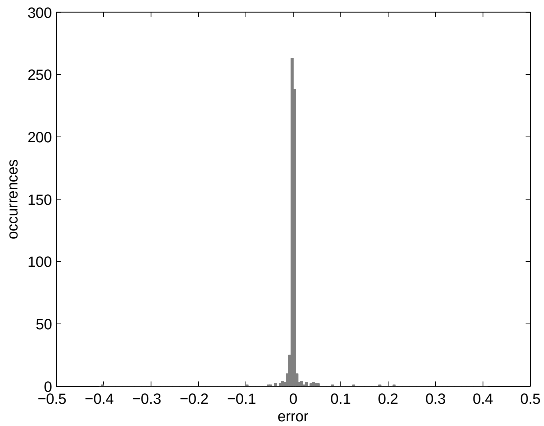


Figure 6: Histogram of errors for the distribution estimates of the HUG network.

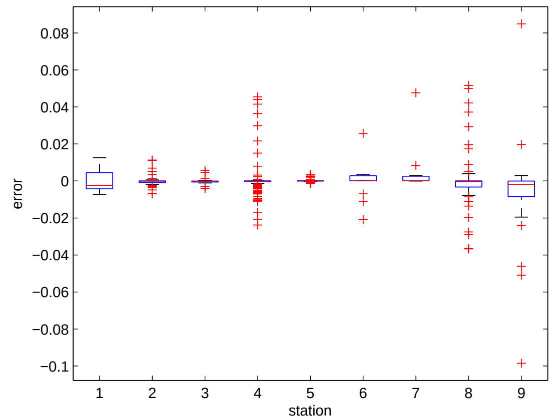


Figure 7: Box plot of the errors for the distribution estimates, omitting the four errors that are larger in magnitude than 0.1

blocking, *Acta Inform.* **15**(1): 19–66.

Brandwajn, A. and Jow, Y. (1985). Tandem exponential queues with finite buffers, *Comp. Networking and Perf. Evaluation* pp. 245–258.

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

Cheah, J. Y. and Smith, M. G. J. (1994). Generalized m/g/c/c state dependent queueing models and pedestrian traffic flows, *Queueing Syst.* **15**: 365–386.

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

Daganzo, C. F. (1996). The nature of freeway gridlock and how to prevent it, *Proceedings of the 13th International Symposium on Transportation and Traffic Theory* pp. 629–646.

Ferrer, J. L. and Barceló, L. (1994). Advanced interactive microscopic traffic simulation for urban and non-urban networks. system description, *Technical report*, Universitat Politcnica de Catalunya. Departamento de EIO. LIOS. Report de Investigacin.

Fone, D., Hollinghurst, S., Temple, M., Round, A., Lester, N., Weightman, A., Roberts, K., Coyle, E., Bevan, G. and 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 Vol. 25, No. 4, pp. 325335* **25**(4): 325–335.

Grassman, W. and Derkic, S. (2000). An analytical solution for a tandem queue with blocking, *Queueing Syst.* **36**: 221–235.

Gupta, S. M. and Kavusturucu, A. (2000). Production systems with interruptions, arbitrary topology and finite buffers, *AnnalsOfOperationsResearch* **93**: 145–176.

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

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

Jackson, J. R. (1957). Networks of waiting lines, *Operations Res.* **5**: 518–521.

Jackson, J. R. (1963). Jobshop-like queuing systems, *Management Sci.* **10**: 131–142.

Jun, J. B., Jacobson, S. H. and Swisher, J. R. (1999). Application of discrete-event simulation in

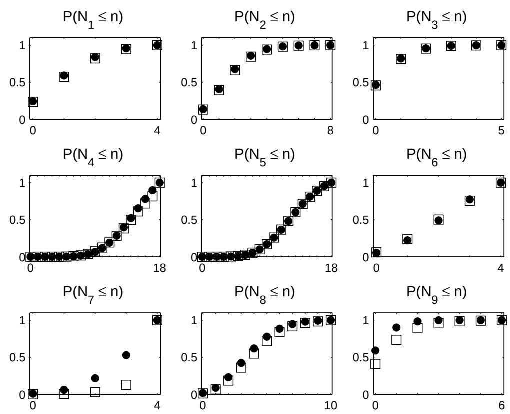


Figure 8: Comparison of the cumulative distribution function, $P(N_i \leq n)$ for all stations. The estimate of our method is represented by a filled circle, whereas the simulation estimates are denoted by empty squares.

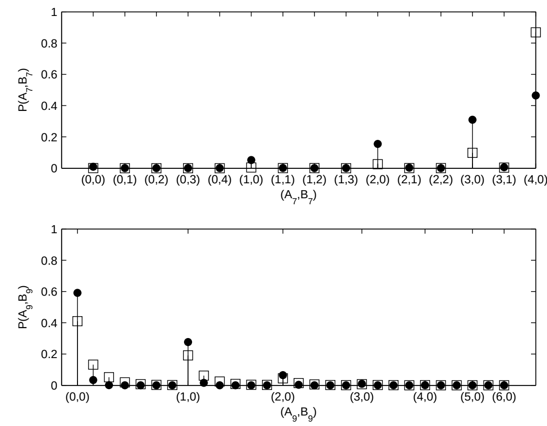


Figure 9: Distributions of stations seven and nine. The estimate of our method is represented by a filled circle, whereas the simulation estimates are denoted by empty squares. The states (a, b) are ordered by increasing number of active jobs and then increasing number of blocked jobs.

health care clinics: A survey, *J. Oper. Res. Soc.* **50**: 109–123.

Jun, K. P. and Perros, H. G. (1988). Approximate analysis of arbitrary configurations of queueing networks with blocking and deadlock, *Queueing Networks with Blocking: Proceedings of the First international workshop* .

Jun, K. P. and Perros, H. G. (1990). An approximate analysis of open tandem queueing networks with blocking and general service times, *Europ. J. Operational Res.* **46**(1): 123–135.

Kerbache, L. and Smith, M. G. J. (1987). The generalized expansion method for open finite queueing networks, *Europ. J. Operational Res.* **32**(3): 448–461.

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

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

Konheim, A. G. and Reiser, M. (1976). Queueing model with finite waiting room and blocking, *J. Assoc. Comput. Mach.* **23**: 328–341.

Konheim, A. G. and Reiser, M. (1978). Finite capacity queueing systems with applications in computer modeling, *SIAM J. Computing* **7**: 210–229.

Korporaal, R., Ridder, A., Kloprogge, P. and Dekker, R. (2000). An analytic model for capacity planning of prisons in the netherlands, *J. Oper. Res. Soc.* **51**(11): 1228–1237.

Langaris, C. and Conolly, B. (1984). On the waiting time of a two-stage queueing system with blocking, *J. Appl. Probab.* **21**(3): 628–638.

Latouche, G. and Neuts, M. F. (1980). Efficient algorithmic solutions to exponential tandem queues with blocking, *SIAM. J. Alg. Disc. Meth* **1**: 93–106.

Lee, H. S., Bouhchouch, A., Dallery, Y. and Frein, Y. (1998). Performance evalutaion of open queueing networks with arbitrary configuration and finite buffers, *Ann. Oper. Res.* **79**: 181–206.

Mackay, M. (2001). Practical experience with bed occupancy management and planning systems: an australian view, *Health Care Management Sci.* **4**: 47–56.

Messmer, A. and Papageorgiou, M. (1990). Metanet: A macroscopic simulation program for motorway networks, *Traffic Eng. and Control* **31**: 466477.

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

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

Perros, H. (1994). *Queueing networks with blocking: Exact and Approximate Solutions*, Oxford Press.

Perros, H. (2003). *Open queueing networks with blocking a personal log. Performance Evaluation: Stories and Perspectives*, Austrian Computer Society.

Schmidt, L. C. and Jackman, J. (2000). Modeling recirculating conveyors with blocking, *Europ. J. Operational Res.* **124**(2): 422–436.

Singh, A. and Smith, M. G. J. (1997). Buffer allocation for an integer nonlinear network design problem, *Comp. and Operations Res.* **24**(5): 453–472.

Stewart, W. J. (1999). *Numerical methods for computing stationary distribution of finite irreducible Markov chains*, Advances in Computational Prob., Kluwer Academic Publishers.

Tahilramani, H., Manjunath, D. and Bose, S. K. (1999). Approximate analysis of open network of GE/GE/m/N queues with transfer blocking, *Seventh IEEE International Symposium on Modeling; Analysis, and Simulation of Computer and Telecommunications Systems (MASCOTS'99)* .

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

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

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