Control and Cybernetics

vol. **43** (2014) No. 1

Open tandem networks with blocking analysis – two approaches[1]

by

Walenty Oniszczuk

Faculty of Computer Science, Bialystok Technical University,
ul. Wiejska 45A, 15-351 Białystok, Poland
w.oniszczuk@pb.edu.pl

Abstract: The paper describes an analytical study of open two-node (tandem) network models with blocking. Here, a specific tandem configuration is chosen: the first node is treated as an infinite server (IS - often referred to as the ample-server), meaning that any incoming task can find at least one empty line for service in this node, and the second node has several parallel lines that can serve input task streams simultaneously. Between these two nodes there is a buffer with finite capacity. In this type of network, if the buffer is full, the accumulation of new tasks by the second node is temporarily suspended (blocking factor) and tasks must wait at the first node until the transmission process is resumed. In this paper, the two-node model is investigated using two different methods. The first is the multi-step exact algorithm, involving a numerical part for solving a set of linear equations, and the second is an approximate algorithm using a product form solution. The numerical part is used for solving a system of linear equations and for calculating the state probability vector. Finally, after comparing both algorithms, some recommendations as to when each method can be used are given.

Keywords: two-node network with blocking, multi-server tandem queues, exact algorithm, product form solution

1. Introduction

In mathematical models of discrete flow systems, which are realistic and effective tools for performance analysis regarding a wide class of systems, such as computer systems and networks, telecommunication networks, transportation networks, production lines, or flexible manufacturing systems, the queuing network models (QNM) with finite capacity queues and blocking are often used (see Balsamo et al., 2003; Badrah et al., 2002; Brandwajn and Jow, 1988; Economou and Fakinos, 1998; Kim et al., 2007; Martin, 2002; Oniszczuk, 2005, 2006, 2009; Sereno, 1999; Zhuang, 1996). Over the years, many publications related to the

[1]Submitted: December 2010; Accepted: March 2014

analysis and application of QNMs with finite capacity queues and blocking in the field of computer science, operations research, traffic engineering or industrial engineering appeared (see Akyildiz, 1988; Balsamo and de Nitto Persone, 1994; Boucherie and van Dijk, 1997; Clo, 1998; Kouvatsos and Almond, 1988; Morrison, 1996; Oniszczuk, 2010; Onvural, 1990; Sharma and Virtamo, 2002).

Most of results of investigations in these areas were selected and ordered in well-known books such as "*Queueing Networks with Blocking. Exact and Approximate Solutions*" (Perros, 1994) and "*Analysis of Queueing Networks with Blocking*" (Balsamo et al., 2001). Similarly, the entire issues of the *Annals of Operation Research*, on *Queueing Networks with Finite Capacity*, Vol. 79 (1998) and *Performance Evaluation*, Vol. 51 (2-4) (2003), were dedicated to queuing networks with blocking, where some sections cover exact analysis, approximate methods and applications. However, there is still great interest in the systems with buffer capacity limitations under different blocking mechanisms (see Amador and Artalejo, 2009; Azadeh et al., 2010; Bose et al., 2006; Bouhchouch et al., 1996; Casale et al., 2008; Gomez-Corral, 2002, 2006; Kwiecień and Filipowicz, 2012; Lenzini et al., 2008; Strelen et al., 1998; van Vuuren et al., 2005). The blocking mechanism restricts the total intensity of input streams by forcing certain limitations on the blocking and synchronization procedures (Kouvatsos et al., 2000; Kwiecień and Filipowicz, 2012; Oniszczuk, 2010; Sharma and Virtamo, 2002; Strelen et al., 1998).

Most studies in the area of two-node (tandem) open networks with blocking (see, e.g. Brandwajn and Jow, 1988; Perros, 1994) assume that each queue is served by a single server, where the first node has an infinite or a finite capacity and the second node has a finite capacity. The state of this queuing network can be described by a pair of variables indicating the number of tasks in the first node and the number of tasks in the second node. Several authors propose approximate methods for single-server networks (or single-server tandems) with blocking, based on the aggregation theorem and on network decomposition by considering various network models and blocking types (see, e.g. Brandwajn and Jow, 1988).

The various closed-form results related to the single-server queuing network include the following two limiting cases: when a task at the first node receives an infinitesimal amount of service, and when the first node is saturated. In the former case, if the task arrives at the tandem point when the second node buffer is not full, the task goes through the first node and it immediately joins the second node. In the case of a saturated first node (the node is never empty), the server is either busy serving or blocked. This case is often described in production systems as a server with an unlimited supply of raw material. In view of this, the second node becomes an *M/M/1* finite waiting capacity queue with an overall arrival rate equal to the first node service rate (single-node approximation). Another special tandem model with blocking assumes that multiple servers serve each queue. In this case, upon completion of service at the first node, a task will get blocked if at that moment the second node is full. A closed-form solution for the queue-length distribution of this model was obtained with

the assumption that the first node is saturated (single-node approximation). This model is equivalent to a queue with state-dependent arrivals. We say that a node is saturated when there is always at least one task waiting for service, i.e. the node is never empty. Another way of studying a tandem configuration is motivated by a kanban scheme, where the first node is assumed to be saturated and it continues to serve tasks during the time when the second node is full (the served tasks remain in the first node). This approach belongs to single-node decompositions. Similarly, other authors studied the tandem configuration with exponential service times and no intermediate buffers, and no queue in front of the first node or where the first node was assumed to have an infinite (or a finite) capacity.

From a practical application point of view, it is very important to investigate a multiple server tandem configuration in two different limiting cases: with heavy traffic (saturation) and with a light tandem load. In the heavy traffic case, the first node is with probability close to 1 never empty (saturation) and it is a "peak" input intensity period or period for serving the tasks accumulated in the first node buffer. In this case it is possible to investigate the maximum possible throughput in the tandem network. When the tandem works on a "light" load, the throughput and hence the network utilization increases as the input stream is increased and tasks with the probability close to 1 find at least one free service line at the first node (conditions similar to an infinite server - IS). This is some kind of simplification for the two-node network functioning when the input stream intensity temporary falls down.

This paper extends the author's previous research on open tandem models with blocking (see Oniszczuk, 2006). The former paper only considered the multiple server two-node queuing networks with blocking separated serving lines, assuming that the first node is under a heavy load. The current article examines an open tandem (two-node approximation) with blocking separated lines at the first node, assuming that the first node works with a light load condition. In both cases, when a departure occurs from the second node, one of the blocked tasks will enter the second node and its associated serving line will become unblocked.

This paper provides the mathematical study of a special type of network configuration (tandem), as shown in Fig. 1. This kind of network has $N+1$ parallel lines at the first node, this being designated as an infinite server (IS), meaning that any input task incoming to this node, can find with probability equal 1 at least one free service line (it is the light load condition where only N serving lines can simultaneously be occupied), and the other node has c parallel servicing lines. Between these nodes there is a common waiting buffer with finite capacity, for example equal m. When the buffer is full, the accumulation of new tasks from the first node is temporarily suspended and the phenomenon called blocking occurs, until the queue empties and allows new inserts. This is the classical mechanism for controlling the intensity of an arriving task stream, which comes to the two-node network. There are also other well-known mechanisms of input process regulation, such as in the systems with truncation, in

which tasks are rejected from service nodes if the waiting buffer is full. In such systems, the rejected tasks, if necessary, are sent back and reprocessed.

In this kind of tandem configuration, no more than $N + m + c$ tasks can be processed simultaneously and the tandem becomes idle, if there are no tasks in both nodes. Assuming that the input stream to the tandem network represents a Poisson process and the service time in both nodes corresponds to a random variable with an exponential distribution, it is a Markovian model of tandem with blocking.

The most common queuing models with blocking assume that the interarrival and service times are exponentially distributed. One of the interesting properties of the exponential distribution is the Markovian or memoryless property, which states that the probability that a job currently in service is completed at some future time t is independent of how long the job has already been in service. It is mainly owing to this special property that the exponential distribution has been the most widely used distribution in the analysis of queuing networks with blocking. In practice, the arrival and service time distributions are not known a priori and they often are not exponential. However, there are also many arrival or service time distributions that fit an exponential. The popularity of the exponential distribution arises out of the fact that it often yields to computationally efficient procedures and obtaining system behavior under this assumption is, in most cases, a relatively easy task.

At the beginning, all possible states of the tandem network are defined, and then the steady state probabilities and the main tandem measures of effectiveness are calculated. Additionally, algorithms for calculation of the blocking probability, delay time in the buffer, blocking time in the node A, the percentage of buffer filling, etc., are shown.

The structure of the paper is as follows. Section 2 specifies the model of the tandem and shows the exact analysis, in Section 3, the product form approximation solution for two-node network with blocking is given. Section 4 describes the procedures for calculating the main measures of effectiveness. Model implementation and numerical examples are described in Section 5. Finally, conclusions are drawn in Section 6.

2. Exact analysis of Markovian tandems with blocking

Let us consider the two-node network with blocking as shown in Fig. 1. The input task stream to the network is assumed to come from a Poisson source with parameter λ to node A and each task is processed on the parallel service lines. Upon service completion at the first node, the tasks are sent to node B. If there are free lines at this node, the service process starts immediately, if not, the tasks must wait in the buffer. If the buffer is full, any task upon service completion at the node A, is forced to wait and blocks this service line.

The general assumptions for this tandem model are:

- the external tasks stream arriving at node A is assumed to be a Poisson stream, with rate $\lambda = 1/a$, where a is the mean inter-arrival time,

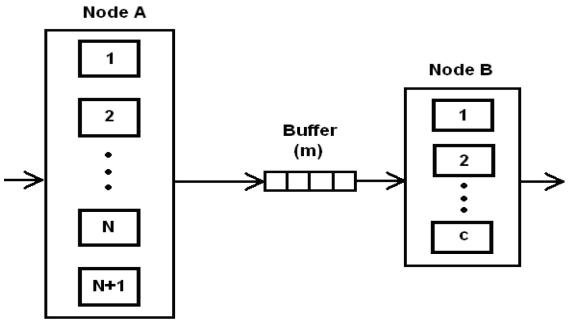

Figure 1. Tandem model with blocking

- node A has $N+1$ parallel service lines, but no more than N tasks can be served simultaneously,
- c is the number of service lines that are available at the node B,
- in both nodes the service time for each task represents an exponentially distributed random variable, with means $s^A = 1/\mu^A$ and $s^B = 1/\mu^B$, where μ is the mean service rate,
- the buffer capacity is finite, for example equal to m.

Under these assumptions, a continuous-time homogeneous Markov chain can represent the tandem network and the model reaches a steady-state condition. It means that the Markov chain has a stationary state distribution.

If there are $(m+c) < N$ conditions, we have a classical tandem with blocking (see Perros, 1994). If the buffer is full, any task upon completion of service at node A, is forced to wait in this service line, because the transfer process from node A depends only on the service process at node B. Physically, blocked tasks stay at node A, but the nature of the service process at node B allows for treating them as located in the additional places in the buffer and belonging to node B. In this case, there can be a maximum of $c + m + N$ tasks assigned to the second node including all tasks in the first node that can be blocked (the maximum number of states in the two-dimensional tandem state space that belong to the second node is equal to $c + m + N$).

In turn, the possible number of non-blocked tasks in the first node is equal to N. According to the initial assumptions, the maximum number of tasks in the two-node network can be $N + m + c$, which means that the current number of tasks that belong to the second node depends on the number of non-blocked tasks in the node A (let it be fixed as i). It means that the current number of possible states at node B (denoted as j) is equal to $j = c + m + N - i$.

This kind of a tandem configuration and state definition can be treated as a Markovian series of service stations with Infinite Server (IS – queue with unlimited service, that is – without a queue in front of the node) at node A

and with buffer enlarged to $m + N$ places in front of node B. If the numbers of tasks located simultaneously at the tandem in the first and second nodes are denoted by i and j, respectively, then a Markov chain with a two-dimensional state space, with unique one path from the state $(0, 0)$ to any state (i, j) and back to the state $(0, 0)$ is defined in this model.

Generally, queuing networks with blocking are difficult to solve, because their steady state probabilities can not be shown to have a product form solution. Hence, most of the techniques that are employed to analyze these networks are in the form of approximations or numerical techniques. Numerical methods are particularly useful in cases for which it is not possible to obtain an analytic solution for the queuing system under study. The queuing system under study is first formulated as a continuous time Markov process with discrete states, and subsequently its steady-state probability vector is calculated using an equation solving technique (e.g. Balsamo et al., 2001; Bolch et al., 1998; Gaver et al., 1984; Oniszczuk, 2006; Stewart, 1994). A queuing network with blocking, under appropriate assumptions, can be formulated as a Markov process and the stationary probability vector can be obtained using numerical methods for a linear system of equations. In general, obtaining the steady state probability vector is a four-step procedure:

1. determine the states and the state space of the network with blocking,
2. enumerate all the transitions that can possibly occur among the states,
3. determine the state transition structure to construct the rate matrix (infinitesimal generator matrix) $\boldsymbol{Q}$,
4. solve the linear system of equations numerically (compute appropriate probability vectors of the Markov chain, from which measures of effectiveness of the queuing network are derived).

It sometimes happens that the infinitesimal generator matrix of a given Markov chain is so highly structured (as in the case of the special type network with blocking) that it is more efficient to write a specific solution procedure for that problem than to use the existing software (for example MARCA, XMARCA or other packages, see Stewart, 1994), this being due to the repetitive nature of the rate matrix (where the resulting rate matrix $\boldsymbol{Q}$ has a block tri-diagonal structure). The process of constructing the full set of the steady-state equations for this special tandem type with blocking requires building the complete two-dimensional state space graph as well as determination of all transition rates from state to state. The framework to describe the state space for open tandem networks as an irregular two-dimensional graph was introduced and formally demonstrated in Oniszczuk (2005).

Figs. 2 and 3 show two parts of this non-trivial state space diagram (in this case the rate matrix is irreducible, aperiodic and has no clearly defined form). The second part of the graph shows the states with blocking and their interpretation. In this case we may denote the state of tandem with blocking by the pair (i, j) where i represents the number of tasks in node A and j denotes the number of tasks in node B. Node B includes the tasks that are both serviced and blocked.

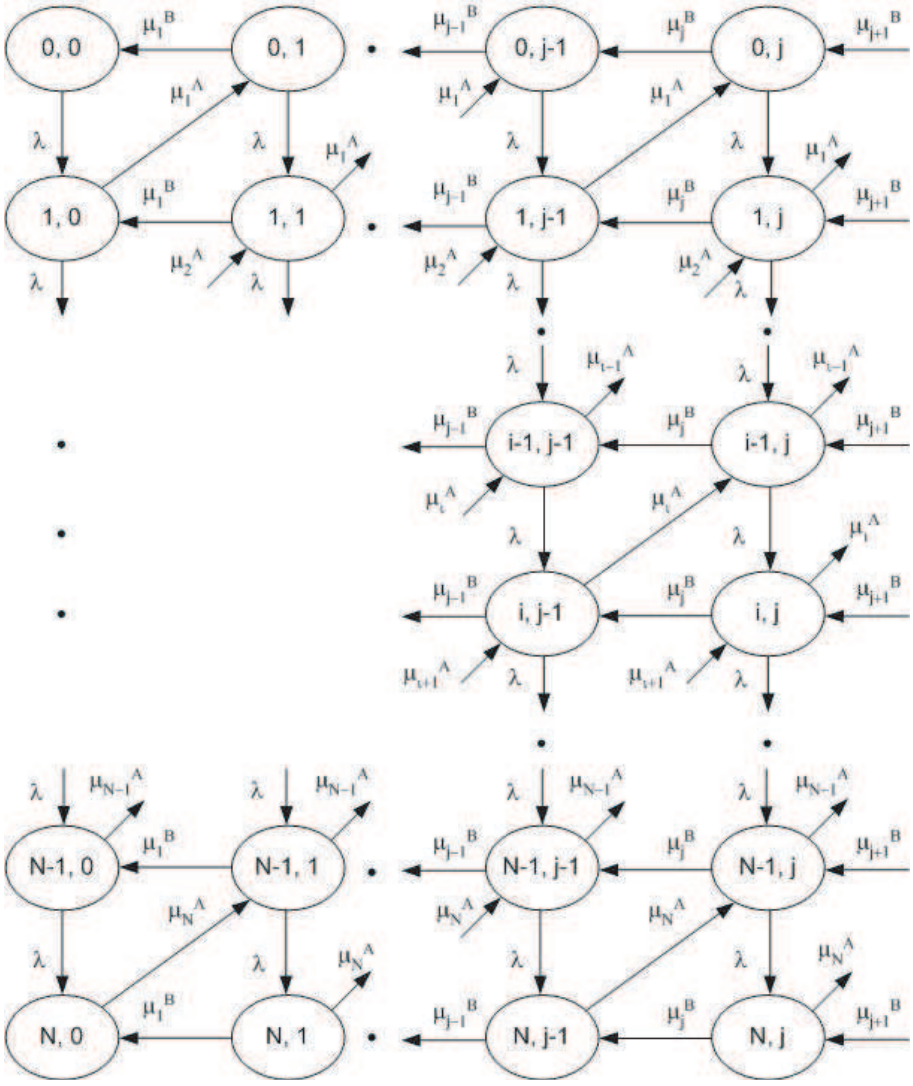

Figure 2. Two-dimensional tandem state space graph (first part)

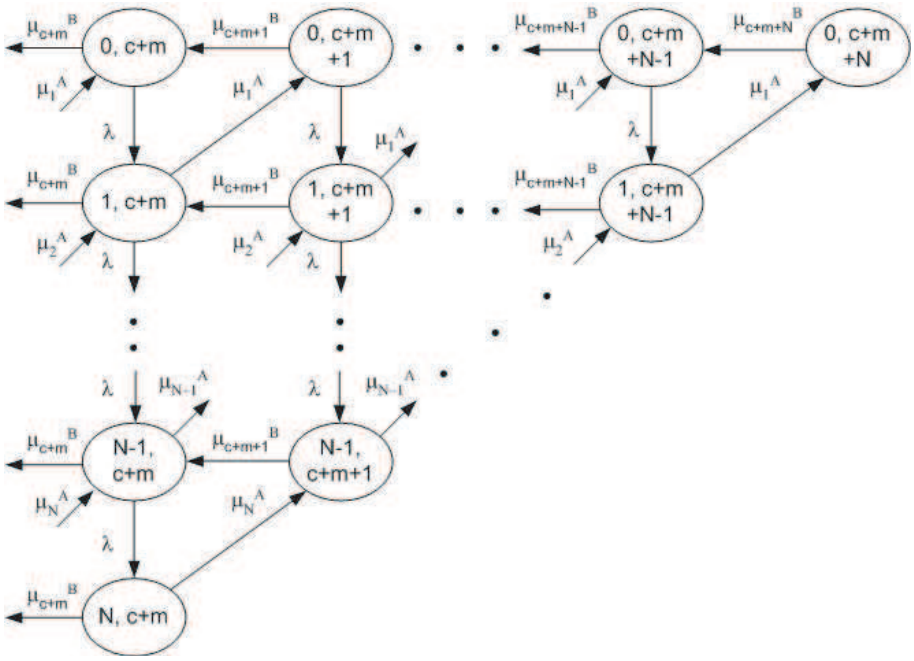

Figure 3. Two-dimensional tandem state space graph (second part)

Based on the analysis of these state space diagrams, the process of constructing the steady-state equations can be divided into several independent steps, which describe several similar, repeatable schema. These steady-state equations are:

$$
\begin{aligned}
0 &= -\lambda \cdot p_{0,0} + \mu_1^B \cdot p_{0,1} \text{ for } i = 0, j = 0 \\
0 &= -(\lambda + \mu_j^B) \cdot p_{0,j} + \mu_1^A \cdot p_{1,j-1} + \mu_{j+1}^B \cdot p_{0,j+1} \text{ for } i = 0, j = 1, ..., c+m \\
0 &= -(\lambda + \mu_i^A) \cdot p_{i,0} + \lambda \cdot p_{i-1,0} + \mu_1^B \cdot p_{i,1} \text{ for } i = 1, ..., N-1, j = 0 \\
0 &= -(\lambda + \mu_j^B + \mu_i^A) \cdot p_{i,j} + \lambda \cdot p_{i-1,j} + \mu_{i+1}^A \dot{p}_{i+1,j-1} + \mu_{j+1}^B \cdot p_{i,j+1} \qquad (1) \\
&\quad \text{for } i = 1, ..., N-1, \quad j = 1, ..., c+m \\
0 &= -\mu_N^A \cdot p_{N,0} + \lambda \cdot p_{N-1,0} + \mu_1^B \cdot p_{N,1} \text{ for } i = N, j = 0 \\
0 &= -(\mu_j^B + \mu_N^A) \cdot p_{N,j} + \lambda \cdot p_{N-1,j} + \mu_{j+1}^B \cdot p_{N,j+1} \text{ for } i = N, \quad j = 1, ..., c+m-1 \\
0 &= -(\mu_{c+m}^B + \mu_N^A) \cdot p_{N,c+m} + \lambda \cdot p_{N-1,c+m} \text{ for } i = N, \quad j = c+m.
\end{aligned}
$$

And for the states with blocking the equations are:

$$
\begin{aligned}
0 &= -(\lambda + \mu_j^B) \cdot p_{0,j} + \mu_1^A \cdot p_{1,j-1} + \mu_{j+1}^B \cdot p_{0,j+1} \\
&\quad \text{for } i = 0, \quad j = c+m+1, ..., c+m+N-1 \\
0 &= -\mu_{c+m+N}^B \cdot p_{0,c+m+N} + \mu_1^A \cdot p_{1,c+m+N-1} \text{ for } i = 0, \quad j = c+m+N \qquad (2) \\
0 &= -(\lambda + \mu_j^B + \mu_i^A) \cdot p_{i,j} + \lambda \cdot p_{i-1,j} + \mu_{i+1}^A \cdot p_{i+1,j-1} + \mu_{j+1}^B \cdot p_{i,j+1} \\
&\quad \text{for } i = 1, ..., N-2, \quad j = c+m+1, ..., c+m+N-1-i \\
0 &= -(\mu_{c+m+N-i}^B + \mu_i^A) \cdot p_{i,j} + \lambda \cdot p_{i-1,j} + \mu_{i+1}^A \cdot p_{i+1,c+m+N-(i+1)} \\
&\quad \text{for } i = 1, ..., N-1, \quad j = c+m+N-i.
\end{aligned}
$$

The process of solving the set of equations given by (1) and (2) with common algorithms, independently of the initial tandem configuration, is not trivial, because a part of the graph has an irregular and triangle shape. There are many methods for the solution of a system of linear algebraic equations but some of these are restricted to certain regular structures of the parameter matrix. In this paper, some methods are proposed, in which the whole graph, column by column, is sequentially re-numbered in order to solve this problem (to get a standard finite-state one-dimensional Markov chain). Additionally, for the triangle part of the graph a special re-numbering method with reduction of column size is applied. This operation is necessary for solving a set of linear equations in MATLAB based on the well-known MATLAB efficient sparse storage schemas and efficient sparsity-preserving algorithms. The state of any Markov chain may be represented as an integer-valued row vector, and this is the means of representation adopted by the above mentioned MATLAB algorithms.

Let some supporting parameter be defined as $k = N + 1$. For the first part of the graph (the states without blocking), the re-numbering algorithm is realized according to the following scheme (*state description* $\rightarrow$ *state number*):

$(0,0)\to 0\cdot k+0$	$(0,1)\to 1\cdot k+0$	...	$(0,c)\to c\cdot k+0$	...	$(0,c+m)\to (c+m)\cdot k$
$(1,0)\to 0\cdot k+1$	$(1,1)\to 1\cdot k+1$	...	$(1,c)\to c\cdot k+1$	...	$(1,c+m)\to (c+m)\cdot k+1$
...	...	...	...	...	...
$(N,0)\to k-1$	$(N,1)\to 2\cdot k-1$	...	$(N,c)\to (c+1)\cdot k-1$	...	$(N,c+m)\to (c+m+1)\cdot k-1$

In turn, for the second part of the state graph with blocking, the column with number $c+m$ is denoted as:

$b_j=(c+m)\cdot k$ for $j=c+m$

and the subsequent column numbers are calculated using the following expression:

$b_j=b_{j-1}+(N+2)-(j-c-m)$ for $j=c+m+1,...,c+m+N$.

In this case, the re-numbering algorithm is (*state description* $\to$ *state number*):

$(0,c+m+1)\to b_{c+m+1}+0$	$(0,c+m+2)\to b_{c+m+2}+0$	...	$(0,c+m+N-1)\to b_{c+m+N-1}+0$	$(0,c+m+N)\to b_{c+m+N}+0$
$(1,c+m+1)\to b_{c+m+1}+1$	$(1,c+m+2)\to b_{c+m+2}+1$	...	$(1,c+m+N-1)\to b_{c+m+N-1}+1$	
...	...	...		
$(N-2,c+m+1)\to b_{c+m+1}+N-2$	$(N-2,c+m+2)\to b_{c+m+2}+N-2$			
$(N-1,c+m+1)\to b_{c+m+1}+N-1$				

The re-numbering algorithm, presented above, allows for transforming the set of equations from (1) and (2) to the following kind of set:

$$0=-\lambda\cdot q_0+\mu_1^B\cdot q_k \quad \text{for } i=0,\quad j=0$$
$$0=-(\lambda+\mu_j^B)\cdot q_{j\cdot k}+\mu_1^A\cdot q_{(j-1).k+1}+\mu_{j+1}^B\cdot q_{(j+1)\cdot k} \text{ for } i=0,\quad j=1,...,c+m$$
$$0=-(\lambda+\mu_i^A)\cdot q_i+\lambda\cdot q_{i-1}+\mu_1^B\cdot q_{k+i} \text{ for } i=1,...,N-1, j=0$$
$$0=-(\lambda+\mu_j^B+\mu_i^A)\cdot q_{j\cdot k+i}+\lambda\cdot q_{j.k+i-1}+\mu_{i+1}^A\cdot q_{(j-1)\cdot k+i+1}+\mu_{j+1}^B\cdot q_{(j+1)\cdot k+i}$$
$$\text{for } i=1,...,N-1, j=1,...,c+m$$
$$0=-\mu_N^A\cdot q_{k-1}+\lambda\cdot q_{k-2}+\mu_1^B\cdot q_{2\cdot k-1} \text{ for } i=N, j=0$$
$$0=-(\mu_j^B+\mu_N^A)\cdot q_{(j+1)\cdot k-1}+\lambda\cdot q_{(j+1)\cdot k-2}+\mu_{j+1}^B\cdot q_{(j+2)\cdot k-1}$$
$$\text{for } i=N, j=1,...,c+m-1$$
$$0=-(\mu_{c+m}^B+\mu_N^A)\cdot q_{(c+m+1)\cdot k-1}+\lambda\cdot q_{(c+m+1)\cdot k-2} \text{ for } i=N, j=c+m \quad (3)$$

And for the states with blocking:

$$0=-(\lambda+\mu_j^B)\cdot q_{b(j)}+\mu_1^A\cdot q_{b(j-1)+1}+\mu_{j+1}^B\cdot q_{b(j+1)}$$
$$\text{for } i=0, j=c+m+1,...,c+m+N-1$$

$$0 = -\mu^B_{c+m+N} \cdot q_{b(c+m+N)} + \mu^A_1 \cdot q_{b(c+m+N-1)+1} \text{ for } i = 0,\ j = c+m+N$$
$$0 = -(\lambda + \mu^B_j + \mu^A_i) \cdot q_{b(j)+i} + \lambda \cdot q_{b(j)+i-1} + \mu^A_{i+1} \cdot q_{b(j-1)+i+1} + \mu^B_{j+1} \cdot q_{b(j+1)+i}$$
$$\text{for } i = 1, ..., N-2,\ j = c+m+1, ..., c+m+N-1-i$$
$$0 = -(\mu^B_{c+m+N-i} + \mu^A_i) \cdot q_{b(j)+i} + \lambda \cdot q_{b(j)+i-1} + \mu^A_{i+1} \cdot q_{b(j-1)+i+1}$$
$$\text{for } i = 1, ..., N-1,\ j = c+m+N-i.$$

This set of linear equations can be solved using classical numerical methods, based on algorithms typical for sparse and diagonal matrices (for example – numerical experiments in MATLAB using efficient sparse storage schemas and efficient sparsity-preserving algorithms). The generation of the rate matrix $\boldsymbol{Q}$ can now be accomplished by going through the list of states and generating all the feasible transitions out of each state and the associated rate of transition. For this kind of Markov process in a steady state, we simply have (see, e.g. Balsamo et al., 2001; Bolch et al., 1998; Gaver et al., 1984; Stewart, 1994):

$$\boldsymbol{xQ} = \boldsymbol{0} \tag{4}$$

where $\boldsymbol{x}$ is the stationary probability vector whose $k-$th element $x_k$is the steady-state probability that the system is in state k. Vector $\boldsymbol{x}$ can be obtained from (4) and the normalizing condition $\sum_{all\ states} \quad x_k = 1$, using equation-solving techniques.

In the next step, the calculated state probabilities are assigned to each state shown on the two-dimensional tandem state graph.

3. An approximate method for tandems with blocking (product form solution)

According to the definition from Section 2, the tandem model represents a Markov chain, therefore based on Jackson's theorem ("if in an open network ergodicity holds for all nodes, then the steady-state probability of the network can be expressed as the product form of the state probabilities of the individual nodes", see Bolch et al., 1998) for queuing networks with exponential service times and a Poisson arrival distribution, the joint probabilities of the tandem states are the product of the marginal probabilities for each node:

$$p_{i,j} = p^A_i \cdot p^B_j. \tag{5}$$

In an open series queue (e.g. tandem) with blocking, each particular node (station) can be described as an independent $M/M/c/L$ finite capacity system, where the maximum number of tasks in the system is L. An arriving task (job) enters the queue if it finds fewer than L tasks in the system. This behavior can be modeled by a birth-death process with:

$$\lambda_k = \begin{cases} \lambda,\ 0 \leq k < L \\ 0,\ k \geqslant L \end{cases} \tag{6}$$

and

$$\mu_k = \begin{cases} k \cdot \mu, \ 0 \leqslant k \leqslant c \\ c \cdot \mu, \ c < k \leqslant L \end{cases} \tag{7}$$

In a classical single-node Markovian queuing model with finite capacity, the steady-state probability of k tasks in the system is given by:

$$p_k = \begin{cases} \frac{\lambda^k}{\mu \cdot 2\mu \cdot 3\mu \cdots k\mu} \cdot p_0 = \frac{\lambda^k}{k! \cdot \mu^k} \cdot p_0, \quad 0 \leqslant k \leqslant c \\ \frac{\lambda^k}{\underbrace{\mu \cdot 2\mu \cdots c\mu}_{c! \cdot \mu^c} \cdot \underbrace{c\mu \cdots c\mu}_{k-c \quad terms}} \cdot p_0 = \frac{\lambda^k}{c! \cdot \mu^k \cdot c^{k-c}} \cdot p_0 = \frac{c^c \cdot \lambda^k}{c! \cdot c^k \cdot \mu^k} \cdot p_0, \quad c < k \leqslant L \end{cases} \tag{8}$$

The boundary condition $\sum_{k=0}^{L} p_k = 1$ will yield p_0; that is:

$$p_0 = \left[\sum_{k=0}^{c} \frac{1}{k!} \cdot \left(\frac{\lambda}{\mu} \right)^k + \sum_{k+c+1}^{L} \frac{c^c}{c! \cdot c^k} \cdot \left(\frac{\lambda}{\mu} \right)^k \right]^{-1}. \tag{9}$$

In the here investigated two-node network with blocking, the maximum number of tasks L in node A is equal to N and in node B this number is equal to $c + m + N$. Before describing the calculation algorithm for tandem marginal probabilities, we need to define the service rates for node A:

$$\mu_1^A = \mu^A, \mu_2^A = 2 \cdot \mu^A, ..., \mu_i^A = i \cdot \mu^A, ..., \mu_N^A = N \cdot \mu^A \tag{10}$$

and node B:

$$\mu_1^B = \mu^B, \mu_2^B = 2 \cdot \mu^B, ..., \mu_c^B = c \cdot \mu^B, ..., \mu_{c+m+N}^B = c \cdot \mu^B. \tag{11}$$

Then, according to formula (8), the marginal steady-state probabilities for each tandem node can be calculated in the following way:

$$p_i^A = \frac{\lambda^i}{i! \cdot (\mu^A)^i} \cdot p_0^A \quad \text{for} \quad i = 0, ..., N \tag{12}$$

$$p_j^B = \frac{\lambda^j}{j! \cdot (\mu^B)^j} \cdot p_0^B \quad \text{for} \quad j = 0, ..., c \tag{13}$$

$$p_j^B = \frac{c^c}{c!} \cdot (\frac{\lambda}{c \cdot \mu^B})^j \cdot p_0^B \quad \text{for} \quad j = c+1, ..., c+m+N. \tag{14}$$

Using these marginal probabilities, all the joint probabilities of the tandem can be calculated as (see formula (5)):

$$p_{i,j} = \frac{\lambda^i \cdot \lambda^j}{i! \cdot j! (\mu^A)^i (\mu^B)^j} \cdot p_{0,0} \ \text{ for } i = 0, ..., N \text{ and } j = 0, ..., c, \tag{15}$$

$$p_{i,j} = \frac{c^c \cdot \lambda^i \cdot \lambda^j}{i! \cdot c!(\mu^A)^i (c \cdot \mu^B)^j} \cdot p_{0,0} \text{ for } i = 0, ...,N \text{ and } j = c+1, ...,c+m+N-i$$

$$\text{also} \quad i + j \leqslant c + m + N. \tag{16}$$

Now we only have to find the value of $p_{0,0}$. This can be accomplished by utilizing the boundary condition:

$$\sum_{i=0}^{N} \sum_{j=0}^{c+m+N-i} p_{i,j} = 1. \tag{17}$$

4. Main measures of effectiveness for a tandem with blocking

The procedures for calculating basic measures of effectiveness use the steady-state probabilities in the following manner:

1. Probability of the tandem blocking p_{bl}:

$$p_{bl} = \sum_{i=0}^{N-1} \sum_{j=c+m+1}^{c+m+N-i} p_{i,j}. \tag{18}$$

2. Idle tandem probability p_{idle}:

$$p_{idle} = p_{0,0}. \tag{19}$$

3. The average number of blocked lines in node A:

$$n_{bl} = \sum_{i=0}^{N-1} \sum_{j=c+m+1}^{c+m+N-i} (j - c - m) \cdot p_{i,j}. \tag{20}$$

4. The average number of active (non-blocked) tasks in node A:

$$lA = \sum_{i=1}^{N} \sum_{j=0}^{c+m+N-i} i \cdot p_{i,j}. \tag{21}$$

5. The average number of tasks in the buffer v:

$$v = \sum_{i=0}^{N} \sum_{j=c+1}^{c+m} (j - c) \cdot p_{i,j} + m \cdot \sum_{i=0}^{N-1} \sum_{j=c+m+1}^{c+m+N-i} p_{i,j}. \tag{22}$$

6. The average number of tasks in node B (buffer + node) n:

$$n = \sum_{i=0}^{N} \sum_{j=1}^{c+m} j \cdot p_{i,j} + (m+c) \cdot \sum_{i=0}^{N-1} \sum_{j=c+m+1}^{c+m+N-i} p_{i,j}. \tag{23}$$

7. The average number of tasks on the service lines in node B:

$$lB = \sum_{i=0}^{N} \sum_{j=1}^{c} j \cdot p_{i,j} + c \cdot \sum_{i=0}^{N-1} \sum_{j=c+1}^{c+m+N-i} p_{i,j}. \tag{24}$$

8. The mean blocking time in node A:

$$t_{bl} = \frac{n_{bl}}{c \cdot \mu^B}. \tag{25}$$

9. The mean waiting time in the buffer:

$$w = \frac{v}{c \cdot \mu^B} \tag{26}$$

10. The mean response time in node B:

$$q = w + \frac{1}{\mu^B}. \tag{27}$$

11. The tandem throughput time:

$$t_{thr} = \frac{1}{\mu^A} + t_{bl} + q. \tag{28}$$

12. The tandem throughput parameter:

$$thr = \frac{N}{t_{thr}}. \tag{29}$$

5. Numerical examples

In this section, we describe the tests of these two approaches on a number of examples. The product form algorithm is simpler than the exact solution based on numerical techniques. Hence, there arises the question of accuracy of the here presented product form solution and the possibility of using the product form algorithms for the investigation of networks with blocking. The results obtained indicate that, in terms of the expected queue lengths, blocking probabilities,

waiting times in the buffer, throughput times and etc., the accuracy of the product form approximation is generally good except for one specific case. To illustrate this fact, in this section the results obtained for four examples were chosen so as to cover a reasonable selection of parameters. Additionally, the investigations will answer the question as to in what cases the product form algorithm can be applied.

First, we consider the tandem with the following configuration: $N = 2$, $c = 2$, $m = 2$, with the inter-arrival and service rates equal to: $\lambda = 4.5$, $\mu^A = 5.5$, $\mu^B = 1.8$. This model has 18 states, but only three of them are with blocking: states (0,5), (0,6), and (1,5). Tables 1 and 2 show the results for the first example.

Table 1. Tandem state probabilities, obtained with the product form solution – *pf*, and the exact solution – *ex*

Type	*State probabilities*								
	(0,0)	(0,1)	(0,2)	(0,3)	(0,4)	(0,5)	(0,6)	(1,0)	(1,1)
pf	0.0193	0.0482	0.0602	0.0752	0.0941	0.1176	0.1470	0.0158	0.0394
ex	0.0238	0.0594	0.0729	0.0822	0.0890	0.0984	0.1143	0.0203	0.0535
% error	-23.32	-23.24	-21.10	-9.31	5.42	16.33	22.24	-28.48	-35.79

Type	*State probabilities*								
	(1,2)	(1,3)	(1,4)	(1,5)	(2,0)	(2,1)	(2,2)	(2,3)	(2,4)
pf	0.0493	0.0616	0.0770	0.0962	0.0088	0.0184	0.0201	0.0230	0.0292
ex	0.0628	0.0666	0.0701	0.0748	0.0126	0.0261	0.0257	0.0259	0.0216
% error	-27.38	-8.12	8.96	22.25	-43.18	-41.85	-27.86	-12.61	26.03

We observe that relative errors for the joint probabilities are off by up to 44%, while the average numbers of blocked lines in node A reach 21%.

We selected our next example to illustrate the performance of our two methods with a large buffer. The model has the following configuration: $N = 5$, $c = 2$, $m = 10$, with the inter-arrival and service rates equal to: $\lambda = 2.3$, $\mu^A = 1.3$, $\mu^B = 0.9$. This model has 93 states, including 15 states with blocking. Showing all the state probabilities for this model has no sense, because most of them have very small values. Thus, only those with the greatest values were chosen for presentation. The results obtained for this experiment are presented in Tables 3 and 4.

Table 2. Main measures of tandem effectiveness, obtained with the product form solution – *pf*, and the exact solution – *ex*

Type	*Main measures of effectiveness*										
	p_{bl}	n_{bl}	v	lA	lB	n	q	w	t_{bl}	t_{thr}	thr
pf	0.3607	0.508	1.288	0.538	1.813	3.102	0.913	0.358	0.141	1.236	1.618
ex	0.2875	0.402	1.111	0.572	1.748	2.859	0.864	0.309	0.112	1.158	1.727
% error	20.29	20.87	13.74	-6.32	3.59	7.83	5.37	13.69	20.57	6.31	-6.74

Table 3. Selected tandem state effectiveness, obtained with the product form solution – *pf*, and the exact solution – *ex*

Type	*State probabilities*							
	(0,10)	(0,11)	(0,12)	(0,13)	(0,14)	(0,15)	(0,16)	(0,17)
pf	0.0100	0.0128	0.0164	0.0209	0.0267	0.0341	0.0436	0.0557
ex	0.0102	0.0129	0.0163	0.0206	0.0261	0.0332	0.0422	0.0538
% error	-2.00	-0.78	0.61	1.44	2.25	2.64	3.21	3.41

Type	*State probabilities (continuation)*							
	(2,8)	(2,9)	(2,10)	(2,11)	(2,12)	(2,13)	(2,14)	(2,15)
pf	0.0097	0.0123	0.0157	0.0200	0.0256	0.0326	0.0418	0.0534
ex	0.0101	0.0128	0.0161	0.0203	0.0256	0.0322	0.0407	0.0515
% error	-4.12	-4.07	-2.55	-1.50	0.00	1.23	2.63	3.63

Table 4. Main measures of tandem effectiveness, obtained with the product form solution – *pf*, and the exact solution – *ex*

Type	*Main measures of effectiveness*										
	p_{bl}	n_{bl}	v	lA	lB	n	q	w	t_{bl}	t_{thr}	thr
pf	0.5833	1.605	8.767	1.363	1.989	10.756	5.982	4.871	0.892	7.643	0.654
ex	0.5665	1.555	8.670	1.376	1.987	10.657	1.633	4.817	0.864	7.561	0.661
% error	2.88	3.12	1.11	-0.95	0.10	0.92	0.90	1.11	3.14	1.07	-1.07

Here, the relative errors for the joint probabilities and for the main measures of effectiveness are below 5%. Note that a large buffer does not cause poor approximation.

The third investigated tandem configuration has the following parameters: $N = 10$, $c = 3$, $m = 4$, with the inter-arrival and service rates equal to: $\lambda = 8.0$, $\mu^A = 4.5$, $\mu^B = 3.3$. This model with a large number of servers at the first node has 143 states, 88 states are without blocking and 55 states include blocking. Table 5 shows the results obtained for this example. We observe that the relative errors for the main measures of effectiveness are close to or equal 0%.

Table 5. Main measures of tandem effectiveness, obtained with the product form solution – *pf*, and the exact solution – *ex*

Type	*Main measures of effectiveness*										
	p_{bl}	n_{bl}	v	lA	lB	n	q	w	t_{bl}	t_{thr}	thr
pf	0.1943	0.677	1.492	1.760	2.399	3.891	0.454	0.151	0.068	0.744	13.435
ex	0.1944	0.677	1.492	1.760	2.400	3.891	0.454	0.151	0.068	0.744	13.435
% error	-0.05	0.00	0.00	0.00	-0.04	0.00	0.00	0.00	0.00	0.00	0.00

Additionally, a new set of experiments was conducted with the third tandem configuration, using the exact solution algorithms, for a wide range of tandem utilization. In this case, the input stream intensity changed within the range of 2.0 to 12.0. The results of these investigations are presented in Fig. 4.

The last investigated tandem configuration has the following parameters: $N = 30$, $c = 10$, $m = 10$, with the inter-arrival and service rates equal to: $\lambda = 6.0$, $\mu^A = 0.5$, $\mu^B = 0.65$. This kind of model has 1 116 states, 651 states are without blocking and 465 states include blocking. The results of the study of the last tandem model are presented in Table 6. We observe that relative errors for the main measures of effectiveness are below 1%.

Table 6. Main measures of tandem effectiveness, obtained with the product form solution – *pf*, and the exact solution – *ex*

Type	*Main measures of effectiveness*										
	p_{bl}	n_{bl}	v	lA	lB	n	q	w	t_{bl}	t_{thr}	thr
pf	0.2497	1.859	4.467	11.918	9.168	13.635	2.225	0.687	0.286	4.511	6.650
ex	0.2515	1.877	4.483	11.917	9.171	13.654	2.228	0.690	0.289	4.517	6.642
% error	-0.72	-0.96	-0.36	0.00	-0.04	-0.14	-0.13	-0.44	-1.05	-0.13	0.12

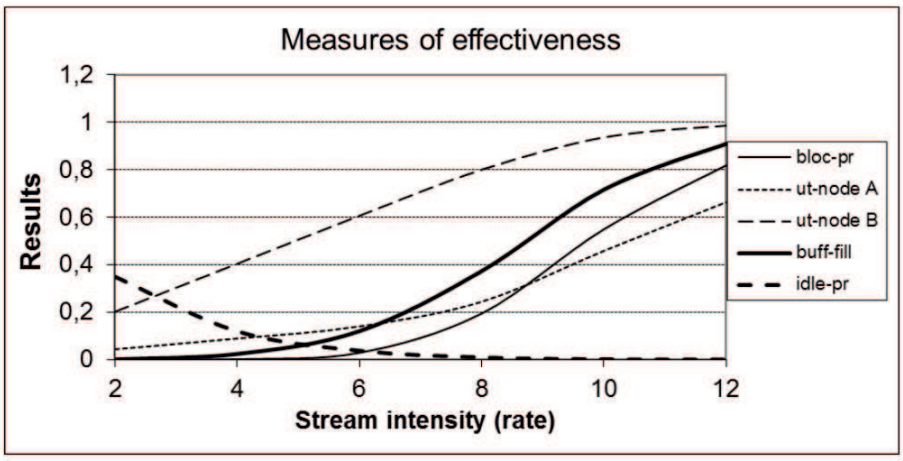

Figure 4. The parameters related to the probability and utilization factors, where *bloc-prob* – node A blocking probability, *ut-node A* – node A utilization factor, *ut-node B* – node B utilization factor, *buff-fill* – the filling buffer coefficient and *idle-pr* – the idle tandem probability

The results of comparison of solutions obtained for these four examples using product form and exact solution show how and where we can use the less complicated algorithm described in Section 3. This product form algorithm is not precise enough for small tandem configurations, but gives results close to the exact algorithm for configurations containing hundreds or thousands of tandem states. In turn, exact analysis of big tandem configurations is more complicated and includes several independent phases – it requires determination of the tandem state space, determination of the state transition structure to construct the rate matrix, as well as solving the linear system of equations using numerical methods. Additionally, we can encounter problems with solving a set of linear equations, because the rate matrix is usually sparse and quite diagonal with a very small intensity (the ratio of the number of non-zero elements to the total number of elements), see Bolch et al. (1998), Stewart (1994). In this case, the simple product form algorithms can be more useful.

6. Conclusions

In this paper, a special configuration of the open two-node tandem network with blocking is investigated. In this model, the first node of the tandem processes the incoming input stream as an infinite server. Assuming Poisson arrivals flow, two different algorithms are proposed. The first algorithm is related to an exact steady-state solution, based on a numerical approach, whereas the second one is based on the product-form solution. Derived from a theoretical analysis, the efficient numerical procedures for calculating the main measures of effectiveness

are proposed. In general, these performance measures are related to the quality of service requirements, such as blocking probability, mean blocking time, mean number of blocked service lines at the first node, etc.

Algorithms presented in this paper can be used as models during the design and modernization of, for example, computer sub-networks, as well as in choosing service strategies, buffer sizes etc.

A comparison of these algorithms shows that the second, approximate, method (the product form solution) is simple and easy to apply, but can only be used for the complex tandem network configuration, which may include hundreds or thousands of states. Its application for smaller configurations may produce serious errors. The first method gives the exact solution, but it requires several independent steps in application. At the beginning, a special state diagram must be constructed, where all intensity rates associated with each node must be determined. Then the set of steady-state equations must be constructed. Finally, the set of linear equations must be solved numerically using suitable methods for sparse and diagonal rate matrices.

Further work has to be done concerning the extension of the analysis to the case where the queuing tandem network with blocking includes multiple classes of jobs, where the set of job classes is partitioned into several disjoint sets, referred to as chains. Each chain is either open or closed. In an open chain, jobs belonging to the chain arrive from outside and depart from the tandem after having been serviced at two nodes. Finally, blocking mechanisms defined for single class open tandem can be extended to multiclass queuing networks. The blocking mechanism can be defined differently for each chain, and for each class of jobs within a chain.

Acknowledgement

The author would like to thank the anonymous referees for their comments, which have helped in improving the presentation and clarity of the paper.

References

AKYILDIZ I. F. (1998) Mean Value Analysis for Blocking Queuing Networks. *IEEE Transaction on Software Engineering* **14** (4), 418-428.

AMADOR J. and ARTALEJO J. R. (2009) Transient analysis of the successful and blocked events in retrial queues. *Telecommunication Systems* **41**, 255-265.

AZADEH A., EBRAHIM R. M. and EIVAZY H. (2010) Parameter optimization of tandem queue systems with finite intermediate buffers via fuzzy simulation. *Performance Evaluation* **67**, 353-360.

BALSAMO S., DE NITTO PERSONE V. (1994) A survey of product form queueing networks with blocking and their equivalences. *Annals of Operations Research* **48** (1/4), 31-61.

BALSAMO S., DE NITTO PERSONE V. and ONVURAL R. (2001) *Analysis of Queueing Networks with Blocking*. Kluwer Academic Publishers, Boston.

BALSAMO S., DE NITTO PERSONE V. and INVERARDI P. (2003) A review on queueing network models with finite capacity queues for software architectures performance predication. *Performance Evaluation* **51** (2-4), 269-288.

BADRAH A., CZACHÓRSKI T., DOMAŃSKA J., FOURNEAU J.-M. and QUESSETTE F. (2002) Performance evaluation of multistage interconnection networks with blocking – discrete and continuous time Markov models. *Archiwum Informatyki Teoretycznej i Stosowanej* **14** (2), 145-162.

BOLCH G., GREINER S., DE MEER H. and TRIVEDI K. S. (1998) *Queueing Networks and Markov Chains. Modeling and Performance Evaluation with Computer Science Applications*. John Wiley, New York.

BOSE A., JIANG X., LIU B. and LI G. (2006) Analysis of manufacturing blocking systems with Network Calculus. *Performance Evaluation* **63**, 1216-1234.

BOUCHERIE R. J. and VAN DIJK N. M. (1997) On the arrival theorem for product form queueing networks with blocking. *Performance Evaluation* **29** (3), 155-176.

BOUHCHOUCH A., FREIN Y. and DALLERY Y. (1996) Performance evaluation of closed tandem queueing networks. *Performance Evaluation* **26**, 115-132.

BRANDWAJN A. and JOW Y-L. L. (1988) An approximate method for tandem queues with blocking. *Operations Research* **36** (1), 73-83.

CASALE G., MUNTZ R. R. and SERAZZI G. (2008) Geometric Bounds: A Noniterative Analysis Technique for Closed Queueing Networks. *IEEE Transactions on Computers* **57** (6), 780-794.

CLO M. C. (1998) MVA for product-form cyclic queueing networks with blocking. *Annals of Operations Research* **79**, 83-96.

ECONOMOU A. and FAKINOS D. (1998) Product form stationary distributions for queueing networks with blocking and rerouting. *Queueing Systems* **30** (3/4), 251-260.

GAVER D. P., JACOBS P. A. and LATOUCHE G. (1984) Finite birth-and-death models in randomly changing environments. *Advances in Applied Probability* **16**, 715-731.

GOMEZ-CORRAL A. (2002) A Tandem Queue with Blocking and Markovian Arrival Process. *Queueing Systems* **41**, 343-370.

GOMEZ-CORRAL A. and MARTOS M. E. (2006) Performance of two-stage tandem queues with blocking: The impact of several flows of signals. *Performance Evaluation* **63**, 910-938.

KIM C. S., KLIMENOK V., TSARENKOV G., BREUER L. and DUDIN A. (2007) The BMAP/G/1 $\rightarrow$ $\cdot$/PH/1/M tandem queue with feedback and losses. *Performance Evaluation* **64**, 802-818.

KOUVATSOS D. and ALMOND J. (1988) Maximum entropy two-station cyclic queues with multiple general servers. *Acta Informatica* **26**, 241-267.

KOUVATSOS D., AVAN I., FRETWELL R. and DIMAKOPOULOS G. (2000) A cost-effective approximation for SRD traffic in arbitrary multi-buffered networks. *Computer Networks* **34**, 97-113.

KWIECIEŃ J. and FILIPOWICZ B. (2012) Firefly algorithm in optimization of queueing systems. *Bulletin of the Polish Academy of Sciences: Technical Sciences* **60** (2), 363-368.

LENZINI L., MINGOZZI E. and STEA G. (2008) A methodology for computing end-to-end delay bounds in FIFO-multiplexing tandems. *Performance Evaluation* **65**, 922-943.

MARTIN J. B. (2002) Large Tandem Queueing Networks with Blocking. *Queueing Systems* **41** (1/2), 45-72.

MORRISON J. A. (1996) Blocking probabilities for multiple class batched arrivals to a shared resource. *Performance Evaluation* **25**, 131-150.

ONISZCZUK W. (2005) *Modele, algorytmy kolejkowe i strategie obsługi w sieciach komputerowych* (*Models, queuing algorithms and service strategies in computer networks*; in Polish). Wydawnictwo Politechniki Białostockiej, Białystok.

ONISZCZUK W. (2006) Tandem Models with Blocking in the Computer Subnetworks Performance Analysis. In: K. Saeed et al., eds., *Biometrics, Computer Security Systems and Artificial Intelligence Applications*. Springer Science+Business Media, 259-267.

ONISZCZUK W. (2009) Semi-Markov-based approach for analysis of open tandem networks with blocking and truncation. *International Journal of Applied Mathematics and Computer Science* **19** (1), 151-163.

ONISZCZUK W. (2010) Loss Tandem Networks with Blocking Analysis – A Semi-Markov Approach. *Bulletin of the Polish Academy of Sciences: Technical Sciences* **58** (4), 673-681.

ONVURAL R. (1990) Survey of closed queuing networks with blocking. *Computer Survey* **22** (2), 83-121.

PERROS H. G. (1994) *Queuing Networks with Blocking. Exact and Approximate Solution*. Oxford University Press, New York.

RAMESH S. and PERROS H. G. (2000) A two-level queueing network model with blocking and non-blocking messages. *Annals of Operations Research* **93** (1/4), 357-372.

SERENO M. (1999) Mean value analysis of product form solution queueing networks with repetitive service blocking. *Performance Evaluation* **36-37**, 19-33.

SHARMA V. and VIRTAMO J. T. (2002) A finite buffer queue with priorities. *Performance Evaluation* **47**, 1-22.

STEWART W. J. (1994) *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press, New Jersey.

STRELEN J. CH., BÄRK B., BECKER J. and JONAS V. (1998) Analysis of queueing networks with blocking using a new aggregation technique. *Annals of Operations Research* **79**, 121-142.

TOLIO T. and GERSHWIN S. B. (1998) Throughput estimation in cyclic queueing networks with blocking. *Annals of Operations Research* **79**, 207-229.

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.

ZHUANG L., BUZACOTT J. A. and LIU X-G. (1994) Approximate mean value performance analysis of cyclic queueing networks with production blocking. *Queueing Systems* **16**, 139-165.

ZHUANG L. (1996) Acceptance instant distributions in product-form closed queueing networks with blocking. *Performance Evaluation* **26**, 133-144.