Response time in a tandem queue with blocking, Markovian arrivals and Phase-type services: Extended Version

B. Van Houdt†* Attahiru Sule Alfa‡

†University of Antwerp, Department of Mathematics and Computer Science,

Performance Analysis of Telecommunication Systems Research Group,

Middelheimlaan, 1, B-2020 Antwerp - Belgium,

benny.vanhoudt@ua.ac.be

‡University of Manitoba, Department of Electrical and Computer Engineering,

Winnipeg, Manitoba, Canada R3T 5V6,

alfa@ee.umanitoba.ca

August 20, 2004

Abstract

A novel approach for obtaining the response time in a discrete time tandem queue with blocking, using matrix analytic methods, is presented. The approach sets up the appropriate Markov chain based on the age of the leading customer in the first queue, together with other auxiliary variables. By this approach the response time is obtained with only minor additional effort after obtaining the stationary distribution of the Markov chain. The queue length distributions, if needed, can still be obtained from the stationary distributions. We also study the stability conditions of this system and carry out several numerical examples that give us insight to how the system behaves.

Index Terms: Tandem queue, blocking, Markovian arrivals, phase type services, response time, Matrix Analytic Method.

[*] B. Van Houdt is a post-doctoral Fellow of the FWO-Flanders.

1 Introduction

In most communication networks we are interested in obtaining the response time of jobs as well as the number of jobs in the system. It has been common practice to first obtain the queue length and then use that to obtain the response time. The two part procedure could be cumbersome in many situations, especially when dealing with tandem queues. This is because in order to obtain the queue length we have to set up the associated Markov chain, obtain its stationary vector and then use that in a process that involves a considerable amount of effort to obtain the response time. In this paper we present a different and novel approach in which we set up the Markov chain, based on the age of the leading job in the first queue and other auxiliary variables. The stationary distribution of this Markov chain will lead us to the response time easily and the queue length can also be computed from it using a simple procedure. We believe that the effort required to compute the two quantities is less using this age process approach we are presenting when compared to the traditional approach of using the Markov chain of queue length. More importantly is that since the response time is sometimes the only key measure of interest, our approach is very favorable in such cases.

Tandem queues with blocking have received considerable attention in the queueing, communications and manufacturing literature because of their pervasiveness and significance in real life. A number of survey papers have been published during the last two decades [7, 20, 10]. Some of the earlier works on this include those of Hunt [11] who first studied the blocking effects in a sequence of waiting lines. Later Avi-Itzhak [1] studied the system with arbitrary input and regular service times. There has been a plethora of studies on this subject and an additional literature survey can be found in [21]. A continuous time tandem queue with blocking, Markovian arrivals (MAP) and no intermediate buffer was considered by Gómez-Corral [6]. Phase type service was assumed at the first infinite queue whereas the other queue is assumed to have a general service time. Results for the joint queue length distribution were provided and the stability issues were partially addressed. Gómez-Corral used matrix analytical methods (MAMs), as we do, to obtain his results, the difference in methodology being that we propose a novel approach that keeps track of the age of the leading customer in the first queue as opposed to the number of customers, as this more easily leads to the response time distribution. MAMs have been used on a variety of occasions when studying tandem queues with blocking, going back to Latouche and Neuts [14].

The bulk of the work done in this area of tandem queues focussed on continuous-time models. As pointed out by Daduna [4], the introduction of the Asynchronous Transfer Mode as a multiplexing technique for broadband integrated services digital networks has increased the studies in the areas of discrete time queueing models. Even though carrying out discrete time analysis of tandem queues may be done, to some extent, in a similar manner as their continuous time counterparts their analysis often introduce some additional challenges. Setting up their transition matrices is more complex, and obtaining the response times of such a system after that requires a more considerable effort. Gün and Makowsky [8, 9] considered a discrete time tandem queue with blocking (and failures), Bernoulli arrivals and phase-type services. They used the MAM approach also and assumed both waiting rooms to be finite. Daduna [4] considered the case with an infinite waiting room for the first queue, but restricted himself to Bernoulli service processes. Desert and Daduna [5] focused on discrete time tandem queues with state dependent Bernoulli service rates and a state dependent Bernoulli arrival stream at the first node. They obtained the joint sojourn time distribution for a customer traversing the tandem system under consideration. In our current paper we consider an infinite waiting room in front of the first server and allow Markovian arrivals (D-MAP), enabling us to model arrival processes that have some elements of correlations, which is more common in the telecommunication field where arrivals are usually bursty. Moreover, while Gün and Makowsky focus on the joint queue length distribution, we provide an algorithm for the total response time of a customer and address the stability issues raised by the infinite queue.

Other related works, in the sense that they focus on the response time as opposed to the joint queue length, are those by van der Mei et al. [22] and Knessl and Tier [12], who both studied the first two moments of the response time in an open two-node queueing network with feedback for the case with an exponential processor sharing node and a FIFO node (while the arrivals at the PS node are Poisson). Chao and Pinedo [3] considered the case of two tandem queues with batch Poisson arrivals and no buffer space in the second queue. They allowed the service times to be general and obtained the expected time in system.

We start with a description of the model under consideration in Section 2, while the GI/M/1 type Markov chain constructed to obtain the performance measures is given in Section 3. Afterward, in Section 4 we indicate how to reduce this GI/M/1 type Markov chain to a QBD. An efficient method to compute the response time and the joint queue length distribution from the steady state vector is presented in Section 5, whereas Section 6 addresses the stability issues surrounding our model. We end by demonstrating the strength of our model through a variety of numerical examples.

2 Model Description

Consider two queues in tandem, where the first queue has an infinite waiting line and the second has a finite one with capacity B. Customers arrive (to the first queue) according to a discrete time Markovian arrival process (D-MAP), characterized by the $l \times l$ sub-stochastic matrices D_0 and D_1. The matrix $D = D_0 + D_1$ is the stochastic matrix of the underlying Markov chain that governs the arrival process. The element $(D_k)_{i,j},\ 1 \leq i, j \leq l,\ k = 0, 1$, represents the probability of making a transition from state i to j with k arrivals. Let γ be the stationary distribution associated with D, then the arrival rate is given by $\lambda = \gamma D_1 \mathbf{1}_l$, where $\mathbf{1}_l$ is a $l \times 1$ column vector of ones. For more details on MAP see [2] and [16].

The service required by a customer in the i-th queue is phase type (PH) distributed with matrix representation (m_i, α_i, T_i), for $i = 1, 2$. It is well known that PH distributions are very good for representing most of the types of services encountered in communication systems [13]. The mean service time of a PH is given as $\mu_i^{-1} = \alpha_i (I_{m_i} - T_i)^{-1} \mathbf{1}_{m_i}$, where I_x is an x dimensional identity matrix. The matrix T_i is sub-stochastic and is of order m_i, while t_i is defined as $\mathbf{1}_{m_i} - T_i \mathbf{1}_{m_i}$. The elements $(\alpha_i)_s$ of the stochastic vector α_i represent the probability that a customer starts his service in phase s. Let $r_i^{(k)}$ be the probability that the service time at node i lasts for k or more units of time, then $r_i^{(k)} = \alpha_i T_i^{k-1} \mathbf{1}_{m_i},\ k \geq 1$. Notice, the minimum service time at node i is at least 1 and the probability that the service time equals exactly k is found as $\alpha_i T_i^{k-1} t_i$. For more details on the phase type distribution see [19].

Whenever a customer finishes service in the first queue it advances to the second queue (at no switching cost), unless the waiting line of the second queue is already fully occupied. In this case the customer remains within the service facility of the first queue until there is a service completion in the second queue. Thereby, preventing any other customers waiting in the waiting line of queue 1 from entering the server (meaning, we adopt the *blocking-after-service* mechanism, see [20, Page 6]). Both queues serve their customers in a FCFS order. All events such as arrivals, transfers from a waiting line to the server and service completions are assumed to occur at instants immediately after the discrete time epochs. This implies, amongst others, that the age of a customer in service at some time epoch n is at least 1.

3 The GI/M/1 type Markov Chain

A Markov chain (MC) that allows us to efficiently obtain the response time distribution of an arbitrary customer, is constructed next. The state space of this MC will be subdivided into an infinite number of groups, called levels. Level zero will contain all the states that correspond to a situation in which the first server is idle. Whereas level i, for $i > 0$, reflects the fact that the first server is occupied by a customer of age i (either because he is being served or because he is blocked by the second queue). To be more specific, the states of level 0 are divided into 2 sets:

- $BI = \{j|\ 1 \leq j \leq l\}$: The MC is said to be in state j of the set BI at time n, if both servers are idle and the arrival process is in state j at time n.
- $FI = \{(b, s_2, j)|\ 0 \leq b \leq B, 1 \leq s_2 \leq m_2, 1 \leq j \leq l\}$: If the first server is idle, the second server is occupied by a customer whose service is in phase s_2 and b customers are waiting to be served by the second server, while the state of the D-MAP is j at time n, then the MC is said to be in state (b, s_2, j) of the set FI at time n.

The states of level i, for $i > 0$, are further subdivided into three sets:

- $SI = \{(s_1, j)|\ 1 \leq s_1 \leq m_1, 1 \leq j \leq l\}$: The MC is said to be in state (s_1, j) of the set SI, at time n, in case the second server is idle, an age i customer is served by server 1, the phase of his service equaling s_1, while the arrival process is, at time $n - i + 1$, in state j.
- $BS = \{(b, s_1, s_2, j)|\ 0 \leq b \leq B, 1 \leq s_v \leq m_v, v = 1, 2; 1 \leq j \leq l\}$: The situation in which, at time n, a customer of age i is being served by server 1, there are b customers waiting for service in the 2nd waiting line, the phase of service in the v-th server equals s_v, for $v = 1, 2$, and the state of the D-MAP at time $n - i + 1$ equals j, will correspond to the state (b, s_1, s_2, j) of level i.
- $BL = \{(s_2, j)|\ 1 \leq s_2 \leq m_2, 1 \leq j \leq l\}$: The scenario where, at time n, there is an age i customer blocked in server 1, a customer is served by server 2, whose current phase is s_2, and the state of the D-MAP at time $n - i + 1$ equals j is represented by state (s_2, j) of the set BL.

Let $|S|$ denote the number of elements in a set S. Define d_t and d_b as $|SI| + |BS| + |BL| = (B + 1)m_1m_2l + (m_1 + m_2)l$ and $|BI| + |FI| = (B + 1)m_2l + l$, respectively. As we shall explain later on, the transition matrix P of this MC has the following form:

$$P = \begin{bmatrix} B_1 & B_0 & 0 & 0 & 0 & \dots \\ B_2 & A_1 & A_0 & 0 & 0 & \dots \\ B_3 & A_2 & A_1 & A_0 & 0 & \dots \\ B_4 & A_3 & A_2 & A_1 & A_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{3.1}$$

where the matrices A_k are $d_t \times d_t$ matrices, B_k, for $k \geq 2$, is a $d_t \times d_b$ matrix, B_0 a $d_b \times d_t$ and B_1 a square matrix of dimension d_b. Notice, the matrix A_k represents the transition probabilities of going from level $i > 0$ to level $i - k + 1$, for $k \leq i$, whereas the matrices B_k are related to transitions from and/or to level 0.

Let us now discuss the matrices A_k and B_k in detail, the structure of P will be apparent from this discussion. Assume the MC is in some state of level $i > 0$ at time n, meaning that an age i customer, referred to as customer c, is occupying server 1. In order to get a transition to level $i + 1$, customer c has to remain in server 1. This is because customers are served in a FCFS order and there are no batch arrivals; hence, the age of the very next customer who arrives after c cannot be larger than i at time $n + 1$. There are three scenarios that would cause customer c to remain in server 1: (i) his service (in server 1) did not finish at time n, (ii) his service finished, but he is blocked by the second queue, or (iii) customer c remains blocked. In case (i), the number of customers in the second queue will either remain the same or decrease by one, depending on whether there is a service completion in server 2. In case (ii) and (iii), the number of customers in the 2nd waiting line has to remain equal to B, implying that there can be no service completion. As a result, we find:

$$A_0 = K_0 \otimes I_l, \tag{3.2}$$

where,

$$K_0 = \left[\begin{array}{c|cccccc|c} T_1 & 0 & 0 & \dots & 0 & 0 & 0 \\ \hline T_1 \otimes t_2 & T_1 \otimes T_2 & 0 & \dots & 0 & 0 & 0 \\ 0 & T_1 \otimes t_2\alpha_2 & T_1 \otimes T_2 & \ddots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ddots & T_1 \otimes T_2 & 0 & 0 \\ 0 & 0 & 0 & \dots & T_1 \otimes t_2\alpha_2 & T_1 \otimes T_2 & t_1 \otimes T_2 \\ \hline 0 & 0 & 0 & \dots & 0 & 0 & T_2 \end{array}\right], \tag{3.3}$$

I_k represents the identity matrix of dimension k, $t_i = \mathbf{1}_{m_i} - T_i \mathbf{1}_{m_i}$, for $i = 1, 2$; and $\mathbf{1}_k$ is a $1 \times k$ vector with each entry set to 1. Notice, this matrix does not depend on the age i of customer c.

Next, we consider the transitions from level i to $i - k + 1$, for $i \geq k \geq 1$. Thus, as before we have a customer, called c, occupying server 1 at time n. To get a transition to level $1 \leq i - k + 1 \leq i$, customer c has to leave server 1. Moreover, a new customer, whose age should equal $i - k + 1$ at time $n + 1$, has to enter the server at time n, call him customer c'. Meaning, the interarrival time between c and c' has to equal k. The fact that customer c leaves server 1 implies that the MC cannot make a transition to one of the states $SI \cup BL$ of level $i - k + 1$. Also, given that the waiting line of server 2 was fully occupied at time n, there should have been a service completion in server 2 (otherwise c would become/remain blocked). Finally, if there still was a vacancy in the waiting line of the second queue at time n, the number of waiting customers there either increases by one or remains the same, depending on whether there is a service completion in server 2. Hence, transitions from level i to $i - k + 1$ are governed by the matrix A_k below (and are thus independent of i):

$$A_k = K_1 \otimes D_0^{k-1} D_1, \tag{3.4}$$

where,

$$K_1 = \left[\begin{array}{c|ccccc|c}
0 & t_1\alpha_1 \otimes \alpha_2 & 0 & \ldots & 0 & 0 & 0 \\
\hline
0 & t_1\alpha_1 \otimes t_2\alpha_2 & t_1\alpha_1 \otimes T_2 & \ldots & 0 & 0 & 0 \\
0 & 0 & t_1\alpha_1 \otimes t_2\alpha_2 & \ddots & 0 & 0 & 0 \\
\vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \ddots & t_1\alpha_1 \otimes t_2\alpha_2 & t_1\alpha_1 \otimes T_2 & 0 \\
0 & 0 & 0 & \ldots & 0 & t_1\alpha_1 \otimes t_2\alpha_2 & 0 \\
\hline
0 & 0 & 0 & \ldots & 0 & \alpha_1 \otimes t_2\alpha_2 & 0
\end{array}\right]. \tag{3.5}$$

Let us now consider the possible transitions from level i to level 0. Again, our customer c, who was occupying server 1 at time n, has to leave the server, while, at time $n+1$, server 1 should be idle. This can only be true if there are no arrivals at time $n-i+1$, $n-i+2$, $\ldots$, n. Thus, the interarrival time between customer c and the very next arrival should be at least $i + 1$ time units. The state of the MC at time $n + 1$ cannot be part of the set BI as customer c has to pass through the second queue (and needs at least one time unit to do so). Therefore, level zero has to be entered (from a higher level) through one of the states of FI. The exact state is once more determined by whether or not there was a service completion in server 2. Notice, if the waiting line of queue 2 was fully occupied at time n, a service completion is necessary as costumer c would otherwise be blocked. Transitions from level i into level 0 are characterized by the matrix B_{i+1}, thus

$$B_k = \left[\begin{array}{c|ccccc} 0 & t_1\alpha_2 & 0 & \dots & 0 & 0 \\ \hline 0 & t_1 \otimes t_2\alpha_2 & t_1 \otimes T_2 & \dots & 0 & 0 \\ 0 & 0 & t_1 \otimes t_2\alpha_2 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & t_1 \otimes t_2\alpha_2 & t_1 \otimes T_2 \\ 0 & 0 & 0 & \dots & 0 & t_1 \otimes t_2\alpha_2 \\ \hline 0 & 0 & 0 & \dots & 0 & t_2\alpha_2 \end{array}\right] \otimes D_0^{k-1}, \tag{3.6}$$

for $k \geq 2$.

Finally, consider the transitions from level 0. Due to the definition of the state space, this means the first server is idle. Depending on whether an arrival occurs at time n, the MC will be at level 0 (no arrival) or 1 (arrival) at time $n+1$. If both servers were idle at time n, then obviously this will still hold for server 2 at time $n+1$. This implies that the MC will be in a state of the set SI, resp. BI, if an, resp. no, arrival occurs at time n. If, on the other hand, server 2 was busy (at time n), the number of waiting customers in its waiting line can either decrease by one or remain identical (depending on whether there is a service completion). Finally, if an arrival occurs at time n, then this customer cannot be blocked in server 1 as he requires service first. As a result, the matrices B_0 and B_1 are found as

$$B_0 = \left[\begin{array}{c|ccccc|c} \alpha_1 & 0 & 0 & \dots & 0 & 0 & 0 \\ \hline \alpha_1 \otimes t_2 & \alpha_1 \otimes T_2 & 0 & \dots & 0 & 0 & 0 \\ 0 & \alpha_1 \otimes t_2\alpha_2 & \alpha_1 \otimes T_2 & \ddots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ddots & \alpha_1 \otimes T_2 & 0 & 0 \\ 0 & 0 & 0 & \dots & \alpha_1 \otimes t_2\alpha_2 & \alpha_1 \otimes T_2 & 0 \end{array}\right] \otimes D_1, \tag{3.7}$$

and

$$B_1 = \left[\begin{array}{c|ccccc} 1 & 0 & 0 & \dots & 0 & 0 \\ \hline t_2 & T_2 & 0 & \dots & 0 & 0 \\ 0 & t_2\alpha_2 & T_2 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & T_2 & 0 \\ 0 & 0 & 0 & \dots & t_2\alpha_2 & T_2 \end{array}\right] \otimes D_0, \tag{3.8}$$

respectively.

Let $\pi = [\pi_0, \pi_1, \pi_2, \dots]$ be the steady state vector of P, where π_0 is a $1 \times d_b$ vector and π_i, for $i > 0$, a $1 \times d_t$ vector. Since, the MC characterized by P is an GI/M/1 type MC, π exists if and only if $\theta\beta > 1$, where $\theta \sum_{k \geq 0} A_k = \theta$, $\theta \mathbf{1}_{d_t} = 1$ and $\beta = \sum_{k \geq 1} k A_k \mathbf{1}_{d_t}$. A detailed discussion of the stability condition $\theta\beta > 1$ is given in Section 6. The steady state vector of an GI/M/1 type MC can be found by solving for the minimal non-negative R in the following non-linear equation iteratively:

$$R = \sum_{k \geq 0} R^k A_k. \tag{3.9}$$

Instead we shall construct a Quasi-Birth-Death (QBD) MC, from which we can compute π much more efficiently (both in terms of the time and memory complexity).

4 The QBD Markov Chain

In order to construct the QBD we add $(B+1)m_2 l$ states to each level i, for $i > 0$, to the state space of P. These additional states are referred to as artificial states. The basic idea behind this construction is to replace a transition from level i to $i-k$, for $k \geq 1$, by $k+1$ transitions, where for each of the first k transitions we decrease the level by one, while for the $(k+1)^{st}$ transition the level will remain identical. Thus, instead of making a transition from level i to $i-k$ at once, the new MC will visit k intermediate states, which shall all be artificial states. The fact that it suffices to add $(B+1)m_2 l$ artificial states to obtain a QBD, is caused by: (i) The geometric nature of the matrices A_k and B_k, for $k \geq 2$ (indeed, these matrices only depend upon k through the geometric term D_0^{k-1}). (ii) The states $SI \cup BL$ of level $i-k$, for $k \geq 1$, are never directly visited from level i. (iii) We can postpone the determination of the initial phase of a possible customer entering server 1 until the $k+1$th transition, during which we return to a non-artificial state. Define d_t^* as $d_t + (B+1)m_2 l$.

Let us now introduce the resulting QBD characterized by the transition matrix P^*:

$$P^* = \begin{bmatrix} B_1 & B_0^* & 0 & 0 & \dots \\ B_2^* & A_1^* & A_0^* & 0 & \dots \\ 0 & A_2^* & A_1^* & A_0^* & \dots \\ 0 & 0 & A_2* & A_1^* & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix}, \tag{4.10}$$

where A_k^*, for $k = 0, 1, 2$; are square matrices of dimension d_t^*, B_2^* is a $d_t^* \times d_b$ matrix, B_0^* a $d_b \times d_t^*$ and B_1 is the same matrix as before. The matrices A_k^*, for $k = 0, 1, 2$; B_0^* and B_2^* are then constructed as

$$A_0^* = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & A_0 \end{array}\right], \quad A_1^* = \left[\begin{array}{c|c|c|c} 0 & 0 & I_{B+1} \otimes \alpha_1 \otimes I_{m_2} \otimes D_1 & 0 \\ \hline 0 & \multicolumn{3}{c}{A_1} \end{array}\right],$$
$$A_2^* = \left[\begin{array}{c|c} I_{(B+1)m_2} \otimes D_0 & 0 \\ \hline B_2^+ & 0 \end{array}\right], \quad B_0^* = \left[\begin{array}{c|c} 0 & B_0 \end{array}\right], \quad B_2^* = \left[\begin{array}{c|c} 0 & I_{(B+1)m_2} \otimes D_0 \\ \hline 0 & B_2^+ \end{array}\right] \tag{4.11}$$

where B_2^+ is found by removing the first l (zero) columns of B_2. It is an easy exercise to see that this QBD MC coincides with P when censored on the non-artificial states. Moreover, this MC is ergodic if and only if P is (if π is the steady state vector of P, one can easily construct a steady state vector π^* of P^* and vice versa).

The key in finding the steady state probability vector $\pi^* = (\pi_0^*, \pi_1^*, \dots)$ of P^*, where π_0^* and π_i^*, for $i > 0$, are $1 \times d_b$ and $1 \times d_t^*$ vectors, respectively, is to solve the following equation:

$$G = A_2^* + A_1^* G + A_0^* G^2. \tag{4.12}$$

We propose to use the Cyclic Reduction algorithm to compute G. This algorithm is very easy to implement, requires a low amount of memory, converges quadratically and is numerically stable [17]. The memory requirements of the CR algorithm are about $5(d_t^*)^2$, while $14(d_t^*)^3$ flops are needed for a single iteration. As the CR algorithm converges quadratically, one typically needs less than 25 iterations, even if the arrival process is very correlated and/or the system is close to instability. Having found G, one computes R as $A_0^*(I - A_1^* - A_0^* G)^{-1}$ [15]. The steady state probability vectors π_i^* are then found as:

$$\begin{aligned} [\pi_0^*,\ \pi_1^*] &= [\pi_0^*,\ \pi_1^*] \begin{bmatrix} B_1 & B_0^* \\ B_2^* & A_1^* + RA_2^* \end{bmatrix}, && (4.13) \\ \pi_i^* &= \pi_{i-1}^* R, && (4.14) \end{aligned}$$

where $i > 1$, π_0^* and π_1^* are normalized as $\pi_0^* \mathbf{1}_{d_b} + \pi_1^*(I-R)^{-1}\mathbf{1}_{d_t^*} = 1$. Denote π_i^*, for $i > 0$, as $[\pi_i^*(0),\ \pi_i^*(1)]$, with $\pi_i^*(0)$ and $\pi_i^*(1)$ a $1 \times (B+1)m_2 l$ and $1 \times d_t$ vector, respectively. Then, we have the following relationship between π and π^*:

$$\begin{aligned} \pi_0 &= \pi_0^*/(1-c) \qquad (4.15) \\ \pi_i &= \pi_i^*(1)/(1-c), \qquad (4.16) \end{aligned}$$

for $i > 0$. The constant c equals $\sum_{i>0} \pi_i^*(0) 1_{d_b}$.

5 Performance Measures

In this section we demonstrate how to get the response time distribution from the steady state vector π. We start by introducing the following set of random variables:

- T_{W_i} : The amount of time a tagged customer has to wait in the ith waiting line, for $i = 1, 2$.
- T_{S_i}: The service time duration of a tagged customer in server i, for $i = 1, 2$.
- T_B: The time that elapses while a tagged customer is blocked in server 1.

Having defined these variables the total response time T_R of a tagged customer is defined as $T_{W_1} + T_{S_1} + T_B + T_{W_2} + T_{S_2}$, while the response time in queue 1, denoted as T_{R_1}, equals $T_{W_1} + T_{S_1} + T_B$. We need two more variables before we can proceed:

- F_N: The number of customers still requiring full service by server 2 before a tagged customer who just left server 1 can start his service.
- F_T: The remaining service time of the customer occupying server 2 when a tagged customer leaves server 1.

Write π_i as $[\pi_i^{SI}, \pi_i^{BS}, \pi_i^{BL}]$ in accordance with the three sets of states of level i, then

$$P[T_{R_1} = r, F_N = b, F_T = h] = \sum_{s_1} \frac{(t_1)_{s_1}}{\lambda} \left\{ 1_{\{b=0\&h=0\}} \left(\sum_j \pi_r^{SI}(s_1, j) \right) + 1_{\{b<B|h=0\}} \left(\sum_{s_2,j} \pi_r^{BS}(b, s_1, s_2, j)(T_2^h t_2)_{s_2} \right) \right\} + \frac{1_{\{b=B\&h=0\}}}{\lambda} \left(\sum_{s_2,j} \pi_r^{BL}(s_2, j)(t_2)_{s_2} \right),$$

where $(x)_i$ denotes the ith component of the vector x. From this it is straightforward to compute the probabilities $P[T_{R_1} + F_T = r, F_N = b]$. The total response time distribution is then found by

$$P[T_R = i] = \sum_b \sum_{r \leq i-b} P[T_{R_1} + F_T = r, F_N = b] P[S_2^{(*b+1)} = i - r], \tag{5.17}$$

where $S_2^{(*x)}$ is the x-fold convolution of the PH service time distribution of server 2 characterized by (m_2, α_2, T_2).

The blocking probability p_{BL} can be computed as

$$p_{BL} = \frac{1}{\lambda} \sum_{r>0} \sum_{s_1,s_2,j} \pi_r^{BS}(B, s_1, s_2, j)(t_1)_{s_1}(T_2)_{s_2}, \tag{5.18}$$

and the blocking time distribution as

$$P[T_B = t] = \frac{1}{\lambda} \sum_{r>0} \sum_{s_1,s_2,j} \pi_r^{BS}(B, s_1, s_2, j)(t_1)_{s_1}(T_2^t t_2)_{s_2}, \tag{5.19}$$

for $t > 0$ and $P[T_B = 0] = 1 - p_{BL}$.

In order to compute the joint queue contents distribution (Q_1, Q_2)[1], we first define $h_{k,a}(j)$ as the probability that k arrivals occur in an interval of length a that started in state j. These vectors can be computed by means of the following recursion:

$$\begin{aligned}
h_{k,0}(j) &= 1_{\{k=0\}}, \\
h_{0,a}(j) &= (D_0^a \mathbf{1}_l)_j, \\
h_{k,a}(j) &= \sum_{j'} \left\{ (D_1)_{j,j'} h_{k-1,a-1}(j') + (D_0)_{j,j'} h_{k,a-1}(j') \right\}.
\end{aligned}$$

By means of these probabilities we can easily compute (Q_1, Q_2):

$$\begin{aligned}
P[Q_1 = q_1, Q_2 = q_2] = \sum_{r \geq q_1} \sum_j \Bigg(& 1_{\{q_2 \neq 0\}} \sum_{s_1,s_2} \pi_r^{BS}(q_2 - 1, s_1, s_2, j) \\
& + 1_{\{q_2=0\}} \sum_{s_1} \pi_r^{SI}(s_1, j) + 1_{\{q_2=B+1\}} \sum_{s_2} \pi_r^{BL}(s_2, j) \Bigg) h_{q_1-1,r-1}(j),
\end{aligned} \tag{5.20}$$

for $q_1 > 0$ and

$$P[Q_1 = 0, Q_2 = q_2] = 1_{\{q_2 \neq 0\}} \sum_{s_2,j} \pi_0^{FI}(q_2 - 1, s_2, j) + 1_{\{q_2=0\}} \sum_j \pi_0^{BI}(j). \tag{5.21}$$

[1] The queue contents is equal to the number of customers that are either in the waiting room or in service.

6 Stability Condition

6.1 The stationarity of the GI/M/1 type Markov Chain

The GI/M/1 type Markov chain introduced in Section 3 is ergodic if and only if $\theta\beta > 1$, where $\theta \sum_{k\geq 0} A_k = \theta$, $\theta \mathbf{1}_{d_t} = 1$ and $\beta = \sum_{k\geq 1} k A_k \mathbf{1}_{d_t}$. From Eq. (3.2) and (3.4) one easily finds the following expressions for β

$$\beta = \tau \otimes (I_l - D_0)^{-1} \mathbf{1}_l, \tag{6.22}$$

where,

$$\tau = \left[t_1^T \mid (t_1 \otimes \mathbf{1}_{m_2})^T,\ (t_1 \otimes \mathbf{1}_{m_2})^T,\ \ldots,\ (t_1 \otimes \mathbf{1}_{m_2})^T,\ (t_1 \otimes t_2)^T \mid t_2^T \right]^T, \tag{6.23}$$

and x^T denotes the transposed of the vector x. The matrix $A = \sum_{k\geq 0} A_k$ can be written as $A_0 + (\sum_{k>0} A_k) = K_0 \otimes I_l + K_1 \otimes (I_l - D_0)^{-1} D_1$ (see Eq. (3.2) and (3.4)). Moreover, it is easily seen that $K_0 + K_1$ is stochastic. Let $\gamma = \gamma(D_0 + D_1)$ and $\gamma \mathbf{1}_l = 1$, then $\gamma = \gamma D_1 (I_l - D_0)^{-1}$. As a result θ, the stochastic invariant vector of A, must equal $\kappa \otimes (\gamma D_1/\lambda)$, where $\lambda = \gamma D_1 \mathbf{1}_l$ is the arrival rate of the D-MAP and κ is the stochastic left invariant vector of $K = K_0 + K_1$:

$$K = \left[\begin{array}{c|cccc|c}
T_1 & t_1\alpha_1 \otimes \alpha_2 & \ldots & 0 & 0 \\
\hline
T_1 \otimes t_2 & T_1 \otimes T_2 + t_1\alpha_1 \otimes t_2\alpha_2 & \ldots & 0 & 0 \\
0 & T_1 \otimes t_2\alpha_2 & \ldots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \ldots & t_1\alpha_1 \otimes T_2 & 0 \\
0 & 0 & \ldots & T_1 \otimes T_2 + t_1\alpha_1 \otimes t_2\alpha_2 & t_1 \otimes T_2 \\
\hline
0 & 0 & \ldots & \alpha_1 \otimes t_2\alpha_2 & T_2
\end{array} \right], \tag{6.24}$$

Notice, K does not depend upon the arrival process. Given these results, we find that the stability condition $\theta\beta > 1$ can be written as

$$\theta\beta = \kappa\tau \, \frac{1}{\lambda} \, \overbrace{\gamma D_1 (I_l - D_0)^{-1} \mathbf{1}_l}^{=1} = k/\lambda > 1, \tag{6.25}$$

for some $k = \kappa\tau \geq 0$ that is independent of the arrival process. Thus, the stability condition is equivalent to $\lambda/k < 1$, meaning the arrival process only influences the stability of the system through its mean arrival rate. The constant k is determined by both service time distributions and the size of the intermediate waiting line B.

We end this subsection by having a closer look at the constant $k = \kappa\tau$. Denote $\kappa = [\kappa_{SI}, \kappa_0, \kappa_1, \ldots, \kappa_B, \kappa_{BL}]$, where κ_{SI} is a $1 \times m_1$ vector, κ_{BL} a $1 \times m_2$ vector and κ_j, for $j = 0, \ldots, B$, a $1 \times m_1 m_2$ vector. Looking at the matrix K it should be clear that

$$\kappa_{SI} + \sum_{j=0}^{B} \kappa_j \left(I_{m_1} \otimes \mathbf{1}_{m_2}\right) = \psi_1(1 - \kappa_{BL}\mathbf{1}_{m_2}), \tag{6.26}$$

where $\psi_1 = \psi_1(T_1 + t_1\alpha_1)$ and $\psi_1\mathbf{1}_{m_1} = 1$. Using this equality we find that k satisfies

$$k = \kappa\tau = \psi_1 t_1(1 - \kappa_{BL}\mathbf{1}_{m_2}) + \kappa_B(t_1 \otimes (t_2 - \mathbf{1}_{m_2})) + \kappa_{BL}t_2. \tag{6.27}$$

The vector κ_{BL} can be written as $\kappa_B(t_1 \otimes T_2(I_{m_2} - T_2)^{-1}) = \kappa_B(t_1 \otimes ((I_{m_2} - T_2)^{-1} - I_{m_2}))$. Therefore, Eq. (6.27) reduces to

$$k \;=\; \psi_1 t_1(1 - \kappa_{BL}\mathbf{1}_{m_2}) + \kappa_B(t_1 \otimes (t_2 - \mathbf{1}_{m_2} + (I_{m_2} - T_2)^{-1}t_2 - t_2)).$$

Now, $(I_{m_2} - T_2)^{-1}t_2 = \mathbf{1}_{m_2}$ and $\psi_1 t_1$ equals $\mu_1 = 1/E[S_1]$, where $E[S_1]$ is the mean service time in server 1. In conclusion, k can be written as

$$k = \mu_1(1 - \kappa_{BL}\mathbf{1}_{m_2}). \tag{6.28}$$

If we consider the system where there are always customers waiting in the waiting line of queue 1, then $\kappa_{BL}\mathbf{1}_{m_2}$ represents the probability that the customer occupying server 1 is blocked. Thus, k equals the output rate of server 1 (in such a system). Therefore, the probability that the second server is busy $(1 - \kappa_{SI}\mathbf{1}_{m_1})$ must equal k/μ_2. As a result, we have

$$k = \mu_1(1 - \kappa_{BL}\mathbf{1}_{m_2}) = \mu_2(1 - \kappa_{SI}\mathbf{1}_{m_1}). \tag{6.29}$$

Remark: Eq. (6.29) suffices to prove that interchanging both service time distributions does not affect the stability of the system. Indeed, if we add a "'bar"' to all the variables of the interchanged system we find: $\bar{\mu}_1 = \mu_2$, $\bar{\mu}_2 = \mu_1$, $\bar{\kappa}_{SI} = \kappa_{BL}$ and $\bar{\kappa}_{BL} = \kappa_{SI}$ (due to the symmetric nature of K). This implies that $\bar{k} = k$. Intuitively, when studying the stability, we may assume that there are always customers ready to be served in the infinite waiting line. Now, using an argument by Melamed [18], we can regard the empty places (holes) in the intermediate buffer as dual customers. For each regular customer that moves through the system, a dual customer receiving identical service moves in the opposite direction. Hence, the maximum stable thoughput of the regular customers and the dual customers must be identical. Clearly, the dual system is identical to the interchanged system. This type of interchangeability result was already established long ago for exponential servers (and Poisson arrivals) [19, Section 5.2].

6.2 Special case: no intermediate buffer ($B = 0$)

In the special case where the size of the intermediate waiting line B equals 0, we can prove that $1/k$ is nothing but the mean of the maximum of both PH service time distributions. This value equals the mean time a customer, called c, spends in server 1 provided that another customer started his service in server 2 at the same time as customer c. The proof goes as follows. Let S_{max} denote the maximum of both PH distributions, then S_{max} is also a PH distribution characterized by $(m_{max} = m_1+m_1m_2+m_2, T_{max}, \alpha_{max} = (0, \alpha_1\otimes\alpha_2, 0))$, where T_{max} is shown below:

$$T_{max} = \begin{bmatrix} T_1 & 0 & 0 \\ T_1 \otimes t_2 & T_1 \otimes T_2 & t_1 \otimes T_2 \\ 0 & 0 & T_2 \end{bmatrix}. \tag{6.30}$$

Now, $t_{max} = \mathbf{1}_{m_{max}} - T_{max}\mathbf{1}_{m_{max}} = [t_1^T, (t_1 \otimes t_2)^T, t_2^T]^T$. For $B = 0$, it is readily seen that $t_{max} = \tau$ and $T_{max} + t_{max}\alpha_{max} = K$. Now, the stochastic invariant vector ψ_{max} of $(T_{max}+t_{max}\alpha_{max})$ multiplied with t_{max} equals $1/E[S_{max}]$; therefore, the stability condition $\theta\beta > 1$ is reduced to $\lambda E[S_{max}] < 1$. This result can be seen as a generalized, discrete time counterpart of [6, Remark 8], where a similar result was obtained for a tandem queue with Poisson arrivals.

6.3 The stability of the queue length process

In this section, we demonstrate that the MC characterized by P is ergodic if and only if the queue contents of the infinite waiting line of queue 1 has a steady state. A MC that describes the queue contents can be formed using the same state space as P. The difference being that the variable i does not reflect the age of a possible customer in server 1, but the number of customers in queue 1 (in either the waiting line or the server). The resulting MC, which we characterize by a transition matrix P^+, is a QBD process. Its corresponding matrices A_0^+, A_1^+ and A_2^+ can be found as:

$$\begin{aligned} A_0^+ &= K_0 \otimes D_1, & (6.31)\\ A_1^+ &= K_0 \otimes D_0 + K_1 \otimes D_1, & (6.32)\\ A_2^+ &= K_1 \otimes D_0. & (6.33) \end{aligned}$$

Therefore, $A^+ = A_0^+ + A_1^+ + A_2^+ = K \otimes D$, where $D = D_0 + D_1$, which implies that θ^+, the left stochastic invariant vector of A^+, equals $\kappa\otimes\gamma$. If we multiply θ^+ with $\beta^+ = A_1^+ + 2A_2^+$ we find:

$$\theta^+\beta^+ = (\kappa \otimes \gamma)\left((K \otimes D_0)\mathbf{1}_{d_t} + (K_1 \otimes D)\mathbf{1}_{d_t}\right) = (1 - \lambda) + \kappa\tau, \tag{6.34}$$

which implies that the queue contents process is stable if $\lambda/k < 1$. This condition is identical to the stability condition of the GI/M/1 MC characterized by P.

7 Numerical Results

In this section we present some numerical examples that provide insight on the system behavior. Any other choice for the input parameters could have been made as long as the dimension d_t^* of the QBD matrices is not too large, say roughly below 1500.

7.1 Influence of the capacity B and the correlation of the D-MAP arrival process

Consider an interrupted Bernoulli process (IBP), where the mean sojourn time in both states equals x_c and an arrival occurs in the on-state with probability x_p. Thus,

$$D_0 = \begin{bmatrix} 1-1/x_c & 1/x_c \\ (1-x_p)/x_c & (1-x_p)(1-1/x_c) \end{bmatrix}, \quad D_1 = \begin{bmatrix} 0 & 0 \\ x_p/x_c & x_p(1-1/x_c) \end{bmatrix}. \quad (7.35)$$

The service time distribution in server 1 is hypergeometric with parameters:

$$\alpha_1 = [1/3, 2/3], \quad T_1 = \begin{bmatrix} 4/5 & 0 \\ 0 & 19/20 \end{bmatrix}, \quad (7.36)$$

Thus, with probability 1/3 and 2/3 the service time is geometrically distributed with a mean of 5 and 20 time units, respectively. The mean of this distribution is 15 time units. The service time distribution of the second server is characterized by

$$\alpha_2 = [1/16, 1/8, 13/16], \quad T_2 = \begin{bmatrix} 3/4 & 1/8 & 0 \\ 1/10 & 4/5 & 0 \\ 0 & 0 & x_l \end{bmatrix}, \quad (7.37)$$

where $x_l = 0.93887$ is chosen such that the mean service time equals 15. We have chosen the mean service time identical in both servers as this ought to create a strong coupling between both queues.

Figures 1 depicts the response time distribution for various capacities B, for $x_p = 1/16$ (meaning that the arrival rate $\lambda = 1/32$), and x_c either 40 or 400. Clearly, the larger x_c the more correlated the arrival process. Obviously, stronger correlated arrivals give rise to slower response times, while adding more capacity B between both servers reduces the response time. However, at some point there is little use in further augmenting the

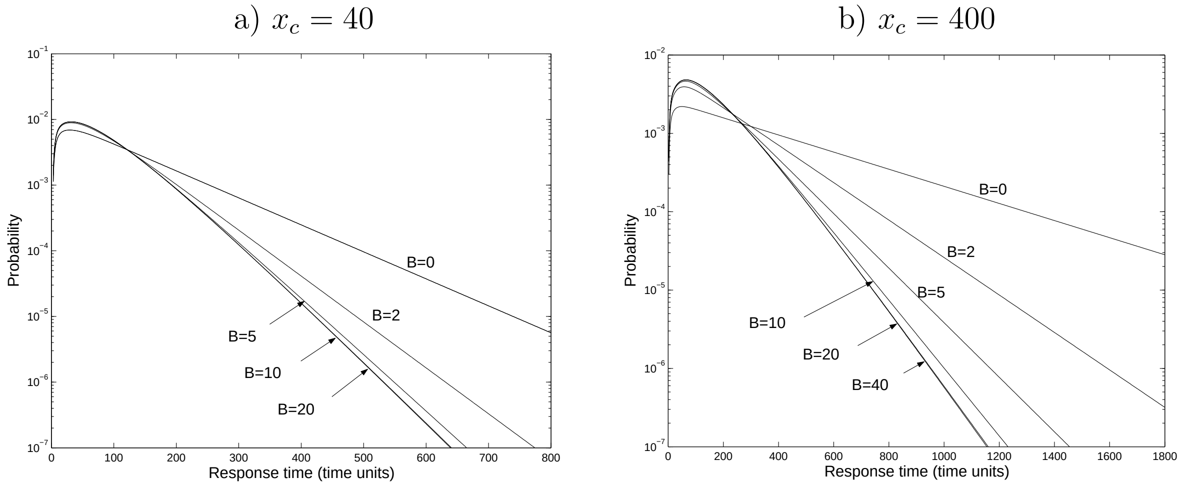

Figure 1: Response time distribution for various capacities B, $x_p = 1/16$ and a) $x_c = 40$, b) $x_c = 400$.

capacity as the response time seems to converge for B large. This is easily understood as the blocking probability tends to decrease to zero while increasing B. The figure further illustrates that the rate of convergence is affected by the correlation of the arrival process: stronger correlated arrival processes more easily justify increasing the capacity B.

7.2 The maximum arrival rate λ and the variation of the service times

We consider the same IBP process as in the previous section. The service time distribution is either geometric (Geo), Erlang-5 (Er5)[2] or hypergeometric (HypGeo) characterized by

$$\alpha_1 = [9/10, 1/10], \quad T_1 = \begin{bmatrix} 4/5 & 0 \\ 0 & 104/105 \end{bmatrix}. \tag{7.38}$$

The mean of each of these service time distributions is 15 time units. The HypGeo distribution is the most variable of the three, followed by Geo and Er5.

Figure 2 shows the maximum stable load ρ_{max}, defined as $E[S_1]\lambda_{max} = 15\lambda_{max}$, as a function of B for different server configurations. λ_{max} is the maximum arrival rate λ for which the system is stable. Recall, $\lambda_{max} = \mu_1(1 - \kappa_{BL}\mathbf{1}_{m_2}) = \mu_2(1 - \kappa_{SI}\mathbf{1}_{m_1})$, see Section 6. The notation (X, Y) indicates that the service time in server 1 and 2 is distributed as X and Y, respectively. Figure 2 clearly demonstrates that more variable service times, that is, HypGeo, give rise to a lower maximum stable input rate λ_{max}. This result stems

[2] That is, the sum of 5 independent and identically distributed geometric random variables.

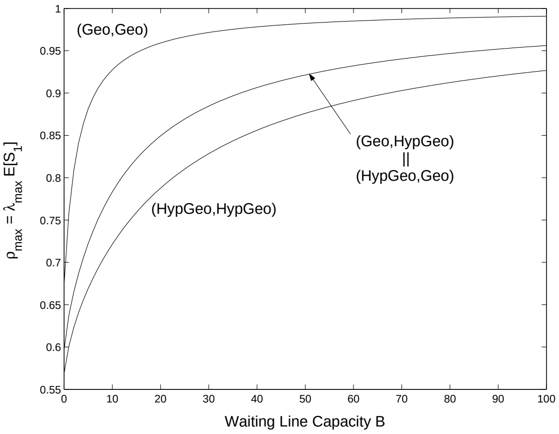

Figure 2: Maximum stable load ρ_{max} as a function of the capacity B for different server configurations.

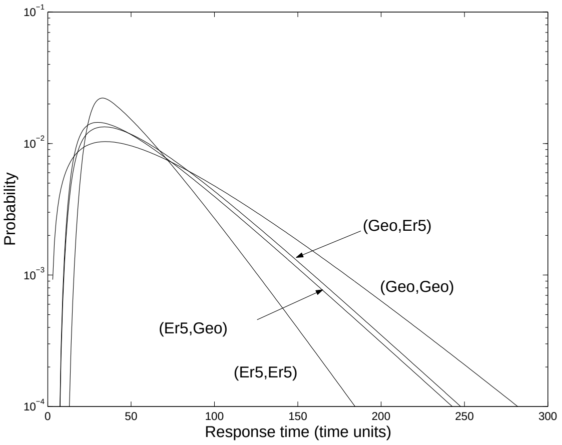

Figure 3: Response time distribution for different server configurations for $B = 10$ and IBP arrivals $(x_c = 40, x_p = 1/16)$.

from the fact that a more variable service time, whether in the first or second server, causes a higher degree of blocking in comparison with a more deterministic service time distribution. As proved in Section 6, interchanging both service time distributions does not alter the system stability. Notice, the actual nature of the D-MAP arrival process is irrelevant as the stability is only affected by the arrival process through its mean.

The response time distribution for different server configurations is depicted in Figure 3. We assume that arrivals occur according to a IBP with $x_c = 40$ and $x_p = 1/16$, see Section 7.1. The capacity of the intermediate waiting line B is assumed to be 10. Figure 3 confirms that the response of the system slows down as the service times become more variable. It further demonstrates that interchanging the service times generally causes a (limited) change in the response time (as opposed to the stability). Placing the more variable server first seems to result in a somewhat slower response. This might be explained by noticing that the output process of server 1 is more bursty in such case, causing a higher blocking probability. On the other hand, less variability in server 2 decreases the blocking probability, so when we interchange both service time distributions both these effects influence the degree of blocking in the system. Various numerical experiments, including Figure 3, seem to indicate that reducing the variation of server 1 should be slightly favored.

Acknowledgment

This work was performed while while the first author was visiting the University of Manitoba and TR Labs in Winnipeg, Manitoba. We would like to thank the University and TR Labs for their hospitality.

References

[1] B. Avi-Itzhak. A sequence of service stations with arbitrary input and regular service times. *Management Science*, 11(5):565–571, 1965.

[2] C. Blondia. A discrete-time batch markovian arrival process as B-ISDN traffic model. *Belgian Journal of Operations Research, Statistics and Computer Science*, 32(3,4), 1993.

[3] X. Chao and M. Pinedo. Batch arrivals to a tandem queue without an intermediate buffer. *Stochastic Models*, 6(4):735–748, 1990.

[4] H. Daduna. *Queueing Networks with Discrete Time Scale*. Springer, Berlin, 2001.

[5] B. Desert and H. Daduna. Discrete time tandem networks of queues: Effects of different regulation schemes for simultaneous events. *Perfomance Evaluation*, 47:73–104, 2002.

[6] A. Gómez-Corral. A tandem queue with blocking and markovian arrival process. *Queueing Systems*, 41:343–370, 2002.

[7] L. Gün. Annotated bibliography of blocking systems. Technical report, Institute for Systems Research ISR 1987-187, 1987.

[8] L. Gün and M. Makowski. Matrix geometric solution for finite capacity queues with phase-type distributions. In *Proceedings of Performance 8́7*, pages 269–282, Brussels, 1987.

[9] L. Gün and M. Makowski. Matrix geometric solution for two node tandem queueing systems with phase-type servers subject to blocking and failures. Technical report, Institute for Systems Research ISR 87-210, 1987.

[10] N. G. Hall and C. Sriskandarajah. A survey of machine scheduling problems with blocking and no-wait in process. *Operations Research*, 44:510–525, 1996.

[11] G.C. Hunt. Sequential arrays of waiting lines. *Operations Research*, 4(6):674–683, 1956.

[12] C. Knessl and C. Tier. Approximation to the moments of the sojourn time in a tandem queue with overtaking. *Stochastic Models*, 6(3):499–524, 1990.

[13] A. Lang and J. L. Arthur. Parameter approximation for phase-type distributions. In *Matrix-Analytic Methods in Stochastic Models, (S. R. Chakravarthy and A. S. Alfa (Editors))*, pages 151–206, New York, 1996. Marcel-Dekker, Inc.

[14] G. Latouche and M.F. Neuts. Efficient algorithmic solutions to exponential tandem queues with blocking. *SIAM J. algebraic and discrete meth.*, 1(1):93–106, 1980.

[15] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods and stochastic modeling.* SIAM, Philadelphia, 1999.

[16] D.M. Lucantoni. New results on the single server queue with a batch markovian arrival process. *Stochastic Models*, 7(1):1–46, 1991.

[17] B. Meini. Solving QBD problems: the cyclic reduction algorithm versus the invariant subspace method. *Advances in Performance Analysis*, 1:215–225, 1998.

[18] B. Melamed. A note on the reversibility and duality of some tandem blocking queueing systems. *Management Science*, 32(12):1648–1650, 1986.

[19] M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models, An Algorithmic Approach.* John Hopkins University Press, 1981.

[20] H. Perros. *Queueing Networks with Blocking.* Oxford University Press, New York, 1994.

[21] M. Schamber. Decomposition methods for finite queue networks with a non-renewal arrival process in discrete time, 1997.

[22] R. D. van der Mei, B. M. M. Gijsen, N. in't Veld, and J. L. van den Berg. Response times in a two-node queueing network with feedback. *Performance Evaluation*, 49:99–110, 2002.