A tandem queue with server slow-down and blocking.

van Foreest, N.; van Ommeren, J.C.; Mandjes, M.R.H.; Scheinhardt, W.

DOI
10.1081/STM-200056037

Publication date
2005

Published in
Stochastic Models

Centrum voor Wiskunde en Informatica

REPORTRAPPORT

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

PNA

A tandem queue with server slow-down and blocking

N.D. van Foreest, M.R.H. Mandjes,
J.C.W. van Ommeren, W.R.W. Scheinhardt

REPORT PNA-E0502 FEBRUARY 2005

CWI is the National Research Institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organization for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

A tandem queue with server slow-down and blocking

ABSTRACT
We consider two variants of a two-station tandem network with blocking. In both variants the first server ceases to work when the queue length at the second station hits a blocking threshold . In addition, in variant 2 the first server decreases its service rate when the second queue exceeds a slow-down threshold, which is smaller than the blocking level. In both variants the arrival process is Poisson and the service times at both stations are exponentially distributed. Note, however, that in case of slow-downs, server 1 works at a high rate, a slow rate, or not at all, depending on whether the second queue is below or above the slow-down threshold or at the blocking threshold, respectively. For variant 1, i.e., only blocking, we concentrate on the geometric decay rate of the number of jobs in the first buffer and prove that for increasing blocking thresholds the sequence of decay rates decreases monotonically and at least geometrically fast to max{rho_ 1, rho_2}, where rho_i is the load at server i. The methods used in the proof also allow us to clarify the asymptotic queue length distribution at the second station. Then we generalize the analysis to variant 2, i.e., slow-down and blocking, and establish analogous results.

2000 Mathematics Subject Classification: 60K25
Keywords and Phrases: tandem -- feedback -- quasi birth-death processes

A Tandem Queue with Server Slow-down and Blocking

N.D. van Foreest[•], M.R.H. Mandjes[◇‡], J.C.W. van Ommeren[•], W.R.W. Scheinhardt[•◇],
[•] University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands
[◇] Center for Mathematics and Computer Science (CWI)
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
[‡] University of Amsterdam (UvA)
Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands

Abstract

We consider two variants of a two-station tandem network with blocking. In both variants the first server ceases to work when the queue length at the second station hits a 'blocking threshold'. In addition, in variant 2 the first server decreases its service rate when the second queue exceeds a 'slow-down threshold', which is smaller than the blocking level. In both variants the arrival process is Poisson and the service times at both stations are exponentially distributed. Note, however, that in case of slow-downs, server 1 works at a high rate, a slow rate, or not at all, depending on whether the second queue is below or above the slow-down threshold or at the blocking threshold, respectively. For variant 1, i.e., only blocking, we concentrate on the geometric decay rate of the number of jobs in the first buffer and prove that for increasing blocking thresholds the sequence of decay rates decreases monotonically and at least geometrically fast to $\max\{\rho_1, \rho_2\}$, where ρ_i is the load at server i. The methods used in the proof also allow us to clarify the asymptotic queue length distribution at the second station. Then we generalize the analysis to variant 2, i.e., slow-down and blocking, and establish analogous results.

1 Introduction

In classical queueing networks service stations do not exchange information about their queue lengths. However, in general such communication might be useful. Suppose for instance that when the queue at some 'downstream' station builds up, this station can protect itself by signalling 'upstream' stations to decrease their service rate. In this way there is *congestion-dependent feedback of information* (not jobs) from downstream stations to upstream stations.

The tandem queue we study here resembles a two-station Jackson tandem queue in which jobs arrive according to a Poisson process with rate λ at the first station and require at the first and second station exponentially distributed service times with mean $1/\mu_1$ and $1/\mu_2$, respectively. Thus, the load on the first and second server is $\rho_1 := \lambda/\mu_1$ and $\rho_2 := \lambda/\mu_2$, respectively. However, we allow the second station to inform the first station about the number of jobs in queue. Immediately after the second station contains n jobs, it signals the first server to stop processing any job in service. We assume that the feedback signal from the second station to the first is not delayed. When the queue length in the second station becomes less than n, the first server may resume service again. Clearly, this blocking mechanism will protect the second station from overflow, at the cost of a stochastically longer queue at the first station.

First, we are interested in the effect on the first station as a function of the *blocking threshold* n. However, due to the presence of the feedback, the stationary joint distribution π_{ij} that the number of jobs in the first and second station is i and j, respectively, does not have a product-form, so that finding a closed-form expression for π_{ij} is difficult. We therefore concentrate on its (asymptotically) dominant structure and consider the *geometric decay rate* of the number of jobs in the first buffer. This quantity, also known as the caudal characteristic, cf. [12], gives insight into the probability of the first queue reaching a high level due to blocking. It turns out that the decay rate of the number of jobs in the first station lies somewhere in the interval $(\rho, 1)$ where $\rho \equiv \max\{\rho_1, \rho_2\}$, a result also obtained in [1]. However, in this paper we also show rigorously that the decay rate as a function of the blocking threshold decreases monotonically and at least geometrically fast to ρ.

As a second topic of interest we estimate the ratio $\pi_{i,j+1}/\pi_{ij}$ when $i \gg 1$, i.e., the ratio of the probability that the number of jobs in the second queue is $j+1$ to the probability that this number is j, while the first queue is large. Thus, our approach also reveals the asymptotic probabilistic structure of the number of jobs in the second station, which is not as simple to see as the decay rate of the first queue.

Third, we study a more complicated type of feedback. Now, when the number of jobs at station 2 is in excess of some threshold m (which should be smaller than the blocking threshold n to be effective), server 1 *slows down*, i.e., it reduces its service rate to $\tilde{\mu}_1$, where $0 < \tilde{\mu}_1 < \mu_1$. Thus, depending on the queue length in station 2, server 1 works at a high rate μ_1, a low rate $\tilde{\mu}_1$, or not at all. In the sequel we distinguish both types of feedback queue by calling the first the *network with blocking* and the second the *network with slow-down and blocking*. The analysis of such queueing networks with service slow-downs has interesting applications in the domain of manufacturing, but also in the design of Ethernet networks, where in point–to–point connections the sending side may react to congestion signals from the receiving side, see e.g. [13]. For the network with slow-down and blocking we can establish analogous results as obtained for the network with blocking. The asymptotic distribution of the number in the second queue turns out to be of particular interest in this case.

Our focus on the asymptotic behavior of π_{ij} has two reasons. First, the resulting expressions are in closed form, contrary to the numerical methods available in the literature. Second, given the rapid convergence of the sequence of networks with blocking when the blocking threshold increases, the asympotic system provides considerable insight in the form of π_{ij} even when n is small or the first queue contains few jobs.

Tandem queues with blocking (but *without slow-down*) received considerable attention over the years. The authors of [2, 3] take z-transforms of the balance equations satisfied by π_{ij} and study the properties of the resulting generating function to establish a stability condition and devise an algorithm to compute π_{ij}. The derivation of the stability condition for this and related models is simplified in [6] by using the methods of Quasi-Birth-Death (QBD) processes. In [1] the authors derive, also by using QBDs, a more efficient numerical procedure to compute π_{ij}. They restrict a number of eigenvalues to a set of (non-overlapping) intervals. After locating the eigenvalues in the bounding intervals, they derive a recursion to obtain the associated eigenvectors. Finally, a suitable linear combination of the eigenvectors should solve the boundary conditions for π_{0j}. Interestingly, by using the bounding intervals derived in [1] for the eigenvalues, our approach extends straightforwardly to a method to compute π_{ij} with the same algorithmic complexity as in [1]. These authors also mention the idea of slow-down but do not analyze the consequences in detail. Kroese et al. [4] also consider a two-station tandem queue with blocking. However, now the rate of the arrival process is set to zero when the first station contains n jobs. The second buffer is assumed infinitely large. For this system the authors compute the decay rate of the number of jobs in the second buffer. They also consider the limiting regime in which $n \to \infty$. Leskelä [8] studies a two-station tandem network with feedback, but now station 2, rather than station 1, provides feedback to the arrival process to change rate as a function of the length of the second queue. He establishes a stability criterion for the system with unlimited first *and* second buffer.

The paper has the following structure. In Section 2 we specify the network with blocking and write it as a QBD process. Next, in Section 3 we present our main results for this network and discuss them from an intuitive point of view. More specifically, we state that the decay rate x_n of the number of jobs in the *first* buffer lies in the interval $(\rho, 1)$ and we establish bounds of the rate at which the sequence $\{x_n\}_n$ converges downward to ρ when $n \to \infty$. In addition we present the asymptotic structure of the distribution of the number of jobs in the second buffer when the first queue is very long. Section 4 contains the proofs of these results, which are based on the theory of QBD processes as dealt with in [7] or [11], and the Perron-Frobenius theorem, cf. [5] or [9]. In Section 5 we consider similar topics for the tandem queue with slow-down and blocking.

2 Model and Preliminaries

We now present the model for the two-node tandem network with blocking, write it as a QBD process, and consider its stability conditions.

Jobs arrive according to a Poisson process with rate λ. Service requirements at the first (second) station are i.i.d. exponentially distributed random variables with mean μ_1^{-1} (μ_2^{-1}), while the two service processes are mutually independent and independent of the arrival process. We assume throughout this paper that $\mu_1 \neq \mu_2$ (see also the comment on this assumption in Section 3). After service completion at the first station, jobs move on to the second. Once service is completed there also, jobs leave the network. Let $X_i^{(n)}(t)$ denote the number of jobs at station i, $i = 1, 2$, at time t (including the job in service) for the system with blocking threshold at n. When $X_2^{(n)}(t)$ is equal to this threshold n, the first server blocks, i.e., its service rate becomes zero. Right after the departure of the job in service at the second station, the first server resumes service (if a job is present there, of course). It is clear that the joint process $\{X_1^{(n)}(t), X_2^{(n)}(t)\} \equiv \{X_1^{(n)}(t), X_2^{(n)}(t), t \geq 0\}$ is a (continuous-time) Markov chain. The state space of this process is $\mathscr{X}^{(n)} = \{(i, j) \,|\, i = 0, 1, 2, \ldots; j = 0, 1, \ldots, n\}$. We present the state transition diagram of $\{X_1^{(n)}(t), X_2^{(n)}(t)\}$ in Figure 1. Finally, let

$$\rho_1 := \lambda/\mu_1,\ \rho_2 := \lambda/\mu_2,\ \text{and}\ \rho := \max\{\rho_1, \rho_2\}, \tag{1}$$

i.e., ρ is the load at the slowest server. Note that when the first (resp. second) server is the bottleneck server, i.e. $\mu_1 < \mu_2$ (resp. $\mu_1 > \mu_2$) we can (and henceforth will) also state this by writing $\rho = \rho_1$ (resp. $\rho = \rho_2$), since we exclude the possibility $\mu_1 = \mu_2$.

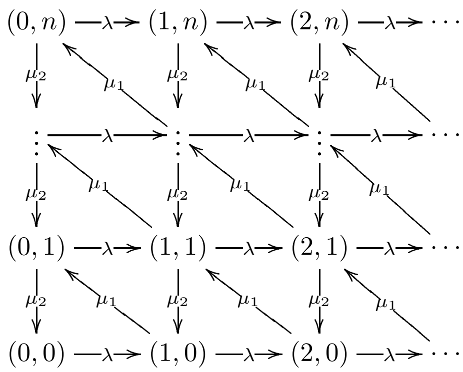

Figure 1: State space and transition rates of the truncated tandem queue.

The Markov process $\{X_1^{(n)}(t), X_2^{(n)}(t)\}$ can be interpreted as a continuous-time QBD process. We identify some common subsets of $\mathscr{X}^{(n)}$ associated specifically to the QBD structure. *Level* i contains all states $(i,j) \in \mathscr{X}^{(n)}$ with i constant. *Phase* j contains the states (i,j) with j constant. Thus, the levels contain the 'vertical' sets of states in Figure 1, whereas the phases contain the 'horizontal' sets of states.

To facilitate the presentation we prefer to uniformize the process $\{X_1^{(n)}(t), X_2^{(n)}(t)\}$ at rate

$$a := \lambda + \mu_1 + \mu_2,$$

and concentrate on the resulting aperiodic discrete-time Markov chain $\{X_{1,k}^{(n)}, X_{2,k}^{(n)}\}$. This procedure allows us to refer directly to a number of results in the literature which we otherwise have to reformulate for the continuous-time model. Evidently, by PASTA, the results we derive for $\{X_{1,k}^{(n)}, X_{2,k}^{(n)}\}$ apply also to $\{X_1^{(n)}(t), X_2^{(n)}(t)\}$.

The matrix of transition probabilities of the QBD $\{X_{1,k}^{(n)}, X_{2,k}^{(n)}\}$ is of the form

$$P^{(n)} = \begin{pmatrix} B^{(n)} & A_0^{(n)} & & & \\ A_2^{(n)} & A_1^{(n)} & A_0^{(n)} & & \\ & \ddots & \ddots & \ddots & \end{pmatrix}. \tag{2a}$$

The $(n+1)\times(n+1)$ matrices in $P^{(n)}$ are given by

$$B^{(n)} = \begin{pmatrix} q+r & & & \\ r & q & & \\ & \ddots & \ddots & \\ & & r & q \end{pmatrix}, \qquad A_0^{(n)} = p\, I^{(n)}, \tag{2b}$$

$$A_1^{(n)} = \begin{pmatrix} r & & & & \\ r & 0 & & & \\ & \ddots & \ddots & & \\ & & r & 0 & \\ & & & r & q \end{pmatrix}, \qquad A_2^{(n)} = \begin{pmatrix} 0 & q & & \\ & \ddots & \ddots & \\ & & 0 & q \\ & & & 0 \end{pmatrix}, \tag{2c}$$

where $p = \lambda/a$, $q = \mu_1/a$, $r = \mu_2/a$, and $I^{(n)}$ is the $(n+1)\times(n+1)$ identity matrix.

Provided a certain stability criterion to be addressed in Theorem 1 below is satisfied, an irreducible QBD chain is positive recurrent. Consequently, its stationary probability vector exists. Let us henceforth consider the system in steady state, and write for brevity $X_i^{(n)}$, $i = 1,2$, for $X_{i,k}^{(n)}$ at an arbitrary point in time. Furthermore, let $\pi_{ij}^{(n)} = \mathbb{P}\{X_1^{(n)} = i, X_2^{(n)} = j\}$, i.e., the steady-state probability that the number of jobs in the first and second station is i and j respectively.

It can be shown that the stationary probability vector $\boldsymbol{\pi}^{(n)}$ can be appropriately partitioned as

$$\boldsymbol{\pi}^{(n)} = \left(\boldsymbol{\pi}_0^{(n)}, \boldsymbol{\pi}_0^{(n)} R^{(n)}, \boldsymbol{\pi}_0^{(n)} \left(R^{(n)}\right)^2, \ldots, \right), \tag{3}$$

where $\boldsymbol{\pi}_0^{(n)} \left(R^{(n)}\right)^i = \left(\pi_{i0}^{(n)}, \pi_{i1}^{(n)}, \ldots, \pi_{in}^{(n)}\right)$ and $R^{(n)}$ is the minimal nonnegative solution of the equation

$$A_0^{(n)} + R^{(n)} A_1^{(n)} + \left(R^{(n)}\right)^2 A_2^{(n)} = R^{(n)}. \tag{4}$$

For our case $R^{(n)}$ has to be computed numerically, for instance with the algorithms derived by [7].

Rather than computing $R^{(n)}$ directly, [12] associates two interesting (probabilistic) quantities to $R^{(n)}$. He starts by observing that when $R^{(n)}$ is irreducible, it satisfies

$$\left(R^{(n)}\right)^i = (x_n)^i \left(\mathbf{u}^{(n)}\right)' \cdot \mathbf{v}^{(n)} + o\left((x_n)^i\right), \quad \text{as } i \to \infty, \tag{5}$$

where $\mathbf{v}^{(n)} = (v_0^{(n)}, \dots, v_n^{(n)})$ and $\mathbf{u}^{(n)}$ are strictly positive left and right eigenvectors of $R^{(n)}$ associated to its largest eigenvalue $x_n \in (0, 1)$. (The prime denotes the transpose of a vector.) The first quantity of interest is

$$\lim_{i\to\infty} \frac{\boldsymbol{\pi}_0^{(n)} \left(R^{(n)}\right)^{i+1} \mathbf{e}}{\boldsymbol{\pi}_0^{(n)} \left(R^{(n)}\right)^i \mathbf{e}} = x_n, \tag{6}$$

where $\mathbf{e}$ is the (column) vector consisting of ones. This says that the ratio of the expected time spent at a high level $i + 1$ to that spent at level i is approximately equal to x_n. In other words, the largest eigenvalue x_n of $R^{(n)}$ is the *geometric decay rate*, which is also known as the caudal characteristic, cf. [12], of the QBD process. Second,

$$\lim_{i\to\infty} \frac{(\boldsymbol{\pi}_0^{(n)} \left(R^{(n)}\right)^i)_j}{\boldsymbol{\pi}_0^{(n)} \left(R^{(n)}\right)^i \mathbf{e}} = v_j^{(n)}, \tag{7}$$

which is to say that (in stationary state) the probability that the chain is in phase j conditional on being in level i, is approximately equal to $v_j^{(n)}$ for large i.

It remains to discuss the stability condition of the chain $\{X_{1,k}^{(n)}, X_{2,k}^{(n)}\}$, which we henceforth assume satisfied. The proof of the next theorem involves the matrix $A^{(n)}(x)$, which is also used in Section 4 and is defined as

$$A^{(n)}(x) = A_0^{(n)} + xA_1^{(n)} + x^2 A_2^{(n)} = \begin{pmatrix} p + rx & qx^2 & & & \\ rx & p & qx^2 & & \\ & \ddots & \ddots & \ddots & \\ & & & rx & p + qx \end{pmatrix}, \quad x \in [0, 1]. \tag{8}$$

Theorem 1. *The chain* $\{X_{1,k}^{(n)}, X_{2,k}^{(n)}\}$ *is positive recurrent if and only if*

$$\frac{\lambda}{\mu_1\mu_2} \frac{{\mu_1}^{n+1} - {\mu_2}^{n+1}}{{\mu_1}^n - {\mu_2}^n} < 1. \tag{9}$$

This condition is equivalent to:

$$n > N(\rho_1, \rho_2) = \frac{\log(1 - \rho_1) - \log(1 - \rho_2)}{\log \rho_2 - \log \rho_1}. \tag{10}$$

Proof. It is simple to see that the QBD $\{X_{1,k}^{(n)}, X_{2,k}^{(n)}\}$ is irreducible and that the number of phases is finite. Moreover, the stochastic matrix $A^{(n)}(1)$ is irreducible. These properties allow us to apply [7, Theorem 7.2.3]. This theorem states that the QBD is positive recurrent iff $\boldsymbol{\alpha} A_0^{(n)}\mathbf{e} < \boldsymbol{\alpha} A_2^{(n)}\mathbf{e}$, where $\boldsymbol{\alpha}$ is the stationary probability vector of $A^{(n)}(1)$. Clearly, $A^{(n)}(1)$ is the stochastic matrix of a simple birth-death process. Hence, the desired solution vector $\boldsymbol{\alpha} = (\alpha_0, \dots, \alpha_n)$ is given by $\alpha_i = \alpha_0\beta^i$, $0 \le i \le n$, where $\beta = \mu_1/\mu_2$ and

$$\alpha_0 = \left(\sum_{i=0}^{n} \beta^i\right)^{-1} = \frac{1 - \beta}{1 - \beta^{n+1}}.$$

The condition $\boldsymbol{\alpha} A_0^{(n)}\mathbf{e} < \boldsymbol{\alpha} A_2^{(n)}\mathbf{e}$ becomes $\lambda < \mu_1 \sum_{i=0}^{n-1} \alpha_i$, which leads to (9). To arrive at (10), we rewrite this as

$$\rho_1 \frac{\beta^{n+1} - 1}{\beta^n - 1} < 1.$$

Assuming $\beta > 1$ (i.e., $\mu_2 < \mu_1$), we can rewrite this to $\rho_2\beta^n - \rho_1 < \beta^n - 1$, which is equivalent to

$$\beta^n > \frac{1 - \rho_1}{1 - \rho_2}.$$

By taking logarithms at both sides (and using $\beta = \rho_2/\rho_1$) we arrive at the result. The case $\beta < 1$ follows analogously. ∎

3 Asymptotic Results for the Tandem Queue with Blocking

In this section we present our main results for the tandem queue with blocking. We choose to present the proofs in the next section, since they are rather lengthy, involving several partial results and lemmas. Here we focus on the results themselves and try to understand them at an intuitive level. Partly this is possible by a comparison with similar notions derived for the standard Jackson tandem network (i.e. without blocking). It may seem that the network with blocking resembles the two-station tandem Jackson network more and more when the blocking threshold n increases. However, this is not true in all respects, as becomes apparent presently.

Before turning to our results, we like to recall our assumption $\mu_1 \neq \mu_2$ throughout this paper. The stability issue is easily settled when $\mu_1 = \mu_2 = \mu$ (condition (9) is then replaced by $\lambda/\mu < n/(n+1)$), but we did not aim at providing an analogue to Theorem 2 below, although we do believe that such an analogue must hold. Furthermore, it is unclear how a possible analogue of Theorem 3 below would read.

As a first point of interest let us see how the blocking level n influences the stability condition (9). Writing g_n for the left hand side of (9), the stability condition $\lim_{n\to\infty} g_n < 1$ is equivalent to the conditions $\rho_1 < 1$ when $\mu_1 < \mu_2$ (or $\rho = \rho_1$), and $\rho_2 < 1$ when $\mu_1 > \mu_2$ (or $\rho = \rho_2$). Thus we arrive at the condition $\rho < 1$, which is also the stability criterion of the two-station tandem Jackson network.

A second characteristic is the geometric decay rate x_n, i.e. the decay rate of the first station's queue length distribution, defined in (6). Interestingly, for certain parameter regimes, the limit of x_n as $n \to \infty$ is *different* from ρ_1, which is the geometric decay rate for the Jackson network. Specifically, we have the following.

Theorem 2.

(i) If the system with threshold at n is stable, i.e. $n > N(\rho_1, \rho_2)$ holds, then the decay rate x_n lies in the interval $(\rho, 1)$, where $\rho \equiv \max(\rho_1, \rho_2)$.

(ii) The sequence $\{x_n\}_{n>N(\rho_1,\rho_2)}$ decreases monotonically to ρ and its elements satisfy the bounds

$$0 < x_n - \rho < \begin{cases} \beta_1\gamma_1\, \alpha_1^n, & \text{if } \rho = \rho_1, \\ \beta_2\gamma_2\, \alpha_2^n, & \text{if } \rho = \rho_2, \end{cases} \tag{11}$$

where for $i = 1, 2,$ the constants α_i are in $(0, 1)$, and β_i and γ_i are positive constants, the precise form of which is presented in Section 4.2

In other words, the asymptotic queue length in the first station is mostly influenced by the bottleneck server. Moreover, the convergence of $x_n \downarrow \rho$ is at least geometrically fast.

At an intuitive level, the first statement is not too difficult to understand. To this end we view the two queues in tandem as one black box at which jobs arrive at rate λ. Since each job receives service at both stations, the slower server in the black box clearly dominates the total number of jobs in the box, wherever these jobs may reside. Thus, the decay rate of the total number of jobs must be bounded below by $\rho = \max\{\rho_1, \rho_2\}$. By 'opening the black box' we see that, as the second buffer is finite, necessarily the first queue is large when the system contains many jobs. Hence the decay rate of the number of jobs in the first station must be greater than or equal to ρ. The other claims however appear less evident, in particular the geometric bounds on the difference between x_n and ρ.

As a third topic of interest we explore the probabilistic structure in the direction of the phases for some given level $i \gg 1$. A convenient notion to consider in the present setting is the ratio of the probability that the chain is in phase $j+1$ to the probability that the chain is in phase j, while the chain is in some high level i.

Theorem 3. *For a stable system the following statements hold.*

(i)

$$\lim_{i\to\infty} \frac{\pi_{i,j+1}^{(n)}}{\pi_{i,j}^{(n)}} = \frac{v_{j+1}^{(n)}}{v_j^{(n)}}. \tag{12a}$$

(ii) When $\rho = \rho_1$, we have

$$\left| \frac{v_{j+1}^{(n)}}{v_j^{(n)}} - \frac{\lambda}{\mu_2} \right| < \frac{\beta_1}{\mu_2\,\rho_1}\,\alpha_1^{n-j} \;+\; \frac{(\mu_1 - \gamma_1^{-1})\,\beta_1\,\gamma_1}{\mu_2\,\rho_1}\,\alpha_1^n, \tag{12b}$$

and when $\rho = \rho_2$,

$$\left| \frac{v_{j+1}^{(n)}}{v_j^{(n)}} - \frac{\mu_1}{\mu_2} \right| < \frac{\beta_2}{\lambda}\,\alpha_2^j \;+\; \frac{1-\rho_2}{1-\rho_1}\,\frac{\beta_2\,\gamma_2}{\rho_2}\,\alpha_2^n. \tag{12c}$$

Here the constants $\alpha_i, \beta_i, \gamma_i, i = 1, 2,$ are the same as in Theorem 2.

The first statement explains why $v_{j+1}^{(n)}/v_j^{(n)}$ is the quantity of interest here, and is an immediate consequence of (7). Clearly, the main relevance of the second statement is that when $\rho = \rho_1$ ($\rho = \rho_2$), the quantity $v_{j+1}^{(n)}/v_j^{(n)}$ converges to a value not far from λ/μ_2 (μ_1/μ_2), as $n \to \infty$, provided that j is not too close to n (0). In particular, the upper bound on the distance between $v_{j+1}^{(n)}/v_j^{(n)}$ and λ/μ_2 (μ_1/μ_2) when $\rho = \rho_1$ ($\rho = \rho_2$) depends on j, also when n grows large.

Again we like to contrast these results to those of the tandem Jackson network at an intuitive level. The stationary distribution of this latter network has product-from, hence, denoting the quantities related to the Jackson tandem queue with a superscript ∞,

$$\frac{\pi_{i,j+1}^{(\infty)}}{\pi_{i,j}^{(\infty)}} = \rho_2, \quad \text{for all } (i,j) \in \mathscr{X}^{(\infty)}. \tag{13}$$

Note that this ratio does *not* depend on which of the two servers is the bottleneck, nor does it depend on j, the queue length in the second station, and finally it holds for all i, and not just for $i \to \infty$ as in Theorem 3.

Now, when the first server is the bottleneck and $i \gg 1$ we conclude from our theorem that the tandem network with blocking behaves similarly as the Jackson tandem network, as is also the case for the decay rate in Theorem 2. Namely, if $n \gg 1$, the geometric decay rate is approximately ρ_1 and $\pi_{i,j+1}^{(n)}/\pi_{i,j}^{(n)} \approx \rho_2$, if also $i \gg 1$.

However, when the second server is the bottleneck, the situation is strikingly different. For the tandem network with blocking we see that the geometric decay rate is larger than ρ_2, not ρ_1. Moreover, when $i \gg 1$, the ratio $\pi_{i,j+1}^{(n)}/\pi_{i,j}^{(n)} \approx \mu_1/\mu_2$, whereas this ratio is λ/μ_2 for the Jackson network. It need not surprise us that the outcomes are different, since the behavior of the system with blocking and $\rho = \rho_2$ will be such that when $i \gg 1$ the number of jobs in the second queue will mostly be high, typically in the neighorhood of n. We can therefore expect other boundary effects than in the case $\rho = \rho_1$ and the Jackson network. This also explains why in this case the ratio should be larger than 1 (and indeed $\mu_1/\mu_2 > 1$ in this case). However, at the moment it is unclear to us how the actual value μ_1/μ_2 can be understood. One may be inclined to reason that, when $i \gg 1$, the arrival rate at the second queue is μ_1. Indeed this is the case, but simply dividing this by μ_2 is not the correct way to find the "local decay rate in the direction of the phases for some large level i". Namely, our quantity of interest is a ratio of stationary probabilities, the determination of which also involves the boundary behavior at $i = 0$. Another way to see that this reasoning is not correct, is that it would then also hold for the case $\rho = \rho_1$ and for the Jackson network, which is apparently not true.

Since we do not fully understand the precise value of μ_1/μ_2, we leave any further intuitive, probabilistic, explanations for Theorems 2 and 3 to future work, and present our (analytic) proofs in Section 4. The analytic approach also finds motivation in that it enables us in Section 5 to explore *networks with slow-down and blocking*, which seem even more complicated to handle probabilistically. As a side result, we also obtain an algorithm to compute x_n by means of bi-section in Section 4.1, see Corollary 12 and Remark 13.

4 Proofs of Theorems 2 and 3

We now successively prove Theorems 2(i), 2(ii) and 3. Although Theorem 2(i) is the least difficult to understand, and indeed known as we mentioned before, cf. [1], the preparations for the proof of this result take up most of the space. However, the machinery used is not difficult and provides us with the tools to give short proofs of Theorems 2(ii) and 3.

4.1 Proof of Theorem 2(i)

4.1.1 Method of proof

Here, and in the remainder of Section 4.1 we fix the blocking threshold n and prove that the decay rate x_n lies in the open interval $(\rho, 1)$. To achieve this, we use the following result stated in [7, Section 9.1]:

Theorem 4. *The decay rate x_n is the unique solution in $(0,1)$ of the equation*

$$x = \xi^{(n)}(x), \tag{14}$$

where $\xi^{(n)}(x)$ is the spectral radius of $A^{(n)}(x)$.

We apply this as follows. Since $A^{(n)}(x)$ is irreducible and nonnegative for $x > 0$, it follows from the Perron-Frobenius theorem that the spectral radius $\xi^{(n)}(x)$ is also the largest (and simple) eigenvalue of $A^{(n)}(x)$. Suppose now that we can find an $(n+1)$-dimensional row vector $\mathbf{v}^{(n)} > 0$, i.e., each component $v_j^{(n)}$ of $\mathbf{v}^{(n)}$ is strictly positive, and $x > 0$ such that

$$\mathbf{v}^{(n)} A^{(n)}(x) = \mathbf{v}^{(n)} x. \tag{15}$$

Then by the Perron-Frobenius theorem, x necessarily solves the equation $x = \xi^{(n)}(x)$, and $\mathbf{v}^{(n)}$ is the left Perron-Frobenius vector of $A^{(n)}(x)$. In fact it is the same vector $\mathbf{v}^{(n)}$ as introduced in Section 2, since this vector satisfies $\mathbf{v}^{(n)}R^{(n)} = x_n\mathbf{v}^{(n)}$ and hence, by (4), also $\mathbf{v}^{(n)}A^{(n)}(x_n) = x_n\mathbf{v}^{(n)}$, so that it must be equal to the left Perron-Frobenius vector of $A^{(n)}(x_n)$, which is unique up to scaling.

Below we use formula (15) to efficiently combine $\xi^{(n)}(x)$ and the components of the Perron-Frobenius eigenvector into a sequence of functions. This will then lead to an even simpler characterization of the decay rate x_n, see Theorem 5, after which we can work out the details and prove Theorem 2(i). Since in this section the blocking threshold n is fixed, we will mostly suppress the dependence on n here. However, we always write x_n for the decay rate.

To introduce the sequence just mentioned, let us interpret (15) as a constraint on x and $\mathbf{v}$ and work out its implications. Thus, assuming that (15) is true and expanding with (8) we find that $x > 0$ and $\mathbf{v} > 0$ should satisfy

$$x = p + rx + \frac{rxv_1}{v_0}, \tag{16a}$$

$$x = \frac{qx^2 v_{j-1}}{v_j} + p + \frac{rxv_{j+1}}{v_j}, \quad 1 \le j < n, \tag{16b}$$

$$x = \frac{qx^2 v_{n-1}}{v_n} + p + qx. \tag{16c}$$

From the first relation we see that for given x and v_0, the value of v_1 follows. But then, the second relation provides $v_2, \ldots, v_n$. Since we are free to choose the norm of $\mathbf{v}$, we can set, arbitrarily, $v_0 \equiv 1$. As a consequence, the first and second relation completely fix $\mathbf{v}$ once x is given. The third relation forms a necessary condition on x such that x and $\mathbf{v}$ indeed form an eigenvalue and eigenvector pair of $A^{(n)}(x)$. In other words, whereas the simultaneous validity of the first and second relation above leaves x free, the third relation fixes it.

To further clarify the structure of (16) and the dependence on x, we now define the following sequence of functions of x:

$$\chi_0(x) := \mu_1 x^2, \tag{17a}$$

$$\chi_j(x) := arx\frac{v_j}{v_{j-1}} = \mu_2 x\frac{v_j}{v_{j-1}}, \qquad 1 \le j \le n, \tag{17b}$$

$$\chi_{n+1}(x) := ax - \lambda - \frac{\mu_1\mu_2 x^3}{\chi_n(x)}. \tag{17c}$$

We define $\chi_0(x)$ and $\chi_{n+1}(x)$ for notational convenience, although they do not relate immediately to $\mathbf{v}$ by (17b). Now, multiply the left and right hand sides of (16) by $a = \lambda + \mu_1 + \mu_2$ and rearrange, to obtain, respectively,

$$\chi_1(x) = ax - \lambda - \frac{\mu_1\mu_2 x^3}{\chi_0(x)} = (\lambda + \mu_1)x - \lambda, \tag{18a}$$

$$\chi_j(x) = ax - \lambda - \frac{\mu_1\mu_2 x^3}{\chi_{j-1}(x)}, \qquad 2 \le j \le n+1, \tag{18b}$$

$$\chi_{n+1}(x) = \mu_1 x. \tag{18c}$$

From the above we conclude the following.

Theorem 5. *Let $x \in (0, 1)$ be such that the sequence $\{\chi_j(x)\}_{0 \le j \le n+1}$ satisfies (18) and each element $\chi_j(x) > 0$. Then x is the unique solution of $\xi^{(n)}(x) = x$, i.e., x equals the geometric decay rate x_n of the tandem queue with blocking at threshold n.*

Proof. When x satisfies the hypothesis, the validity of (15) follows by constructing $\mathbf{v}$ according to (17b). Regarding the positivity of $\mathbf{v}$, which we do not require in the definition (17) of $\chi_j(x)$, the conditions $x > 0$ and $\chi_j > 0$ imply that v_j and v_{j-1} have the same sign. Hence, as all $\chi_j > 0$, it is straightforward to construct $\mathbf{v} > 0$. ∎

Remark 6. It is apparent from (18) that the desired x can be expressed as a root of a polynomial. However, this insight might not provide the easiest method to characterize the decay rate. With the approach below we can achieve our goals with elementary methods. Hence, we do not try to bound the decay rate by locating or bounding the roots of polynomials.

Our search for the decay rate x_n thus motivates a study of the structure of the sequence $\{\chi_j(x)\}_{0 \le j \le n+1}$. First we explore the properties of this sequence, fixing x, in Section 4.1.2. Finally, in Section 4.1.3, we vary x such that, by combining and exploiting these properties, we arrive at the proof of Theorem 2(i), based on Theorem 5.

4.1.2 The sequence $\{\chi_j(x)\}$ with x fixed

For fixed x, (18) clearly shows that the elements of $\{\chi_j(x)\}_{0\le j\le n+1}$ satisfy a recurrence relation. Let, again for fixed x, the mapping T be given by

$$T : \eta \mapsto ax - \lambda - \frac{\mu_1\mu_2 x^3}{\eta}. \tag{19}$$

Then we can write

$$\chi_{j+1}(x) = T\left(\chi_j(x)\right), \qquad \text{for } 0 \le j \le n. \tag{20}$$

It turns out that T is the key to understanding the structure of $\{\chi_j\}$, and thereby to obtaining the decay rate.

The mapping $\eta \to T(\eta)$ is a hyperbolic linear fractional transformation, see, e.g., [10]. It is infinitely differentiable everywhere except in the origin, and it has an inverse, given by

$$T^{(-1)} : \eta \mapsto \frac{\mu_1\mu_2 x^3}{ax - \lambda - \eta}.$$

The equation $\eta = T(\eta)$ reveals that T has two fixed points: η_+ and η_-. These points are the solutions of the quadratic (in η) equation $\eta^2 - (ax-\lambda)\eta + \mu_1\mu_2 x^3 = 0$ so that

$$\eta_\pm = \frac{ax-\lambda}{2} \pm \frac{1}{2}\sqrt{(ax-\lambda)^2 - 4\mu_1\mu_2 x^3}. \tag{21}$$

In Section 4.1.3 we show that only real-valued $\eta_\pm$ are of importance for our purposes. Hence, it suffices to take x such that the discriminant

$$D(x) = (ax-\lambda)^2 - 4\mu_1\mu_2 x^3 > 0. \tag{22}$$

The behavior of the sequence of iterates $\ldots, T^{(-1)}(\eta), T^{(0)}(\eta) := \eta, T^{(1)}(\eta), \ldots$ for $\eta \in (\eta_-, \eta_+)$ is also of interest. The next lemma formalizes what might be anticipated from Figure 2.

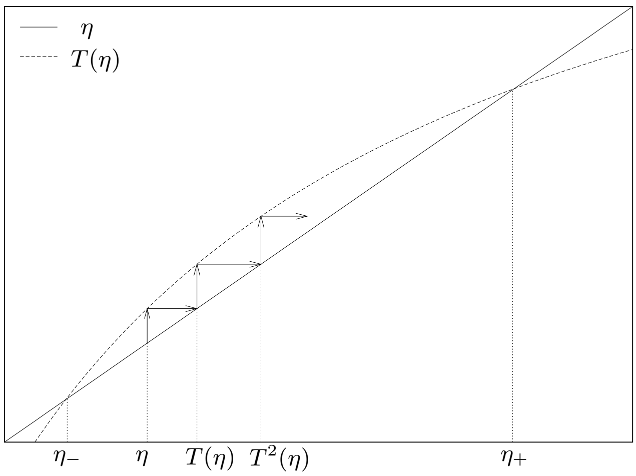

Figure 2: Some properties of the mapping $\eta \to T(\eta)$. The variable η is set out along the horizontal axis. The solid line refers to the identity.

Lemma 7. *If x such that $D(x) > 0$ (which implies that η_- and η_+ are real) and $\eta \in (\eta_-, \eta_+)$, then*

$$\eta_- = \lim_{i\to\infty} T^{(-i)}(\eta) < T^{(-1)}(\eta) < \eta < T(\eta) < \lim_{j\to\infty} T^{(j)}(\eta) = \eta_+,$$

$$\eta_+ - T^{(j)}(\eta) < \left(\frac{\eta_-}{\eta}\right)^j (\eta_+ - \eta), \qquad j > 0,$$

$$T^{(-i)}(\eta) - \eta_- < \left(\frac{\eta}{\eta_+}\right)^i (\eta - \eta_-), \qquad i > 0.$$

Proof. First, from (21) we have

$$\eta_+ + \eta_- = ax - \lambda, \qquad \text{and} \qquad \eta_-\eta_+ = \mu_1\mu_2 x^3.$$

Now, as $\eta \in (\eta_-, \eta_+)$, it follows that

$$\eta_+ - T(\eta) = \eta_+ - (ax - \lambda) + \frac{\mu_1\mu_2 x^3}{\eta} = -\eta_- + \frac{\eta_-\eta_+}{\eta} = \frac{\eta_-}{\eta}(\eta_+ - \eta),$$

Clearly, η_-/η and $\eta_+ - \eta$ are positive, which implies $\eta_+ > T(\eta)$. Moreover, $\eta_-/\eta < 1$ so that $\eta_+ - T(\eta) < \eta_+ - \eta$. Therefore, for all $\eta \in (\eta_-, \eta_+)$ we have $\eta_- < \eta < T(\eta) < \eta_+$. Concerning the convergence rate to η_+, note that

$$\begin{aligned}\eta_+ - T^{(2)}(\eta) &= \frac{\eta_-}{T(\eta)}(\eta_+ - T(\eta)) \\ &= \frac{\eta_-^2}{T(\eta)\eta}(\eta_+ - \eta) < \left(\frac{\eta_-}{\eta}\right)^2(\eta_+ - \eta).\end{aligned}$$

By induction, $T^{(j)}(\eta) \to \eta_+$ at least geometrically fast.

By similar computations we obtain

$$T^{(-1)}(\eta) - \eta_- = \frac{T^{(-1)}(\eta)}{\eta_+}(\eta - \eta_-) > 0.$$

So, $\eta_- < T^{(-1)}(\eta) < \eta < \eta_+$ whenever $\eta \in (\eta_-, \eta_+)$, and $T^{(-i)}(\eta) - \eta_- = (\eta - \eta_-)\prod_{k=1}^{i} T^{(-k)}(\eta)/\eta_+^i$, which is strictly smaller than $(\eta/\eta_+)^i(\eta - \eta_-)$. ∎

4.1.3 Varying x — the result

From now on we will view $\chi_j(x)$ again as a function of x. We start with pointing out an interesting, and perhaps unexpected, relation between the stability of the QBD chain $\{X_{1,k}^{(n)}, X_{2,k}^{(n)}\}$ and the derivative of χ_{n+1} with respect to x.

Lemma 8. *The stability condition* (9) *for the Markov chain* $\{X_{1,k}^{(n)}, X_{2,k}^{(n)}\}$ *is satisfied if and only if*

$$\mu_1 > \chi'_{n+1}(1).$$

Proof. First of all, the differentiability of T implies (by the chain rule) that $\chi_{n+1}(x)$ has a derivative. Next, from (18) it is immediate that $\chi_j(1) = \mu_1$ for all $j = 0, \ldots, n+1$. Hence, from (18) and writing $\beta = \mu_1/\mu_2$ as in the proof of Theorem 1, we find by induction

$$\chi'_j(1) = (\lambda + \mu_1)\frac{1-\beta^{-j}}{1-\beta^{-1}} - 2\mu_2\frac{1-\beta^{-j+1}}{1-\beta^{-1}}, \quad 0 \le j \le n+1.$$

The condition $\chi'_{n+1}(1) < \mu_1$ is therefore equivalent to

$$\lambda\frac{1-\beta^{-(n+1)}}{1-\beta^{-1}} + \mu_1\beta^{-1}\frac{1-\beta^{-n}}{1-\beta^{-1}} - 2\mu_2\frac{1-\beta^{-n}}{1-\beta^{-1}} < 0.$$

After a bit of algebra we see that this condition is precisely (9). ∎

Let us now concentrate on the fixed points η_+ and η_- of T. From their definition (21) it can be seen that they are also functions of x. To provide further intuition about these functions, we plot in Figure 3 their graphs together with $\chi_2(x)$ and $\chi_3(x)$.

Lemma 9. *First, the functions* $x \to \eta_\pm(x)$ *are real valued and positive on* $[\rho, 1]$. *Second,*

$$\eta_-(x) < \chi_0(x) = \mu_1 x^2 < \eta_+(x), \quad \textit{if } x \in (\rho, 1). \tag{23}$$

Third,

$$\chi_0(\rho_1) = \lambda\rho_1 = \eta_-(\rho_1), \tag{24a}$$

$$\chi_0(\rho_2) \in (\eta_-(\rho_2), \eta_+(\rho_2)) = (\lambda\rho_2, \mu_1\rho_2), \quad \textit{if } \rho = \rho_2, \tag{24b}$$

$$\chi_0(1) = \mu_1 = \begin{cases} \eta_-(1), & \textit{if } \rho = \rho_1, \\ \eta_+(1), & \textit{if } \rho = \rho_2. \end{cases} \tag{24c}$$

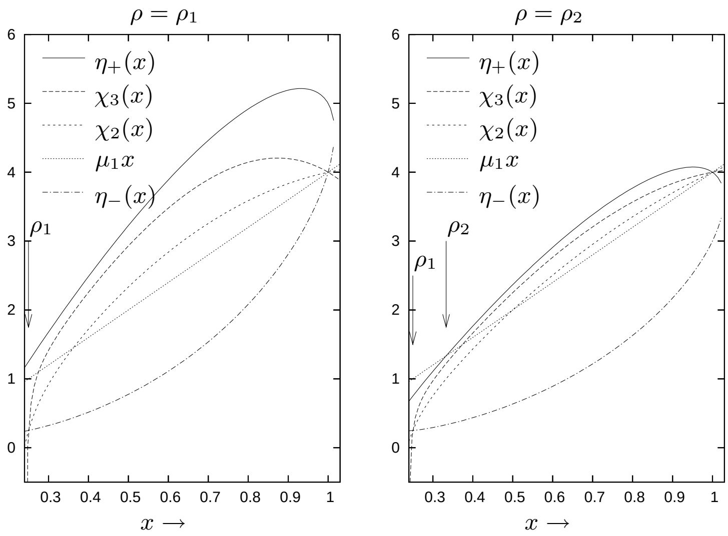

Figure 3: Plots of the functions $\chi_2(x)$, $\chi_3(x)$, and $\eta_\pm(x)$. In the left panel $\lambda = 1, \mu_1 = 4, \mu_2 = 5$, while in the right $\lambda = 1, \mu_1 = 4, \mu_2 = 3$.

Proof. For the first claim we focus on the discriminant $D(x) = (ax - \lambda)^2 - 4\mu_1\mu_2 x^3$ in the definition of $\eta_\pm(x)$. Clearly, $D(x)$, being a cubic polynomial, can have at most three real roots: ξ_1, ξ_2, and ξ_3, say. By simple computations we see that $D(0) > 0$, $D(\lambda/a) < 0$, $D(\rho_2) > 0$, $D(\rho_1) > 0$, $D(1) \geq 0$, and $\lim_{x\to\infty} D(x) = -\infty$. It follows that $0 < \xi_1 < \lambda/a < \xi_2 < \min\{\rho_1, \rho_2\} \leq \max\{\rho_1, \rho_2\} < 1 \leq \xi_3$. So, on $[\rho, 1]$ the discriminant $D(x)$ is positive, and $\eta_\pm(x)$ are real valued. It is now simple to check that $\eta_\pm(x) > 0$ for $x \in [\rho, 1]$.

To prove the second claim, rewrite the inequality $\eta_-(x) < \mu_1 x^2 < \eta_+(x)$ to

$$(2\mu_1 x^2 - (ax - \lambda))^2 \leq (ax - \lambda)^2 - 4\mu_1\mu_2 x^3.$$

After some algebra and using the positivity of x we find the above to be equivalent to $\lambda(1 - x) < \mu_1\, x(1 - x)$. This is clearly true for all $x \in (\rho_1, 1)$ and, hence, for all $x \in (\rho, 1)$.

Verifying the third claim is simple. ∎

With the above observations it is straightforward to apply Lemma 7 to the functions $\chi_j(x)$, $0 \leq j \leq n + 1$. For later purposes we formulate this intermediate result in somewhat greater generality than is necessary for the moment. The generalization consists of extending $\{\chi_j(x)\}_{0\leq j\leq n+1}$ to a doubly infinite sequence $\{\chi_j(x)\}_{j\in\mathbb{Z}}$ by continuing in (20) the iterative operation of T and $T^{(-1)}$ beyond χ_{n+1} and χ_0, respectively. Thus, define for $j \geq 1$,

$$\chi_j(x) := T^{(j)}\left(\chi_0(x)\right) = T\left(T^{(j-1)}\left(\chi_0(x)\right)\right) = T\left(\chi_{j-1}(x)\right),$$
$$\chi_{-j}(x) := T^{(-j)}\left(\chi_0(x)\right) = T^{(-1)}\left(T^{(-j+1)}\left(\chi_0(x)\right)\right) = T^{(-1)}\left(\chi_{-j+1}(x)\right).$$

This extension allows us to state the following.

Lemma 10. *Whenever $x \in (\rho, 1)$,*

$$\begin{aligned} \eta_-(x) < \ldots < \chi_{-i}(x) < \ldots < \chi_0(x) < \chi_1(x) < \ldots \\ < \chi_{n+1}(x) < \ldots < \chi_j(x) < \ldots < \eta_+(x), \end{aligned} \tag{25}$$

for $i > 0$ and $j > n + 1$. Moreover, $\chi_{-i}(x) \to \eta_-(x)$ and $\chi_j(x) \to \eta_+(x)$ geometrically fast for $i, j \to \infty$.

Proof. As, by Lemma 9, $x \in (\rho, 1)$ implies that $\chi_0(x) \in (\eta_-(x), \eta_+(x))$, we can use $\chi_0(x)$ as the 'starting point' for (the iterates of) T and $T^{(-1)}$ and apply Lemma 7. ∎

As a last intermediate result we consider the concavity of the sequence of functions $\chi_j(x)$, $2 \leq j \leq n + 1$, and $\eta_+(x)$. Proving that $\eta_+(x)$ is concave is not immediate as the discriminant (22) need *not* be concave on $(\rho, 1)$.

Lemma 11. *The functions $\chi_j(x)$, $2 \le j \le n+1$, and $\eta_+(x)$ are strictly concave on $(\rho, 1)$. The function $\eta_-(x)$ is strictly convex on $(\rho, 1)$.*

Proof. We assert by induction that $\chi_j''(x) < 0$ for all $x \in (\rho, 1)$ and $j \ge 2$. First, $\chi_1(x) = (\lambda + \mu_1)x - \lambda$ is concave. Now, for $j \ge 2$, we have by (18),

$$\frac{\chi_j''(x)}{\mu_1\mu_2} = \frac{x}{\chi_{j-1}(x)} \left(-6 + 6\,\frac{x\,\chi_{j-1}'(x)}{\chi_{j-1}(x)} - 2\left(\frac{x\,\chi_{j-1}'(x)}{\chi_{j-1}(x)}\right)^2 + \frac{x^2\chi_{j-1}''(x)}{\chi_{j-1}(x)} \right).$$

Let $y(x) = x\chi_{j-1}'(x)/\chi_{j-1}(x)$ and write the first three terms within the brackets as the parabola $-6+6y-2y^2$. It is simple to see that, as both roots are not real, this parabola is negative for all y. The fourth term in the expression above cannot be positive as $\chi_{j-1}(x) > 0$ for $x \in (\rho, 1)$ and $\chi_{j-1}''(x) \le 0$, by the induction hypothesis. Hence, $\chi_j''(x) < 0$.

Now, for any $x, y \in [\rho, 1]$, and $\alpha \in (0, 1)$ take the limit $j \to \infty$ of both sides of

$$\chi_j(\alpha x + (1-\alpha)y) > \alpha\chi_j(x) + (1-\alpha)\chi_j(y),$$

and conclude that $\eta_+(x)$ is also strictly concave. Finally, since $\eta_-(x) = ax - \lambda - \eta_+(x)$, it follows that $\eta_-(x)$ is strictly convex. ∎

By now we have identified all required intermediate results so that we can bound x_n from below.

Proof of Theorem 2(i). We prove that the conditions of Theorem 5 are satisfied. Regarding the positivity of the numbers $\chi_j(x)$ for $x \in (\rho, 1)$ we have by Lemmas 9 and 10 that $\chi_j(x) > \eta_-(x) > 0$ for $j = 0, \ldots, n+1$. It remains to prove that the function $\chi_{n+1}(x)$ intersects the line $\mu_1 x$ somewhere in the interval $(\rho, 1)$. First, from (24a) $\chi_{n+1}(\rho_1) = \eta_-(\rho_1) = \lambda\rho_1 < \mu_1\rho_1$. Also, when $\rho = \rho_2$, $\chi_0(\rho_2) \in (\eta_-(\rho_2), \eta_+(\rho_2))$, which implies by (25) that $\chi_{n+1}(\rho_2) < \eta_+(\rho_2) = \mu_1\rho_2$. Hence, $\chi_{n+1}(\rho) < \mu_1\rho$. On the other hand, $\chi_{n+1}(1) = \mu_1$ and $\chi_{n+1}'(1) < \mu_1$, by Lemma 8. Consequently, the concavity of $\chi_{n+1}(\cdot)$ implies there exists a unique $x \in (\rho, 1)$ such that $\chi_{n+1}(x) = \mu_1 x$. ∎

As a direct by-product of the above proof and the uniqueness of the solution of $\mu_1 x = \chi_{n+1}(x)$ in $(0, 1)$ we obtain

Corollary 12. *$\chi_{n+1}(x) < \mu_1 x$ for all $x \in (\rho, x_n)$ and $\chi_{n+1}(x) > \mu_1 x$ for all $x \in (x_n, 1)$.*

Remark 13. This corollary shows that we can find x_n numerically by the method of bisection. Take the first estimate $x_{n,1}$ of x_n as $(\rho+1)/2$. Compute $\chi_j(x_{n,1})$ for $j = 0, \ldots, n+1$. If $\chi_{n+1}(x_{n,1}) > \mu_1 x_{n,1}$ then $x_{n,1}$ must be too large by the corollary, whereas if $\chi_{n+1}(x_{n,1}) < \mu_1 x_{n,1}$, the estimate $x_{n,1}$ must be too small. Based on this result we take for the next estimate, $x_{n,2}$, either $(\rho + x_{n,1})/2$ or $(x_{n,1}+1)/2$, and so on. Clearly, the sequence $\{x_{n,m}\}_{m\ge1}$ converges to x_n.

At this point the computation of x_1, that is, the geometric decay rate when the second station has no waiting room, is very simple indeed. The equation $\chi_2(x) = \mu_1 x$ reduces to

$$\frac{(x-1)\,\left(\mu_2\,\mu_1\,x^2 - \lambda\,a\,x + \lambda^2\right)}{\chi_1(x)} = 0.$$

Since $x_1 \in (\rho, 1)$ we conclude that

$$x_1 = \frac{\lambda\,a}{2\,\mu_1\,\mu_2}\left(1 + \sqrt{1 - \frac{4\,\mu_1\,\mu_2}{a^2}}\right). \tag{26}$$

4.2 Proof of Theorem 2(ii)

As opposed to the previous section, where the blocking threshold n was fixed, in this section the dependence on n plays a central role, since we study the limiting behavior of the sequence of decay rates $\{x_n\}$ when n increases to ∞. We first quote the result from Section 3, complemented with the expressions for the constants α_i, β_i and γ_i, $i = 1, 2$. The subsequent proof rests heavily upon the functions $\eta_\pm(x)$ and $\chi_j(x)$ from Section 4.1, and their properties.

Theorem 2.

(ii) The sequence $\{x_n\}_{n>N(\rho_1,\rho_2)}$ *decreases monotonically to* ρ *and its elements satisfy the bounds*

$$0 < x_n - \rho < \begin{cases} \beta_1\,\gamma_1\,\alpha_1^n, & \text{if } \rho = \rho_1, \\ \beta_2\,\gamma_2\,\alpha_2^n, & \text{if } \rho = \rho_2, \end{cases}$$

where the constants

$$\alpha_1 := \max_{x\in[\rho_1,1]}\left\{\frac{\mu_1\,x}{\eta_+(x)}\right\}, \qquad \alpha_2 := \max_{x\in[\rho_2,1]}\left\{\frac{\eta_-(x)}{\chi_1(x)}\right\},$$

$$\beta_1 := \max_{x\in[\rho_1,1]}\{\mu_1\,x - \eta_-(x)\}, \qquad \beta_2 := \max_{x\in[\rho_2,1]}\{\eta_+(x) - \chi_1(x)\},$$

$$\gamma_1 := \left(\lambda + \mu_1 - \frac{\eta_-(x_1) - \eta_-(\rho_1)}{x_1 - \rho_1}\right)^{-1}, \qquad \gamma_2 := \left(\frac{\eta_+(x_1) - \eta_+(\rho_2)}{x_1 - \rho_2} - \mu_1\right)^{-1},$$

are positive, $\alpha_i < 1$*,* $i = 1, 2$*, and* x_1 *is given by (26).*

The maxima involved do not occur at the boundaries of the intervals but in the interiors, as is clear from Figure 3 for a concrete case. The form of the solutions obtained by taking the derivative with respect to x are cumbersome; we choose not to display these here.

Proof. We first show that $\{x_n\}$ is decreasing, that is, $x_n \notin [x_m, 1)$ whenever $n > m$. By (25) we see that $\chi_{j+1}(x) > \chi_j(x)$ for all $j \geq 0$ and $x \in (\rho, 1)$. Combining this with Corollary 12 for $x \in (x_m, 1)$ and noting that $\chi_{m+1}(x_m) = \mu_1 x_m$ we conclude that for $x \in [x_m, 1)$

$$\chi_{n+1}(x) > \chi_{m+1}(x) \geq \mu_1 x.$$

As no $x \in [x_m, 1]$ can solve the equation $\chi_{n+1}(x) = \mu_1 x$, it must be that $x_n < x_m$.

With regard to the convergence of $\{x_n\}$ to ρ, we consider first the case $\rho = \rho_2$. Let $\delta_n := x_n - \rho_2$, which is positive for all $n > N$. From Lemma 7,

$$\left(\frac{\eta_-(x_n)}{\chi_1(x_n)}\right)^n (\eta_+(x_n) - \chi_1(x_n)) > \eta_+(x_n) - \chi_{n+1}(x_n). \tag{27}$$

As $\eta_+(\cdot)$ is strictly concave on $(\rho_2, 1)$ and $x_n < x_1 < 1$ (for $n > 1$) we can bound $\eta_+(x_n)$ by

$$\eta_+(x_n) > \eta_+(\rho_2) + \frac{\eta_+(x_1) - \eta_+(\rho_2)}{x_1 - \rho_2}\,\delta_n. \tag{28}$$

Therefore, using $\eta_+(\rho_2) = \mu_1\rho_2$ and $\chi_{n+1}(x_n) = \mu_1\,x_n = \mu_1(\rho_2 + \delta_n)$, the right hand side of (27) satisfies,

$$\eta_+(x_n) - \chi_{n+1}(x_n) > \left(\frac{\eta_+(x_1) - \eta_+(\rho_2)}{x_1 - \rho_2} - \mu_1\right)\delta_n.$$

Hence, with (27),

$$\delta_n < \left(\frac{\eta_+(x_1) - \eta_+(\rho_2)}{x_1 - \rho_2} - \mu_1\right)^{-1}\left(\frac{\eta_-(x_n)}{\chi_1(x_n)}\right)^n (\eta_+(x_n) - \chi_1(x_n)),$$

from which the case for $\rho = \rho_2$ of (11) follows.

For $\rho = \rho_1$, let $\delta_n = x_n - \rho_1 > 0$. Clearly, as $\eta_-(\cdot)$ is convex,

$$\eta_-(x_n) < \eta_-(\rho_1) + \frac{\eta_-(x_1) - \eta_-(\rho_1)}{x_1 - \rho_1}\delta_n \tag{29}$$

Therefore, by Lemma 7 and using that $\chi_1(x_n) = (\lambda + \mu_1)(\rho_1 + \delta_n) - \lambda$ and $\eta_-(\rho_1) = \lambda\,\rho_1$, we obtain

$$\begin{aligned}\left(\frac{\chi_{n+1}(x_n)}{\eta_+(x_n)}\right)^n (\chi_{n+1}(x_n) - \eta_-(x_n)) &> \chi_1(x_n) - \eta_-(x_n) \\ &> \left(\lambda + \mu_1 - \frac{\eta_-(x_1) - \eta_-(\rho_1)}{x_1 - \rho_1}\right)\delta_n.\end{aligned}$$

Moreover,

$$\frac{\chi_{n+1}(x_n)}{\eta_+(x_n)} = \frac{\mu_1\,x_n}{\eta_+(x_n)} \leq \max_{x\in[\rho_1,1]}\left\{\frac{\mu_1\,x}{\eta_+(x)}\right\} =: \alpha_1, \tag{30}$$

and, likewise,

$$\chi_{n+1}(x_n) - \eta_-(x_n) \le \max_{x\in[\rho_1,1]} \{\mu_1\, x - \eta_-(x)\} =: \beta_1. \tag{31}$$

The positivity of the constants, except γ_1 and γ_2, as well as the fact that $\alpha_i < 1$, follows from Lemma 10. For γ_2, observe that

$$\frac{\eta_+(x_1) - \eta_+(\rho_2)}{x_1 - \rho_2} - \mu_1 = \frac{\eta_+(x_1) - \eta_+(\rho_2)}{x_1 - \rho_2} - \frac{\eta_+(1) - \eta_+(\rho_2)}{1 - \rho_2} > 0,$$

since η_+ is strictly concave and $\rho_2 < x_1 < 1$. Similar reasoning applies to γ_1. ∎

4.3 Proof of Theorem 3

Below we restate Theorem 3 for convenience. The last equality in part (i) was not in the original statement in Section 3, since there the functions $\chi_j(x)$ were not introduced yet. It explains why we are interested to gain insight into the effect of an increasing blocking threshold n on the values of $\{\chi_j(x_n)\}_{1\le j\le n}$. In Figure 4 we plot the graphs of the sequences $\{\chi_j(x_5)\}_{1\le j\le 5}$, $\{\chi_j(x_{10})\}_{1\le j\le 10}$, and $\{\chi_j(x_{20})\}_{1\le j\le 20}$ for $\rho = \rho_1$ and $\rho = \rho_2$, respectively. To obtain x_5, x_{10} and x_{20} we follow the procedure specified in Remark 13. These graphs suggest that most of the elements of $\{\chi_j(x_n)\}_{1\le j\le n}$ are close to $\eta_-(x_n)$ or $\eta_+(x_n)$ when $\rho = \rho_1$ or $\rho = \rho_2$.

Theorem 3. *For a stable system, the following statements hold.*

(i)

$$\lim_{i\to\infty} \frac{\pi_{i,j+1}^{(n)}}{\pi_{i,j}^{(n)}} = \frac{v_{j+1}^{(n)}}{v_j^{(n)}} = \frac{\chi_{j+1}(x_n)}{\mu_2\, x_n}.$$

(ii) When $\rho = \rho_1$, we have

$$\left| \frac{v_{j+1}^{(n)}}{v_j^{(n)}} - \frac{\lambda}{\mu_2} \right| < \frac{\beta_1}{\mu_2\, \rho_1}\, \alpha_1^{n-j} \;+\; \frac{(\mu_1 - \gamma_1^{-1})\, \beta_1\, \gamma_1}{\mu_2\, \rho_1}\, \alpha_1^n,$$

and when $\rho = \rho_2$,

$$\left| \frac{v_{j+1}^{(n)}}{v_j^{(n)}} - \frac{\mu_1}{\mu_2} \right| < \frac{\beta_2}{\lambda}\, \alpha_2^j \;+\; \frac{1-\rho_2}{1-\rho_1}\, \frac{\beta_2\, \gamma_2}{\rho_2}\, \alpha_2^n,$$

where the constants $\alpha_i, \beta_i, \gamma_i$ are as defined earlier in Section 4.2.

Proof. Statement (i) is immediate from (7) and (17b).

For (ii) we first prove the result for $\rho = \rho_2$. Observe that by the triangle inequality and the inequality $\mu_2\, x_n > \lambda$,

$$\begin{aligned} \left| \frac{v_{j+1}^{(n)}}{v_j^{(n)}} - \frac{\mu_1}{\mu_2} \right| &= \left| \frac{\chi_{j+1}(x_n)}{\mu_2\, x_n} - \frac{\eta_+(\rho_2)}{\mu_2\, \rho_2} \right| \\ &= \frac{|\rho_2\, \chi_{j+1}(x_n) - x_n\, \eta_+(\rho_2)|}{\mu_2\, x_n\, \rho_2} \\ &< \frac{|\chi_{j+1}(x_n) - \eta_+(x_n)|}{\lambda} + \frac{|\rho_2\, \eta_+(x_n) - x_n\, \eta_+(\rho_2)|}{\lambda\, \rho_2}. \end{aligned} \tag{32}$$

Clearly, by applying the second statement of Lemma 7 to $\eta = \chi_1(x_n)$, we have

$$0 < \eta_+(x_n) - \chi_{j+1}(x_n) < \left(\frac{\eta_-(x_n)}{\chi_1(x_n)} \right)^j (\eta_+(x_n) - \chi_1(x_n)) \,.$$

For the second term, we observe that $\eta_+(x_n) > x_n\, \mu_1$, since η_+ is strictly concave and $\eta_+(x) = \mu_1 x$ for $x = \rho_2, 1$. Hence, $\rho_2\, \eta_+(x_n) > \rho_2\, x_n\, \mu_1 = x_n\, \eta_+(\rho_2)$, so that,

$$\eta_+(x_n) < \eta_+(\rho_2) + \delta_n\, \eta_+'(\rho_2) = \eta_+(\rho_2) + \delta_n\, \frac{\mu_1 + \mu_2 - 2\lambda}{1 - \rho_1}, \tag{33}$$

where we recall that $\delta_n = x_n - \rho_2$. Hence, after some calculations,

$$0 < \rho_2\, \eta_+(x_n) - x_n\, \eta_+(\rho_2) < \big(\rho_2\, \eta_+'(\rho_2) - \eta_+(\rho_2)\big)\, \delta_n = \lambda \frac{1-\rho_2}{1-\rho_1}\, \delta_n.$$

The rest follows immediately from Theorem 2(ii).

When $\rho = \rho_1$, so that $x_n > \rho_1$, consider

$$\left|\frac{v_{j+1}^{(n)}}{v_j^{(n)}} - \frac{\lambda}{\mu_2}\right| = \left|\frac{\chi_{j+1}(x_n)}{\mu_2\, x_n} - \frac{\eta_-(\rho_1)}{\mu_2 \rho_1}\right|$$
$$< \frac{|\chi_{j+1}(x_n) - \eta_-(x_n)|}{\mu_2\, \rho_1} + \frac{|\rho_1\, \eta_-(x_n) - x_n\, \eta_-(\rho_1)|}{\mu_2\, \rho_1^2}.$$

For the first term we apply the third statement of Lemma 7 with $\eta = \chi_{n+1}(x_n)$ and $i = n - j$, to find

$$\chi_{j+1}(x_n) - \eta_-(x_n) < \left(\frac{\chi_{n+1}(x_n)}{\eta_+(x_n)}\right)^{n-j} (\chi_{n+1}(x_n) - \eta_-(x_n)),$$

after which we only need to apply (30) and (31). For the second term we use $x_n = \rho_1 + \delta_n$ and $\eta_-(\rho_1) = \lambda\rho_1$ to arrive at

$$\frac{|\eta_-(x_n) - \eta_-(\rho_1) - \lambda\delta_n|}{\mu_2\rho_1}.$$

Since $\eta_-(x)$ is convex, and $\eta_-'(\rho_1) > \lambda$, the absolute signs are not needed, so that we can arrive at our result using (29) and the fact that $\delta_n < \beta_1\gamma_1\, \alpha_1^n$. ∎

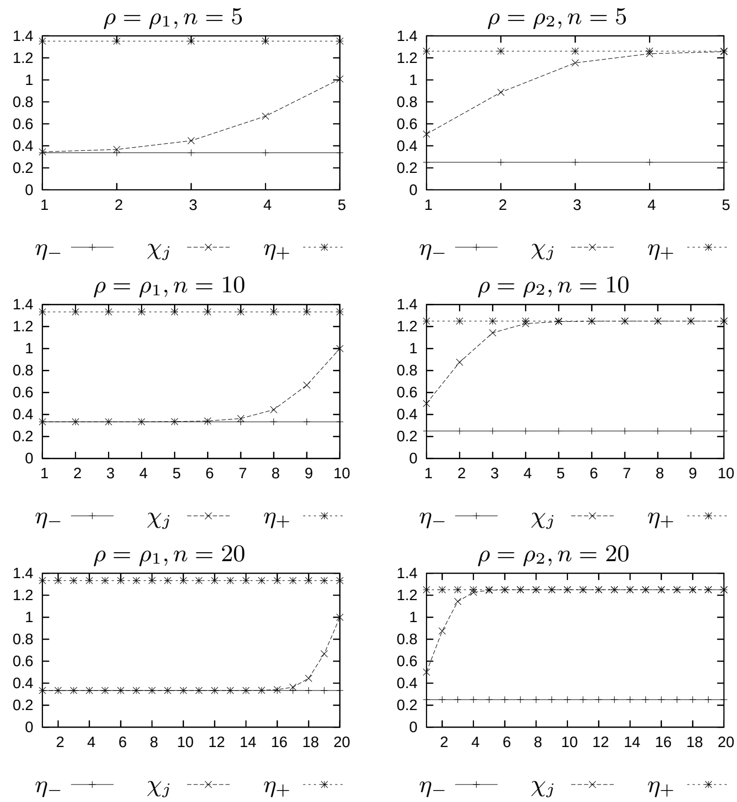

Figure 4: Graphs of the sequence $\{\chi_j(x_n)\}_{1\le j\le n}$ for $n = 5$, 10, and 20. At the left $\rho = \rho_1$ ($\lambda = 1$, $\mu_1 = 3$ and $\mu_2 = 4$), whereas at the right $\rho = \rho_2$ ($\lambda = 1$, $\mu_1 = 5$ and $\mu_2 = 4$). The phase j increases along the x-axis; the value of $\chi_j(x_n)$ is set out along the y-axis. For clarity we connect subsequent terms $\chi_j(x_n)$ by lines.

5 The Tandem Queue with Slow-down and Blocking

Consider now a network in which the second server signals the first to slow down, i.e., to work at rate $\tilde{\mu}_1 < \mu_1$ instead of at rate μ_1, when the second station contains m or more jobs, where, of course, $m < n$. Figure 5 shows the state transition diagram of the resulting queueing process.

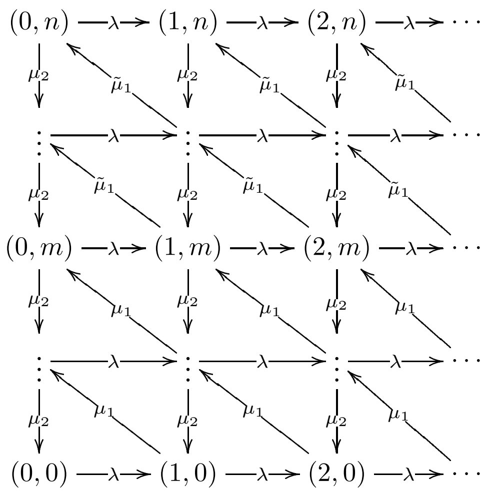

Figure 5: State space and transition rates of the two-station tandem queue with slow-down and blocking. Note that above phase m server 1 works at rate $\tilde{\mu}_1$ rather than at μ_1.

In this section we assume the following ordering of parameters:

$$\lambda < \mu_2 < \tilde{\mu}_1 < \mu_1, \text{ or, equivalently, } \rho_1 < \tilde{\rho}_1 < \rho_2 < 1, \tag{34}$$

where $\tilde{\rho}_1 := \lambda/\tilde{\mu}_1$. Observe that as a consequence, $\rho = \rho_2$ in this section. Henceforth we do no longer use ρ, but always ρ_2. With this ordering we generalize Theorem 2(i) to the present case and restate Theorems 2(ii) and 3 in somewhat weaker form. The methods of proof are similar to those of Section 4. Due to these similarities we only show the main steps to arrive at the results stated here. The details may sometimes be slightly more involved algebraically, but are seldom more complicated conceptually.

Remark 14. It would, of course, be interesting to consider other orderings of the system parameters such as, for instance, $0 < \tilde{\mu}_1 < \lambda < \mu_2 < \mu_1$. However, Lemma 16 below does not immediately carry over to these cases as its proof depends crucially on the ordering (34). We conjecture, based on numerical experiments, that similar results can be obtained for all cases. Thus, 'case checking', i.e., proving each step of the line of reasoning below for every possible ordering of parameters (provided the chain is stable), seems a possible method to obtain stronger results. However, this approach is, admittedly, not elegant, neither might it reveal much of the structure of the problem. It remains for further research to find the general underlying principle; here we concentrate on the ordering specified in (34).

Since $\tilde{\mu}_1 < \mu_1$ we can again uniformize the related continuous-time Markov chain $\{X_1^{(n,m)}(t), X_2^{(n,m)}(t)\}$ at rate $a = \lambda + \mu_1 + \mu_2$ to obtain an aperiodic discrete-time QBD chain $\{X_{1,k}^{(n,m)}, X_{2,k}^{(n,m)}\}$. The matrix of transition probabilities $P^{(n,m)}$ has the same form as $P^{(n)}$ in (2), but whereas $B^{(n,m)} = B^{(n)}$ and $A_0^{(n,m)} = A_0^{(n)}$, $A_1^{(n,m)}$ becomes, with $\tilde{q} = \tilde{\mu}_1/a$,

$$A_1^{(n,m)} = \begin{pmatrix} r & & & & & & & \\ r & 0 & & & & & & \\ & \ddots & \ddots & & & & & \\ & & r & 0 & & & & \\ & & & r & q-\tilde{q} & & & \\ & & & & \ddots & \ddots & & \\ & & & & & r & q-\tilde{q} & \\ & & & & & & r & q \end{pmatrix}, \tag{35}$$

where at the m-th row the changes occur, and $A_2^{(n,m)}$ has the same form as $A_2^{(n)}$, however $\tilde{q}$ replaces q in rows $m, \dots, n-1$. Finally, let $A^{(n,m)}(x) := A_0^{(n,m)} + xA_1^{(n,m)} + x^2 A_2^{(n,m)}$.

Concerning the stability of the chain we follow the approach of Theorem 1 to derive a necessary and sufficient stability condition. In accordance with our expectations for a system with the ordering (34), this condition reduces to $\lambda < \mu_2$ when $n \to \infty$.

Theorem 15. *Let $\beta = \mu_1/\mu_2$ and $\tilde\beta = \tilde\mu_1/\mu_2$. The two-station tandem network with slow-down threshold m and blocking at $n \geq m$ is positive recurrent if and only if*

$$\lambda < \frac{\mu_1\,(1-\beta^m)(1-\tilde\beta) + \tilde\mu_1\,\beta^m\,(1-\beta)(1-\tilde\beta^{n-m})}{(1-\beta^m)(1-\tilde\beta) + \beta^m\,(1-\beta)(1-\tilde\beta^{n-m+1})}. \tag{36}$$

Proof. Following the proof of Theorem 1, the normalized solution of $\boldsymbol{\alpha} A^{(n,m)}(1) = \boldsymbol{\alpha}$ has the form,

$$\alpha_i = \begin{cases} \alpha_0\beta^i, & \text{if } i \leq m-1, \\ \alpha_0\beta^m\tilde\beta^{i-m}, & \text{if } m \leq i \leq n, \end{cases}$$

and

$$\alpha_0^{-1} = \frac{1-\beta^m}{1-\beta} + \beta^m\frac{1-\tilde\beta^{n+1-m}}{1-\tilde\beta}.$$

The inequality $\boldsymbol{\alpha} A_0^{(n,m)}\mathbf{e} < \boldsymbol{\alpha} A_2^{(n,m)}\mathbf{e}$ becomes

$$\lambda < \alpha_0\left(\mu_1\sum_{i=0}^{m-1}\beta^i + \tilde\mu_1\beta^m\sum_{i=m}^{n-1}\tilde\beta^{i-m}\right) = \alpha_0\left(\mu_1\frac{1-\beta^m}{1-\beta} + \tilde\mu_1\beta^m\frac{1-\tilde\beta^{n-m}}{1-\tilde\beta}\right).$$

∎

The next step is to rewrite the equation

$$\mathbf{v}^{(n,m)}A^{(n,m)}(x) = \mathbf{v}^{(n,m)}x, \tag{37}$$

and derive a sequence $\{\chi_j(x\}_{1\leq j\leq n}$ in terms of mappings similar to T defined in (19). With this aim, let $\chi_j(x) = \mu_2\,x\,v_j/v_{j-1}$ as in (17b). However, contrary to (18) we now need *three*, rather than one, mappings to cast (37) into a sequence $\{\chi_j(x)\}_{1\leq j\leq n}$, namely T as in (19), and

$$S:\eta\mapsto \tilde a x - \lambda - \frac{\mu_1\mu_2x^3}{\eta}, \qquad \tilde T:\eta\mapsto \tilde a x - \lambda - \frac{\tilde\mu_1\mu_2x^3}{\eta}, \tag{38}$$

where $\tilde a = \lambda + \tilde\mu_1 + \mu_2$. Again setting $v_0 = 1$ and introducing $\chi_0(x)$ and $\chi_{n+1}(x)$ for convenience, we have

$$\chi_j(x) := \begin{cases} \mu_1x^2, & \text{if } j = 0, \\ T\left(\chi_{j-1}(x)\right), & \text{if } 1 \leq j \leq m, \\ S\left(\chi_m(x)\right), & \text{if } j = m+1, \\ \tilde T\left(\chi_{j-1}(x)\right), & \text{if } m+2 \leq j \leq n+1. \end{cases}$$

Loosely speaking, S moves $\chi_m(x)$ across the slow-down threshold at m to the iterate $\chi_{m+1}(x)$ on which $\tilde T$ can start operating. The condition on x of the last coordinate of the vector equation $\mathbf{v}^{(n,m)}A^{(n,m)}(x) = \mathbf{v}^{(n,m)}x$ is,

$$\chi_{n+1}(x) = \tilde\mu_1 x, \tag{39}$$

rather than $\chi_{n+1}(x) = \mu_1 x$ as in (18c).

Theorem 5 carries over immediately. Thus, if we can find $x \in (0,1)$ such that each element of the sequence $\{\chi_j(x)\}_{0\leq j\leq n+1}$ is positive and $\chi_{n+1}(x) = \tilde\mu_1 x$, then x is the decay rate we are searching for.

To establish that the elements of $\{\chi_j(x)\}_{0\leq j\leq n+1}$ are positive we would like to apply Lemma 10. Supposing that $\chi_0(x) \in (\eta_-(x), \eta_+(x))$, it follows that the elements of $\{\chi_j(x)\}_{0\leq j\leq m}$ all lie in the interval $(\eta_-(x), \eta_+(x))$, hence are positive. However, it is not immediately obvious that $S\left(\chi_m(x)\right)$ lies somewhere in between the fixed points $\tilde\eta_-(x)$ and $\tilde\eta_+(x)$ (regarded as functions of x) of $\tilde T$. Now realize that $\chi_0(x) < \chi_m(x) < \eta_+(x)$, and therefore by (38), that $S\left(\chi_0(x)\right) < S\left(\chi_m(x)\right) < S\left(\eta_+(x)\right)$. Below we prove that $\tilde\eta_-(x) < S\left(\chi_0(x)\right)$ and $S\left(\eta_+(x)\right) \leq \tilde\eta_+(x)$ so that S maps any element in $(\chi_0(x), \eta_+(x))$, and *in particular* $\chi_m(x)$, into the interval $(\tilde\eta_-(x), \tilde\eta_+(x))$. Therefore, Lemma 10, which applies to equally well to $\tilde T$ due to the ordering (34), ensures that also the elements of $\{\chi_j(x)\}_{m+2\leq j\leq n+1}$ lie within the interval $(\tilde\eta_-(x), \tilde\eta_+(x))$. Finally, due to the ordering (34) Lemma 9 implies that $\tilde\eta_-(x) > 0$ for $x \in [\rho_2, 1]$, thereby guaranteeing the positivity of all elements of the sequence $\{\chi_j(x)\}_{0\leq j\leq n+1}$ for $x \in [\rho_2, 1]$.

Lemma 16. *For all $x \in (\rho_2, 1)$:*

$$\tilde\eta_-(x) < S\left(\chi_0(x)\right) \quad \textit{and} \quad S\left(\eta_+(x)\right) \leq \tilde\eta_+(x). \tag{40}$$

Proof. Let us start with proving the first inequality. As $\lambda < \mu_2 < \tilde{\mu}_1$ it follows from Lemma 9 that $\tilde{\eta}_-(x) < \tilde{\mu}_1 x^2$. Hence, $\mu_1 \tilde{\eta}_-(x)/\tilde{\mu}_1 < \mu_1 x^2 = \chi_0(x)$. Applying S to both sides and noting that $S(\mu_1 \tilde{\eta}_-(x)/\tilde{\mu}_1) = \tilde{T}(\tilde{\eta}_-(x)) = \tilde{\eta}_-(x)$ gives the result.

Concerning the second inequality in (40) observe that this is equivalent to

$$\eta_+(x) + (\tilde{\mu}_1 - \mu_1)x = S(\eta_+(x)) \leq \tilde{\eta}_+(x). \tag{41}$$

Clearly, in case $\tilde{\mu}_1 = \mu_1$, the left hand side and the right hand side are equal. Next, if the derivative with respect to $\tilde{\mu}_1$ of the left hand side of (41) is larger than the derivative of the right hand side then, as $\tilde{\mu}_1 < \mu_1$, the inequality must hold.

Thus, we like to show that when $x \in (\rho_2, 1)$,

$$x > \frac{\partial \tilde{\eta}_+(x)}{\partial \tilde{\mu}_1} = \frac{x}{2} + \frac{1}{2}\frac{(\tilde{a}x - \lambda)x - 2\mu_2 x^3}{\sqrt{(\tilde{a}x-\lambda)^2 - 4\tilde{\mu}_1\mu_2 x^3}}.$$

Rewrite this to

$$\sqrt{(\tilde{a}x-\lambda)^2 - 4\tilde{\mu}_1\mu_2 x^3} > \tilde{a}x - \lambda - 2\mu_2 x^2.$$

This inequality is implied by

$$(\tilde{a}x-\lambda)^2 - 4\tilde{\mu}_1\mu_2 x^3 > (\tilde{a}x-\lambda)^2 - 4\mu_2 x^2(\tilde{a}x-\lambda) + 4\mu_2^2 x^4,$$

which in turn reduces to

$$\lambda(x-1) > \mu_2 x(x-1).$$

This is true since $x \in (\rho_2, 1)$. ∎

As counterpart of Theorem 2 we obtain the following.

Theorem 17. *If $\rho_1 < \tilde{\rho}_1 < \rho_2 < 1$ and the blocking threshold n and slow-down threshold $m \leq n$ are such that the chain $\{X_{1,k}^{(n,m)}, X_{2,k}^{(n,m)}\}$ is stable, the sequence $\{x_{n,m}\}_n$ decreases monotonically to ρ_2 for m fixed.*

Proof. The positivity of the elements of $\{\chi_j(x_{n,m}\}_{1\leq j\leq n+1}$ is settled by the discussion leading to Lemma 16.

To prove that there exists a unique $x \in (\rho_2, 1)$ such that $\chi_{n+1}(x) = \tilde{\mu}_1 x$, we reason an in the proof of Theorem 2(i) in Section 4.1.3. Observe that: (i) $\chi_0(\rho_2) < \eta_+(\rho_2) \Rightarrow \chi_j(\rho_2) < \tilde{\eta}_+(\rho_2) = \tilde{\mu}_1\rho_2$ for all $j > m$; (ii) $\chi_{n+1}(1) = \tilde{\mu}_1$; (iii) Condition (36) is equivalent to $\chi'_{n+1}(1) < \tilde{\mu}_1$; (iv) $\chi''_{n+1}(x) < 0$, i.e., $\chi_{n+1}(x)$ is strictly concave, for $x \in (\rho_2, 1)$.

By similar reasoning as in the first part of the proof of Theorem 2(ii) it can be seen that $\{x_{n,m}\}$ decreases monotonically. Finally, pertaining to the convergence to ρ_2, the sequence $\{x_{n,m}\}$, being bounded and decreasing, has a unique limit point ζ in $\mathbb{R}$. Suppose that $\zeta > \rho_2$. Then, since, $\tilde{\eta}_+(\zeta) > \tilde{\mu}_1\zeta$ and $\lim_{j\to\infty}\chi_j(x) = \tilde{\eta}_+(x)$ for all $x \in (\rho_2, 1)$, there exists $M > 0$ such that for all $j > M$, $\chi_j(\zeta) > \tilde{\mu}_1\zeta$. On the other hand, we derived above that $\chi_j(\rho_2) < \tilde{\mu}_1\rho_2$ for $j > m$. The continuity of $\chi_j(x)$ implies that there exists $x_{j-1} \in (\rho_2, \zeta)$ such that $\chi_j(x_{j-1}) = \tilde{\mu}_1 x_{j-1}$. This contradicts $\zeta > \rho_2$. ∎

It proves difficult to bound the rate of convergence of the sequence of decay rates $\{x_{n,m}\}$, which thereby prevents us from generalizing (11) to the present case. As a result, we also cannot carry over Theorem 3. However, we can achieve the following slightly weaker result in which we appropriately scale the slow-down threshold m as a function of the blocking threshold n.

Theorem 18. *Let the slow-down threshold m scale as $m(n) = \alpha n$ for a fixed $\alpha \in (0,1)$ and write $\pi^{(n,m)}(i,j)$ for $\pi_{i,j}^{(n,m)}$. Then,*

$$\lim_{n\to\infty}\lim_{i\to\infty}\frac{\pi^{(n,m)}(i,\lfloor yn\rfloor)}{\pi^{(n,m)}(i,\lfloor yn\rfloor-1)} = \begin{cases} \frac{\eta_+(\rho_2)}{\mu_2\rho_2} = \frac{\mu_1}{\mu_2}, & \textit{if } y \in (0,\alpha], \\ \frac{\tilde{\eta}_+(\rho_2)}{\mu_2\rho_2} = \frac{\tilde{\mu}_1}{\mu_2}, & \textit{if } y \in (\alpha,1), \end{cases} \tag{42}$$

where $\lfloor x\rfloor$ denotes the largest integer smaller than or equal to x.

In Theorem 3 we could bound this ratio for any *fixed* phase j, $j \leq n$, for $n \to \infty$. Here we scale the phase $j(n)$ as a function of n. In fact, the proof below makes clear that we establish the point-wise limit of the functions $\chi_{j(n)}(x_{n,m})/\mu_2 x_{n,m}$ for $n \to \infty$ rather than for j fixed.

Proof. Recall that

$$\lim_{i\to\infty} \frac{\pi^{(n,m)}(i, \lfloor yn \rfloor)}{\pi^{(n,m)}(i, \lfloor yn \rfloor - 1)} = \frac{v^{(n,m)}(\lfloor yn \rfloor)}{v^{(n,m)}(\lfloor yn \rfloor - 1)} = \frac{\chi_{\lfloor yn \rfloor}(x_{n,m})}{\mu_2\, x_{n,m}},$$

and concentrate on the right hand side.

First, let $y \in (0, \alpha]$. Clearly, it follows from Theorem 17 that $x_{n,m} \to \rho_2$ for $n \to \infty$, and therefore, by applying Lemma 10, $\chi_{\lfloor yn \rfloor}(x_{n,m}) \to \eta_+(\rho_2)$. In particular, $\chi_{\lfloor \alpha n \rfloor}(x_{n,m}) \to \eta_+(\rho_2) = \mu_1\,\rho_2$ so that, by (38),

$$\lim_{n\to\infty} S\left(\chi_{\lfloor \alpha n \rfloor}(x_{n,m})\right) = \tilde{a}\rho_2 - \lambda - \frac{\mu_1\mu_2\rho_2^3}{\eta_+(\rho_2)} = \tilde{\mu}_1\,\rho_2 = \tilde{\eta}_+(\rho_2).$$

Now let $y \in (\alpha, 1)$. As $S\left(\chi_{\lfloor \alpha n \rfloor}(x_{n,m})\right) < \chi_{\lfloor yn \rfloor}(x_{n,m}) < \tilde{\eta}_+(x_{n,m})$, and the left and right hand side converge to $\tilde{\eta}_+(\rho_2)$ for $n \to \infty$, the functions $\chi_{\lfloor yn \rfloor}(x_{n,m})$ have the same limit. ∎

In terms of the Perron-Frobenius vector $\mathbf{v}^{(n,m)}$ of $R^{(n,m)}$ this results means the following,

$$\frac{v_j^{(n,m)}}{v_{j-1}^{(n,m)}} \approx \begin{cases} \mu_1/\mu_2 & \text{if } j < m(n) \\ \tilde{\mu}_1/\mu_2 & \text{if } j \geq m(n). \end{cases}$$

Thus, a 'kink' appears in the graph of ratio of the consecutive components of $\mathbf{v}^{(n,m)}$.

Remark 19. The approach to prove the results in this section generalizes to any number of slow-down thresholds when the adapted rates $\tilde{\mu}_1, \tilde{\tilde{\mu}}_1, \ldots$, form a decreasing sequence bounded below by μ_2.

6 Acknowledgments

The authors are grateful to Erik van Doorn for helpful discussions. The authors also thank two anonymous referees for their suggestions to further improve the presentation of the results.

References

[1] W. K. Grassman and S. Drekic. An analytical solution for a tandem queue with blocking. *Queueing Systems*, 36:221–235, 2000.

[2] A.G. Konheim and M. Reiser. A queuing model with finite waiting room and blocking. *Journal of the ACM*, 23(2):328–341, 1976.

[3] A.G. Konheim and M. Reiser. Finite capacity queuing systems with applications in computer modeling. *SIAM Journal of Computing*, 7(2):210–229, 1978.

[4] D.P. Kroese, W.R.W. Scheinhardt, and P.G. Taylor. Spectral properties of the tandem Jackson network, seen as a quasi-birth-and-death process. *Annals of Applied Probability*, 14(4):2057-2089, 2004.

[5] P. Lancaster and M. Tismenetsky. *The Theory of Matrices with Applications*. Academic Press, 2nd edition, 1985.

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

[7] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. SIAM, 1999.

[8] L. Leskelä. Stabilization of an overloaded queueing network using measurement-based admission control. Technical Report A470, Department of Engineering, Physics and Mathematics, Helsinki University of Technology, 2004.

[9] C.D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, 2000.

[10] T. Needham. *Visual Complex Analysis*. Oxford University Press Inc., 2000.

[11] M. F. Neuts. *Matrix-Geometric Solutions in Stochastic Models, An Algorithmic Approach*. Dover Publ., Inc., 1981.

[12] M.F. Neuts. The caudal characteristic curve of queues. *Advances in Applied Probability*, 18:221–254, 1986.

[13] W. Noureddine and F. Tobagi. Selective back-pressure in switched Ethernet LANs. *Proceedings of Globecom '99*, 1256–1263, 1999.