Simple bounds and monotonicity results for finite multi-server exponential tandem queues

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Faculty of Mathematics and Computing Science

Memorandum COSOR 87-37

Simple bounds and monotonicity results for finite multi-server exponential tandem queues

by

N.M. van Dijk and J. van der Wal

Revised June 1988

Eindhoven, The Netherlands

July 1987

SIMPLE BOUNDS AND MONOTONICITY RESULTS FOR FINITE MULTI-SERVER EXPONENTIAL TANDEM QUEUES

N.M. van Dijk [1] *and J. van der Wal* [2]

ABSTRACT

Simple and computationally attractive lower and upper bounds are presented for the call congestion such as those representing multi-server loss or delay stations. Numerical computations indicate a potential usefulness of the bounds for quick engineering purposes. The bounds correspond to product-form modifications and are intuitively appealing. A formal proof of the bounds and related monotonicity results will be presented. The technique of this proof, which is based on Markov reward theory, is of interest in itself and seems promising for further application. The extension to the non-exponential case is discussed. For multi-server loss stations the bounds are argued to be insensitive.

Keywords

Tandem queues, product form, call congestion, bounds, monotonicity results, Markov chains.

[1] Free University, Department of Economics, Postbox 7161, 1007 MC Amsterdam.
[2] Eindhoven University of Technology, Department of Mathematics and Computing Science, Postbox 513, 5600 MB Eindhoven.

1. Introduction

It is well-known that tandem queues with restricted accommodation do not exhibit the celebrated product form, see e.g. Bhat [3] and Lipper and Sengupta [11]. Therefore both numerical and approximate methods have been widely studied. These however may be computational expensive and the accuracy of the approximations is not guaranteed. Sometimes one may just be interested in conservative but secured bounds which are easy to obtain to get a fast impression of the system performance.

This paper further explores the methodology introduced in Van Dijk and Lamond [6] to obtain computationally attractive bounds for finite single-server exponential tandem queues. Here the results will be extended to general service capacities to allow for multi server-loss and delay stations. We obtain simple bounds for the call congestion or the throughput of finite tandem queues. These bounds correspond to product-form modifications of the original tandem queue for which it is intuitively clear that they yield an upper and a lower bound.

We provide a formal proof of these bounds based on Markov reward theory which is of interest in itself.

Numerical calculations are given and indicate a practical usefulness as reasonable first indicator of the performance.

The bounds can be seen as monotonicity properties of queueing networks. This topic currently receives considerable attention.

Basically there seem to be four approaches to establish monotonicity results. An analytic one based on comparison of one-step transition operators (cf. e.g. Stoyan [15], Whitt [19] and Hordijk and Ridder [10]). A second approach is to study the equilibrium distribution itself, e.g. in the product-form case (cf. Robertazzi and Lazar [12], Suri [16] and Yao [21]). The third technique is a probabilistic one using coupling and sample path comparison and makes it possible to treat other than exponential or phase type distributions. This seems to be the most powerful approach although the method failed at unexpected moments. See e.g. Adan and Van der Wal [1,2] Van Dijk, Tsoucas and Walrand [7], Shanthikumar and Yao [13,14] and Tsoucas and Walrand [17]. The fourth method is the one used in this paper based on Markov reward theory, cf. Van Dijk [4,5] and Van Dijk and Lamond [6] and Van der Wal [18].

The paper is organized as follows. In section 2 we present the model and formulate the bounds. Preliminary results needed to prove the bounds are given in section 3. The details of the formal proofs for the lower and upper bound are given in sections 4 and 5. Section 6 contains some numerical results. Section 7 discusses some extensions, in particular to the non-exponential case.

2. Model and bounds

The original model.

Jobs arrive at a two node (stage) tandem queue according to a Poisson process with parameter λ. At each node there is a constraint on the number of jobs which can be allowed at a time, say M at node 1 and N at node 2. Let m and n denote the number of jobs at nodes 1 and 2 respectively. An arriving job is rejected (and lost) if node 1 is saturated (i.e., if $m = M$). A job which completes a service at node 1 while node 2 is saturated (i.e., while $n = N$) is not accepted at node 2 and has to restart a complete new service at node 1. Otherwise an arriving job is accepted at node 1 (respectively node 2) at which it requests an amount of service. The total service capacity (that is, the rate at which jobs are being served) at node 1 is given by $\Phi(m)$ and at node 2 by $\Psi(n)$. The natural assumption is made that $\Phi(m)$ and $\Psi(n)$ are non-decreasing in m and n while $\Phi(0) = \Psi(0) = 0$. Further, all service requirements including the repetitions at node 1, are assumed to be independent and exponentially distributed with means μ^{-1} at node 1 and ν^{-1} at node 2. The description above includes service stations such as:

a) multi-server loss stations with M and N available servers as by $\Phi(m) = m$, $\Psi(n) = n$ for all m, n;

b) multi-server delay (e.g. FCFS-) stations with $r < M$ and $s < N$ servers by $\Phi(m) = m$ for $m \le r$, $\Phi(m) = \Phi(r)$ for $r < m \le M$ and $\Psi(n) = n$ for $n \le s$, $\Psi(n) = \Psi(s)$ for $s < n \le N$;

c) single-server processor sharing stations by $\Phi(m) = 1$, $\Psi(n) = 1$ for all m, n.

Furthermore, due to the memoryless property of exponential services the repeating protocol at node 1 can also be interpreted as if node 1 stops serving as long as node 2 is saturated, which is the standard protocol in communication systems.

We will now consider two modifications of the original model. These modifications, called "lower bound" and "upper bound" model, differ from the original model only in their blocking protocol in the way described below.

The lower bound model.

An arriving job is rejected (and lost) only if the total number of jobs $m + n$ already present is equal to $M + N$. A job which completes its service at node 1 is always instantly accepted at node 2. Either node can accommodate $M + N$ jobs. The service capacities $\Phi(m)$ and $\Psi(n)$ are equal to that of the original model for $m \le M$ and $n \le N$ but can be chosen to be any non-decreasing function for $M \le m \le M + N$ and $N \le n \le M + N$.

The upper bound model.

An arriving job is rejected (and lost) not only if the first node is saturated ($m = M$) but also when the second node is saturated ($n = N$). Furthermore, a job, which completes its service at node 1 (respectively 2) while the other node is saturated, has to restart a complete new service at that node. The service capacity $\Phi(m)$ and $\Psi(n)$ correspond to that of the original model for any $m \leq M$ and $n \leq N$.

Intuitively, in the lower bound model the total number $M + N$ of available places is used more efficiently since these can be allocated to either node in a dynamic manner wherever needed. We may thus expect a less frequent rejection of arriving jobs. Conversely, upon arrival of a job the upper bound model is seen fully congested also if node 2 is saturated. On top of that, the service repetitions at node 2 cause a higher congestion at both nodes. We may thus expect a more frequent rejection of arriving jobs.

The corresponding state spaces of the lower bound (L), original (O) and upper bound (U) model become

$$R_L = \{(m, n) \mid m + n \leq M + N\} \ ,$$

$$R_O = \{(m, n) \mid m \leq M, \ n \leq N\} \text{ and}$$

$$R_U = \{(m, n) \mid m \leq M, \ n \leq N, \ m + n \neq M + N\}$$

and the blockings for the different models are summarized in Table 1 below.

Model (L, O, U)	Arrival at node 1	Transition from node 1 to 2	Departure at node 2	State space
Lower	$m + n = M + N$	--	--	R_L
Original	$m = M$	$n = N$	--	R_0
Upper	$m = M$ ór $n = N$	$n = N$	$m = M$	R_U

Table 1

Let $\pi_L(n_1, n_2)$ and $\pi_U(n_1, n_2)$ be the steady state probability of state (m, n) for the lower and upper bound model respectively. Then from Hordijk and Van Dijk [8], as based upon a so-called notion of "centre-local-balance", we can conclude that with c_L and c_U normalizing constants and $\prod_{k=1}^{0} \Phi(k) = \prod_{k=1}^{0} \Psi(k) = 1$,

$$\pi_L(m, n) = c_L(\frac{\lambda}{\mu})^m(\frac{\lambda}{\nu})^n[\prod_{k=1}^{m}\Phi(k)]^{-1}[\prod_{k=1}^{n}\Psi(k)]^{-1} \quad , (m, n) \in R_L$$
$$\pi_U(m, n) = c_U(\frac{\lambda}{\mu})^m(\frac{\lambda}{\nu})^n[\prod_{k=1}^{m}\Phi(k)]^{-1}[\prod_{k=1}^{n}\Psi(k)]^{-1} \quad , (m, n) \in R_U \tag{2.1}$$

It may be noted that the stationarity of these probabilities can also be verified easily by substitution in the global balance equations. However, it has essentially been this notion of "centre-local-balance" which has led to the above product-form modifications as potential candidates in the first place.

Let B denote the call congestion of the original model that is

$$B = \text{the steady state probability that an arriving job is rejected} \ ,$$

and define B_L and B_U similarly for the lower and upper bound model. By virtue of the "Poisson arrivals see time averages" - theorem (cf. Wolff [20]), B_L and B_U can thus be calculated by PASTA

$$B_L = \pi_L(m + n = M + N), \quad B_U = \pi_U(m = M \text{ or } n = N) \ . \tag{2.2}$$

The probabilities B_L and B_U are easily obtained from (2.1).

Hence, by proving that

$$B_L \le B \le B_U \tag{2.3}$$

we will thus have established computationally attractive bounds for the call congestion B as well as the throughput T, via the relation $T = \lambda(1 - B)$.

The inequalities (2.3) will be formally established in sections 3, 4 and 5.

3. Preliminary results

We now describe some preliminary results for Markov reward theory in order to prove inequality (2.3). For convenience, we introduce the notation

$$\mu(m) = \mu\Phi(m), \quad \nu(n) = \nu\Psi(n) \ .$$

We use the subindex α to indicate the α-model, that is the lower bound ($\alpha = L$), original ($\alpha = O$) and upper bound ($\alpha = U$) model.

We first note that due to the exponentiality assumptions, the underlying queueing process of the α-model constitutes an irreducible finite Markov jump process (or continuous-time

Markov chain), say with jumprate (or infinitesimal generator) matrix G_α. Without restriction of generality assume that $\lambda + \mu(M+N) + \nu(M+N) \leq 1$, which can always be achieved by scaling the time scale. The discrete time Markov chain with one-step transition matrix defined by $P_\alpha = (I + G_\alpha)$, then has exactly the same stationary probability (row-) vector π_α as the above Markov jump process, which for either case is uniquely determined, up to normalization, by $\pi_\alpha G_\alpha = 0$. For analyzing the call congestion we may thus restrict the attention to the discrete-time Markov chain with transition matrix P_α.

With the α-model we thus associate a discrete-time Markov chain with one-step transition probabilities $p_\alpha(m, n; m', n')$ for a transition from state (m, n) into state (m', n') given by:

$$p_\alpha(m, n; m+1, n) = \lambda \delta_\alpha^1(m, n)$$

$$p_\alpha(m, n; m-1, n+1) = \mu(m)\, \delta_\alpha^2(m, n)$$

$$p_\alpha(m, n; m, n-1) = \nu(n)\, \delta_\alpha^3(m, n)$$

and

$$p_\alpha(m, n; m, n) = 1 - \lambda \delta_\alpha^1(m, n) - \mu(m)\, \delta_\alpha^2(m, n) - \nu(n)\, \delta_\alpha^3(m, n)$$

where $\delta_\alpha^i(m, n)$ is equal to 1 if that transition is possible for the α-model and equal to 0 otherwise, $i = 1, 2, 3$. (e.g., $\delta_U^1(0, n) = 1$ if $n < N$ but 0 if $n = N$).

All other transition probabilities are equal to 0.

Along with this chain, we include a one-step reward $r_\alpha(m, n)$ in state (m, n) as defined by

$$r_\alpha(m, n) = p_\alpha(m, n; m, n-1) = \begin{cases} \nu(n) & , \text{if } \alpha = L \\ \nu(n) & , \text{if } \alpha = O \\ \nu(n)\, 1_{\{m<M\}} & , \text{if } \alpha = U \end{cases} .$$

Here 1_A denotes the indicator of the event A.

Thus, $r_\alpha(m, n)$ denotes the probability that a job leaves the system. Now let us denote the expected total reward over k-periods for initial state (m, n) by $V_\alpha^k(m, n)$. Then, by virtue of the Markov property, for any $(m, n) \in R_\alpha$ and $k \geq 1$ we obtain:

$$V_\alpha^k(m, n) = r_\alpha(m, n) + \Sigma_{m', n'}\, p_\alpha(m, n; m', n')\, V_\alpha^{k-1}(m', n') \ , \qquad (3.1)$$

with $V_\alpha^0(m, n) = 0$. From the definition of r_α it is clear that V_α^k is just the expected number of departures form the system within the first k periods.

From the above interpretation of V_α^k (or a standard tauberian theorem), we may therefore conclude that for arbitrary $(m, n) \in R_\alpha$:

$$\lambda(1 - B_\alpha) = \lim_{k\to\infty} \frac{1}{k} V_\alpha^k(m, n) \ . \tag{3.2}$$

As a result, in order to prove $B_L \le B_O$ it thus suffices to show that for some $(m, n) \in R_L \cap R_O = R_O$:

$$V_L^k(m, n) \ge V_0^k(m, n) \quad \text{for all } k \ge 0 \ , \tag{3.3}$$

and in order to prove $B_O \le B_U$ that for some $(m, n) \in R_O \cap R_U = R_U$:

$$V_0^k(m, n) \ge V_U^k(m, n) \quad \text{for all } k \ge 0 \ . \tag{3.4}$$

When comparing an α– and β–model with $R_\alpha \subset R_\beta$, we obtain for $(m, n) \in R_\alpha \cap R_\beta = R_\alpha$ from the recursion (3.1):

$$V_\beta^k(m, n) - V_\alpha^k(m, n) =$$

$$r_\beta(m, n) - r_\alpha(m, n) + \Sigma_{m',n'} p_\alpha(m, n; m', n') [V_\beta^{k-1}(m', n') - V_\alpha^{k-1}(m', n')]$$

$$+ \Sigma_{m',n'} [p_\beta(m, n; m', n') - p_\alpha(m, n; m', n')] V_\beta^{k-1}(m', n') \ .$$

By induction to k and using that $V_\beta^0(\cdot) = V_\alpha^0(\cdot) = 0$, we can conclude that for any $(m, n) \in R_\alpha \cap R_\beta = R_\alpha$:

$$V_\beta^k(m, n) \ge V_\alpha^k(m, n) \tag{3.5}$$

if for all $k \ge 0$ and $(m, n) \in R_\alpha$:

$$r_\beta(m, n) - r_\alpha(m, n) + \Sigma_{m',n'} [p_\beta(m, n; m', n')] V_\beta^k(m', n') \ge 0 \ . \tag{3.6}$$

This will be applied in sections 4 and 5 in order to verify (3.3) and (3.4).

4. Proof of lower bound

In view of the discussion in section 3 we need to prove (3.3) in order to establish that $B_L < B_O$. To this end, substitute $\alpha = 0$ and $\beta = L$ in (3.5) and (3.6) where it is noted that $R_O \subset R_L$. Since also $r_L = r_O$, it thus suffices to verify that for any $(m, n) \in R_O$ and $k \ge 0$:

$$\Sigma_{m',n'} [p_L(m, n; m', n') - p_O(m, n; m', n')] V_L^k(m', n') =$$

$$\lambda 1_{\{m=M,n<N\}}[V_L^k(M+1,n)-V_L^k(M,n)]$$
$$+\mu(m)1_{\{n=N\}}[V_L^k(m-1,N+1)-V_L^k(m,N)]\geq 0\ .$$

This is the essence of the proof since it transforms the comparison of two different models into monotonicity properties for one of them.

It turns out that in order to prove these properties some monotonicity results are needed which are given in the following lemma. The lemma states the intuitive result that the total number of departures from the system upto period k increases, but by at most 1, if initially there is one more job in one of the queues or if a job is moved from queue 1 to queue 2. It is clear that the lower bounds in (4.1) and (4.3) ensure the above inequality and thus complete the proof of $B_L \leq B_Q$.

Lemma 4.1. For all k and within R_L:

$$0\leq V_L^k(m+1,n)-V_L^k(m,n)\leq 1 \tag{4.1}$$

$$0\leq V_L^k(m,n+1)-V_L^k(m,n)\leq 1 \tag{4.2}$$

$$0\leq V_L^k(m-1,n+1)-V_L^k(m,n)\leq 1 \tag{4.3}$$

Proof. The proof will be given by induction on k. Since $V_L^k(\cdot)=0$, clearly (4.1)-(4.3) hold for $k=0$. Suppose they hold for $k=l$. Then we need to establish them for $k=l+1$. In the rest of the proof we suppress the subindex L.

(i) <u>(4.1) for $k=l+1$</u>.

From the recursion (3.1) we derive for $m+n<M+N$:

$$V^{l+1}(m,n)=\nu(n)+\lambda V^l(m+1,n)+\mu(m)\,V^l(m-1,n+1)$$
$$+\nu(n)\,V^l(m,n-1)+[1-\lambda-\mu(m)-\nu(n)]\,V^l(m,n) \tag{4.4}$$

and

$$V^{l+1}(m+1,n)=\nu(n)+\lambda 1_{\{m+n+1<M+N\}}V^l(m+2,n)$$
$$+\mu(m+1)\,V^l(m,n+1)+\nu(n)\,V^l(m+1,n-1)$$
$$+[1-\lambda 1_{\{m+n+1<M+N\}}-\mu(m+1)-\nu(n)]\,V^l(m+1,n)\ . \tag{4.5}$$

The following expressions are now to be substituted. On the right side of (4.4) the second term can be written as

$$\lambda 1_{\{m+n+1<M+N\}}V^l(m+1,n)+\lambda 1_{\{m+n+1=M+N\}}V^l(m+1,n)\ ,$$

and the last term as

$$[1-\lambda-\mu(m+1)-\nu(n)]\,V^l(m,\,n)+[\mu(m+1)-\mu(m)]\,V^l(m,\,n)\ .$$

Also, on the right side of (4.5) the third term can be written as

$$\mu(m)\,V^l(m,\,n+1)+[\mu(m+1)-\mu(m)]\,V^l(m,\,n+1)$$

and the last term as

$$[1-\lambda-\mu(m+1)-\nu(n)]\,V^l(m+1,\,n)+\lambda\,1_{\{m+n+1=N+N\}}\,V^l(m+1,\,n)\ .$$

Then by substracting (4.4) from (4.5) and collecting terms with the same coefficients, we find that

$$V^{l+1}(m+1,\,n)-V^{l+1}(m,\,n)=\lambda\,1_{\{m+n+1<M+N\}}[V^l(m+2,\,n)-V^l(m+1,\,n)]$$

$$+\lambda\,1_{\{m+n+1=M+N\}}[V^l(m+1,\,n)-V^l(m+1,\,n)]+\mu(m)[V^l(m,\,n+1)-V^l(m-1,\,n+1)]$$

$$+[\mu(m+1)-\mu(m)][V^l(m,\,n+1)-V^l(m,\,n)]+\nu(n)[V^l(m+1,\,n-1)-V^l(m,\,n-1)]$$

$$+[1-\lambda-\mu(m+1)-\nu(n)][V^l(m+1,\,n)-V^l(m,\,n)]\ . \qquad (4.6)$$

Note that the second term on the right in (4.6) equals 0, and that coefficients are nonnegative and sum up to 1. So the induction hypotheses (4.1) and (4.2) for $k=l$ directly imply (4.1) for $k=l+1$.

(4.2) for $k=l+1$.

Similarly to (4.6) one can derive for $M+n<M+N$:

$$V^{l+1}(m,\,n+1)-V^{l+1}(m,\,n)$$

$$=[\nu(n+1)-\nu(n)]+\lambda\,1_{\{m+n+1<M+N\}}[V^l(m+1,\,n+1)-V^l(m+1,\,n)]$$

$$+\lambda\,1_{\{m+n+1=M+N\}}[V^l(m,\,n+1)-V^l(m+1,\,n)]$$

$$+\mu(m)[V^l(m-1,\,n+2)-V^l(m-1,\,n+1)]+\nu(n)[V^l(m,\,n)-V^l(m,\,n-1)]$$

$$+[\nu(n+1)-\nu(n)][V^l(m,\,n)-V^l(m,\,n)]$$

$$+[1-\lambda-\mu(m)-\nu(n+1)][V^l(m,\,n+1)-V^l(m,\,n)]\ .$$

We easily see that the induction hypotheses (4.2) and (4.3) for $k=l$ imply (4.2) for $k=l+1$.

The most complicated case is (4.3). It is for this case that we need the right hand inequalities in (4.1)-(4.3).

(4.3) for $k = l + 1$.

As before it can be proven that for $m \geq 1$:

$$V^{l+1}(m-1, n+1) - V^{l+1}(m, n)$$
$$= [\nu(n+1) - \nu(n)] + \lambda\, 1_{\{m+n+1<M+N\}}[V^l(m, n+1) - V^l(m+1, n)]$$
$$+ \mu(m-1)\,[V^l(m-2, n+2) - V^l(m-1, n+1)$$
$$+ [\mu(m) - \mu(m-1)][V^l(m-1, n+1) - V^l(m-1, n+1)]$$
$$+ \nu(n)[V^l(m-1, n) - V^l(m, n-1)] + [\nu(n+1) - \nu(n)][V^l(m-1, n) - V^l(m, n)]$$
$$+ [1 - \lambda\, 1_{\{m+n+1<M+N\}} - \mu(m) - \nu(n+1)][V^l(m-1, n+1) - V^l(m, n)]$$

Note that the one but last term is non-positive. This term has to be combined with the immediate reward $\nu(n+1) - \nu(n)$. Together with the induction assumption for $k = l$ this yields (4.3) for $k = l + 1$.

5. Proof of the upper bound

According to section 3 again, we need to prove (3.4) in order to establish that $B_O \leq B_U$. To this end, substitute $\alpha = U$ and $\beta = O$ in (3.5) and (3.6) where it is noted that $R_U \subset R_O$. Then it suffices to show that for any $(m, n) \in R_U$ and all $k \geq 0$:

$$r_O(m, n) - r_U(m, n) + \Sigma_{m',n'}[p_O(m, n; m', n') - p_U(m, n; m', n')]\, V_O^k(m', n')$$
$$= \nu(n)\, 1_{\{m=M\}} + \lambda\, 1_{\{m<M, n=N\}}[V_O^k(m+1, n) - V_O^k(m, n)]$$
$$+ \nu(n)\, 1_{\{m=M, n>0\}}[V_O^k(m, n-1) - V_O^k(m, n)] \geq 0\ .$$

The lower estimate from (5.1) and the upper estimate from (5.2) in lemma 5.1 below ensure the latter inequality and thus complete the proof of $B_O \leq B_U$.

Lemma 5.1. For all k and within R_0:

$$0 \leq V_O^k(m+1, n) - V_O^k(m, n) \leq 1 \tag{5.1}$$

$$0 \leq V_O^k(m, n+1) - V_O^k(m, n) \leq 1 \tag{5.2}$$

Proof: Again the proof will follow by induction on k. Since $V_O^0(\cdot) = 0$, clearly (5.1) and (5.2) hold for $k = 0$. Suppose that they hold for $k = l$. Then we will establish them for $k = l + 1$. In the rest of the proof we suppress the subindex O.

(5.1) for $k = l + 1$

Similarly to the derivation of (4.6), we derive

$$V^{l+1}(m+1, n) - V^{l+1}(m, n) = \lambda\, 1_{\{m+1<M\}}[V^l(m+2, n) - V^l(m+1, n)]$$
$$+ \lambda\, 1_{\{m+1=M\}}[V^l(M, n) - V^l(M, n)] + \mu(m)\, 1_{\{n<N\}}[V^l(m, n+1) - V^l(m-1, n+1)]$$
$$+ [\mu(m+1) - \mu(m)]\, 1_{\{n<N\}}[V^l(m, n+1) - V^l(m, n)]$$
$$+ \nu(n)[V^l(m+1, n-1) - V^l(m, n-1)]$$
$$+ [1 - \lambda - \mu(m+1)\, 1_{\{n<N\}} - \nu(n)][V^l(m+1, n) - V^l(m, n)] \ .$$

Since all coefficients are nonnegative and sum up to 1, the induction hypotheses (5.1) and (5.2) for $k = l$ directly imply (5.1) for $k = l + 1$.

(5.2) for $k = l + 1$

As before we find

$$V^{l+1}(m, n+1) - V^{l+1}(m, n)$$
$$= [\nu(n+1) - \nu(n)] + \lambda\, 1_{\{m<M\}}[V^l(m+1, n+1) - V^l(m+1, n)]$$
$$+ \mu(m)\, 1_{\{n+1<N\}}[V^l(m-1, n+2) - V^l(m-1, n+1)]$$
$$+ [\mu(m)]\, 1_{\{n+1=N\}}[V^l(m, n+1) - V^l(m-1, n+1)]$$
$$+ \nu(n)[V^l(m, n) - V^l(m, n-1)] + [\nu(n+1) - \nu(n)][V^l(m, n) - V^l(m, n)]$$
$$+ [1 - \lambda\, 1_{\{m<M\}} - \mu(m) - \nu(n+1)][V^l(m, n+1) - V^l(m, n)] \ .$$

From this one easily verifies (5.2) for $k = l + 1$.

Remark 5.1. In Stoyan [15] general monotonicity results have been established based on the condition that the one step transition mechanism is monotone. Unfortunately, there results could not be applied for proving the Lemmas 4.1 and 5.1 since this condition fails for the present system. More precisely, with the one-step expectation operator T_L defined upon functions $g(\cdot\,, \cdot)$ by:

$$(T_L g)(m, n) = \Sigma_{(m', n')} p_L(m, n; m', n')\, g(m', n')$$

one can find functions $g(\cdot\,, \cdot)$ such that (cf. Lemma 4.1)

$$g(m+1, n) - g(m, n) \geq 0 \ ,$$

$$g(m, n+1) - g(m, n) \geq 0 \ , \text{ and}$$

$$g(m-1, n+1) - g((m, n) \geq 0 \ ,$$

while these inequalities do no longer hold with g replaced by $(T_L g)$. That is, T_L does not preserve monotonicity in the required directions. For example, with $M = N = 1$ and $g(0, 0) = 0$ and $g(m, n) = 1$ if $(m, n) \neq (0, 0)$, these inequalities are satisfied. But with $\lambda = \mu(m) = \nu(n) = \frac{1}{3}$ for all $m, n > 0$, we obtain $(T_L g)(0, 1) - (T_L g)(1, 0) = -\frac{1}{3}$.

6. Numerical results

In order to get an idea of the quality of the bounds, a number of examples has been calculated for a wide range of parameter values. The results are displayed in Tables 2, 3 and 4. Here in $\rho = \lambda/\mu$ and $\sigma = \lambda/\nu$, r and s denote the number of servers at node 1 and 2 and M and N denote the total capacities in nodes 1 and 2 as before. For the calculation of the lower bound we have used $\Phi(m) = \Phi(M)$ for $m > M$ and $\Psi(n) = \Psi(N)$ for $n > N$.

ρ	σ	r	s	M	N	B_L	B_U
1	1	1	1	1	1	.500	.667
1	1	1	1	2	2	.333	.500
1	1	1	1	3	3	.250	.400
1	1	1	1	5	5	.167	.286
1	1	1	1	10	10	.091	.167
1	1	1	1	40	20	.032	.070
.5	.5	1	1	10	5	.000	.016
10	2	1	1	10	5	.906	.947
10	10	1	1	10	5	.900	.909
2	.5	1	1	10	5	.500	.504

Table 2 Single-server case ($r = s = 1$)

ρ	σ	r	s	M	N	B_L	B_U
1	1	2	2	2	2	.133	.333
1	1	3	3	3	3	.018	.118
1	1	4	4	4	4	.001	.030
15	15	5	5	5	5	.727	.819
10	10	4	4	4	4	.683	.785
10	10	6	6	6	6	.511	.653
10	10	10	10	10	10	.199	.353
10	10	13	13	13	13	.051	.156
10	10	15	15	15	15	.012	.070
10	10	20	20	20	20	.000	.004

Table 3 Multi-server loss case ($r = M$, $s = N$)

ρ	σ	r	s	M	N	B_L	B_U
1	1	2	2	3	3	.044	.167
1	1	2	2	4	4	.014	.083
1	1	2	2	5	5	.004	.042
1	1	3	3	4	4	.003	.040
1	1	3	3	6	6	.000	.004
10	10	2	2	4	4	.826	.889
10	10	3	3	4	4	.747	.830
10	10	4	4	6	6	.646	.755
10	10	6	6	10	10	.446	.581
10	10	10	10	15	15	.102	.188
10	10	15	15	20	20	.000	.009

Table 4 Multi-server delay case ($1 < r < M$, $1 < s < N$)

Throughout the bounds turn out to be fairly accurate when the call congestion is large and to provide reasonable indicators when the call congestion is low, as is the more realistic case. The last two examples in Table 2 show that in some non symmetric cases the bounds are excellent.

In particular the bounds are useful to get a rough but fast impression of the performance. For instance, in the single server case they instantly show that in the case $\rho = \sigma = 1$ having 2 waiting places instead of 1 the congestion drops from at least .5 to at most .5, and for 5 waiting places B drops to at most .286. Or, for the multi server loss case with $\rho = \sigma = 1$ the congestion B drops from at least .133 to at most .118 if the number of servers per station is increased from 2 to 3. And in Table 4, rows 2, 3 and 4, we see that the effect of adding a server and adding a waiting place somtimes is the same.

7. Extensions

7.1. Non-exponential case

Among other references it can be concluded from Hordijk and Van Dijk [9], that "job-local-balance" guarantees that the lower and upper bound model are "insensitive". That is, for special service disciplines (such as a multi-server loss, processor sharing or last-come-first-served preemptive disciplin) the steady state distribution given in section 2 for the exponential case remains valid for non-exponential services with means μ^{-1} and ν^{-1}. The call congestions of the lower and upper bound model under any of these disciplines are thus equal to the call congestions B_L and B_U as calculated for the exponential case in section 2, regardless of the distributional forms of the services. Shortly that is, B_L and B_U are "insensitive call congestions". Intuitively therefore, although the original model is not insensitive, one might expect that the values B_L and B_U are bounds for the call congestion of the original model under such disciplines, regardless of the distributional forms of the services. Shortly, that B_L and B_U are "insensitive bounds". Counterintuitively, however, this is not generally true, as will be illustrated and discussed below.

Counterintuitive example with LCFS.

Let $M = N = 1$, $\lambda = 2$, $\mu = 100$ and $\nu = 1$. The service time distribution is exponential at node 1 but Erlang-2 at node 2. Assume that the service discipline at both nodes is last-come first-served preemptive with a single server. (This assumption is irrelevant for the original model but does play a role for the lower bound model in which 2 jobs at a node are allowed). Then by standard calculus and with rounding in the fourth decimal, one finds

$$B = .5595 \text{ and } B_L = .5723 \ .$$

Roughly, this counterintuitive result can be accounted for by the fact that in a last-come first-served preemptive discipline the residual service time of a job present can be replaced by a total service time of a new job. As a consequence, when comparing two systems, where in one a job is rejected (original) while in the other accepted (lower bound model), the accepted job in the latter model will lead to a larger mean time up to a next completion if the mean service time exceeds the mean residual service time of a job (such as for Erlang or deterministic services). The acceptation of an extra job may thus lead to a longer saturation period and therefore extra rejections later on. A similar feature will occur with processor-sharing disciplines.

Multi-server loss case (insensitive bounds).

When once accepted jobs can never be interrupted in their service by other jobs, the counterintuitive conflict discussed above seems to be avoided. The following conjecture therefore comes up.

> "For finite tandem queues with multi-server loss stations (that is, with M and N representing servers), the values B_L and B_U as computed by (2.1) and (2.2) with $\Phi(m) = m$, $\Psi(n) = n$ for all m and n, are insensitive bounds for the call congestion of any original model with mean service times μ^{-1} and ν^{-1} at node 1 and 2 respectively."

For the case of non-exponential services at only the second node, this conjecture is formally proven in Van Dijk [5] by the same technique as applied in this paper. This proof, however, is notationally much more complex and therefore restricted to only one non-exponential node. For the general case essentially the same proof can be expected. Note that the results in Table 3 are thus claimed (and partially proved) to be valid for general service distributions with means μ^{-1} and ν^{-1}.

Another conjecture is:

> For finite tandem queues with general service time distributions and FCFS disciplin the models L and U provide lower and upper bounds for the call congestion.

The proof for this result will have to be given by means of a sample path approach. Note however that the blocking probabilities B_L and B_U are as hard to obtain as the one for the original model.

7.2. Other performance measures

By taking other appropriate one step rewards one might hope to establish other performance quantities such as the average number of jobs in the system. In this case one should use $r(m, n) = m + n$. The proof however will be hard since the obvious upper bound of 1 in the Lemmas 4.1 and 5.1 has to be replaced by another one which will be a complicated function of $\mu(m)$ and $\nu(n)$.

7.3. Tandem queues with breakdowns

By combining the results from this paper with those from Van Dijk [4], simple lower and upper bounds can also be provided when the nodes can independently have a breakdown.

Evaluation

A bounding methodology is investigated for finite exponential tandem queues with general service capacities. This methodology is based upon product-form modifications which guarantee computationally attractive expressions. Simple and intuitively appealing bounds have so been proposed for the call congestion. Numerical support indicated a potential usefulness for quick engineering purposes. In particular, for multi-server loss stations the bounds are claimed to be insensitive to service distributional forms, i.e., to depend on the services only through their means.

A formal proof of the bounds is presented. The technique of this proof is based on Markov reward theory and monotonicity results. This technique is of interest in itself and seems a useful addition to standard techniques for proving monotonicity results in queueing networks.

Both the methodology and technique of the proof seem promising for further exploitation.

References

[1] I.J.B.F. Adan and J. van der Wal, "Monotonicity of the throughput of a closed queueing network in the number of jobs", Memorandum 87-03, Department of Mathematics and Computing Science, Eindhoven University of Technology (1987).

[2] I.J.B.F. Adan and J. van der Wal, "Monotonicity of the throughput in single server production and assembly networks with respect to the buffer sizes." To appear: Proceedings 1th International workshop on queuing systems with blocking.

[3] U.N. Bhat, "Finite capacity assembly-like queues" Queueing Systems 1, 85-101 (1986).

[4] N.M. van Dijk, "Simple bounds for queueing systems with breakdowns", Perf. Evaluation 8, 117-128 (1988).

[5] N.M. van Dijk, "A formal proof for the insensitivity of simple bounds for finite multi-server non-exponential tandem queues", Stochastic Processes 27, 261-277 (1988).

[6] N.M. van Dijk and B.F. Lamond, "Bounds for the call congestion of finite single-server exponential tandem queues", Working paper 1078, University of British Columbia. To appear: Opns. Res.

[7] N.M. van Dijk, P. Tsoucas and J. Walrand, "Simple bounds and monotonicity of the call congestion of finite multiserver delay systems", Research report no. 146, Free University, Amsterdam. To appear: Probability in the Engineering and Informational Sciences.

[8] A. Hordijk and N. van Dijk, "Networks of queues with blocking", Performance '81, North-Holland, 51-65, (1981).

[9] A. Hordijk and N. van Dijk, "Adjoint process, job-local-balance and insensitivity for stochastic networks". Bull. 44-th Session Int.Stat.Inst., 50, 776-788 (1983).

[10] A. Hordijk and A. Ridder, "Stochastic inequalities for an overflow model", J. Appl. Probability 24, 696-708, (1987).

[11] E.H. Lipper and B. Sengupta, "Assembly-like queues with finite capacity: Bounds, asymptotics and approximations", Queueing Systems 1, 67-83 (1986).

[12] T.S. Robertazzi and A.A. Lazar, "On the modelling and optimal flow control of the Jacksonian network". Perf. Evaluation 5, 29-43 (1985).

[13] J.G. Shanthikumar and D.D. Yao, "Stochastic monotonicity of the queue lengths in closed queueing networks", Research Report, University of California, Berkeley. To appear: Opns. Res.

[14] J.G. Shanthikumar and D.D. Yao, "General queueing networks: Representation and stochastic monotonicity", Proceedings of the 26th IEEE Conference on Decision and Control, 1084-1087 (1987).

[15] D. Stoyan, "Comparison method for queues and other stochastic models", Wiley, New York (1983).

[16] R. Suri, "A concept of monotonicity and its characterization for closed queueing networks", Opns. Res. 33, 606-624 (1985).

[17] P. Tsoucas and J. Walrand, "Monotonicity of throughput in non-Markovian networks". To appear: J. Appl. Probability.

[18] J. van der Wal, "Monotonicity of the throughput of a closed exponential network in the number of jobs", Research report COSOR 85-21, Eindhoven University of Technology (1985).

[19] W. Whitt, "Comparing counting processes and queues", Adv. Appl. Probability 13, 207-220 (1981)

[20] R.W. Wolff, "Poisson arrivals see time averages", Opns. Res. 30, 223-231 (1982).

[21] D.D. Yao, "Some properties of the throughput function of closed networks of queues, Oper. Res. Letters 3, 313-317 (1985).