AN APPROXIMATION METHOD FOR TANDEM QUEUES WITH BLOCKING

ALEXANDRE BRANDWAJN

University of California, Santa Cruz, California, and Pallas International Corporation, San Jose, California

YUNG-LI LILY JOW

Amdahl Corporation, Sunnyvale, California

(Received February 1986; revisions received October 1986; accepted April 1987)

We propose an approximate analysis of open systems of tandem queues with blocking caused by finite buffers between servers. Our approach relies on the use of marginal probability distributions ("state equivalence") coupled with an approximate evaluation of the conditional probabilities introduced through the equivalence. The method iterates over consecutive pairs of servers using the solution of a two-queue system as a building block. It produces performance measures for individual servers as well as an approximation to joint queue-length probability distributions for pairs of neighboring stations. Experience indicates that the number of iterations required for the method grows moderately with the number of nodes in the network. We give examples to demonstrate the accuracy and the convergence properties of the proposed approximation.

Systems of tandem queues with blocking caused by finite buffers between servers are important as models of production systems and communication networks. The exact analytic solution of such queueing models is unavailable except for a few special cases, involving a small number of servers or customers (e.g., Gordon and Newell 1967 and Konheim and Reiser 1976, 1978). Thus, much of the considerable literature on tandem queues with finite buffers has proposed several approximation methods (see, for example, Altiok 1982; Altiok and Stidham 1982; Balsamo and Iazeolla 1983; Boxma and Konheim 1981; Foster and Perros 1980, Gershwin 1987, Gershwin and Schick 1983; Gordon and Newell; Hillier and Boling 1967; Hunt 1956; Konheim and Reiser 1976, 1978; Labetoulle and Pujolle 1980; Pittel 1976; Suri and Diehl 1984; and Takahashi, Miyahara and Hasegawa 1980). These methods differ in their accuracy and complexity, but generally aim at producing performance measures only for individual servers. The approach in Suri and Diehl directly produces only the average total transit time through the network. Some of the most accurate methods (Boxma and Konheim) have a computational cost that increases rapidly with the number of stations in the network. Also, seemingly all these researchers consider only load-independent service rates. Note that load-dependent service rates may arise either because the system modeled exhibits explicitly such dependencies or, when the tandem system is a part of a larger model, as a result of model transformations (for example, Brandwajn 1985 and Chandy, Herzog and Woo 1975).

As discussed in Altiok and Stidham and in Suri and Diehl, there are two common definitions of blocking due to finite buffers. In the first case, a server is blocked if, at completion time, the downstream buffer is full. This definition seems to be abstracted from production systems. In the second case, a server is not allowed to start service until space is available in the downstream buffer. This definition is better suited for the modeling of communication systems when the "service" includes transmission of data to the next station. Except in special cases, these two definitions are not equivalent (see Altiok and Stidham).

The method we propose is applicable to both types of blocking. Because our interest in tandem queues stems from communications modeling, and for the sake of simplicity, we cast our exposition in terms of the second type of blocking. We also focus on state-dependent exponential servers, and only mention the possible extension to more general service time distributions.

Our approach is based on the use of marginal probability distributions ("equivalence") together with an approximate evaluation of the conditional probabilities introduced through such an equivalence. The method uses the solution of a two-queue system as a building block, in an iteration over pairs of adjacent

Subject classification: 684 approximate solution, 705 open tandem queues.

0030-364X/87/3601-0073 $01.25

stations. It produces, in addition to performance measures for individual servers, an approximation to joint queue length probability distributions for pairs of neighboring stations. The computational complexity of each iteration is linear in the number of servers, and experimentation suggests that the number of iterations grows moderately as the number of nodes in the network increases.

Section 1 describes our approach. In Section 2, we discuss the accuracy of the approximation proposed and present numerical examples to illustrate its performance. Finally, we conclude in Section 3 by briefly indicating possible extensions to the method.

1. Approximation Method

Consider the open system of K queues (nodes) in series, shown in Figure 1. Each node is assumed to possess a limited buffer space so that there may be at most M_i users (including the one in service) at node i, $i = 1, \ldots, K$. We denote by n_i, for $n_i \leq M_i$, the current number of users at server i.

Arrivals to the system are assumed to come from a quasi-Poisson source with parameter $\lambda(n_1)$, a function of the number of users at the first node. When the buffer at this node is full, i.e., $n_1 = M_1$, the source shuts down, and it resumes operation as soon as n_1 drops below M_1. This is equivalent to keeping the source active and turning away the arrivals ("lost calls").

The service at node i is assumed to be Markovian with rate $\mu_i(n_i)$, a function of the current number of users at this server (state-dependent exponential), and is allowed to start only when the next queue is not full ($n_{i+1} < M_{i+1}$, $i = 1, \ldots, K - 1$).

Denote by $p(n_1, \ldots, n_i, n_{i+1}, \ldots, n_K)$ the stationary state probability of the tandem network. To simplify notation, let us consider the service rate at node i ($i < K$) as a function of the number of users at nodes i and $i + 1$, namely, $\mu_i(n_i, n_{i+1})$. In this way, blocking in our tandem network is represented through the condition: $\mu_i(n_i, n_{i+1}) = 0$ if $n_{i+1} = M_{i+1}$. We also have, of course, $\mu_i(n_i, n_{i+1}) = 0$ if $n_i = 0$. The balance equations for the network can be written as

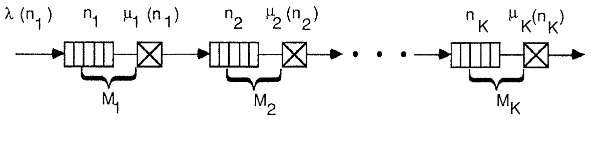

Figure 1. Tandem queues with finite buffers.

$$\left\{\lambda(n_1) + \sum_{i=1}^{K-1} \mu_i(n_1, n_{i+1}) + \mu_K(n_K)\right\}$$

$$\begin{aligned} &p(n_1, \ldots, n_i, n_{i+1}, \ldots, n_K) \\ &= \lambda(n_1 - 1)p(n_1 - 1, \ldots, n_i, n_{i+1}, \ldots, n_K) \\ &\quad + \mu_K(n_K + 1)p(n_1, \ldots, n_i, n_{i+1}, \ldots, n_K + 1) \\ &\quad + \sum_{i=1}^{K-1} \mu_i(n_i + 1, n_{i+1} - 1) \\ &\quad \times p(n_1, \ldots, n_i + 1, n_{i+1} - 1, \ldots, n_K), \end{aligned}$$

where we assume that $p(n_1, \ldots, n_i, n_{i+1}, \ldots, n_K) = 0$ for impossible states (any n_j, $j = 1, \ldots, K$, negative or greater than its corresponding maximum value).

Let us focus our attention on a pair of nodes, i and $i + 1$, and denote by $p(n_i, n_{i+1})$ the stationary joint probability distribution for the numbers of users at these two nodes. $p(n_i, n_{i+1})$ is a marginal probability derived from $p(n_1, \ldots n_i, n_{i+1}, \ldots n_K)$ by summing over all n_j for $j = 1, \ldots, i - 1, i + 2, \ldots, K$:

$$p(n_i, n_{i+1}) = \sum_{j \neq i, i+1} \sum_{n_j = 0}^{M_j} p(n_1, \ldots, n_i, n_{i+1}, \ldots, n_K).$$

Performing this summation on the balance equations of the K station tandem network and expressing $p(n_1, \ldots, n_K)$ as $\text{Prob}\{n_1, \ldots, n_K \mid n_i, n_{i+1}\} \cdot p(n_i, n_{i+1})$, we get

$$\begin{aligned} &\{a_i(n_i, n_{i+1}) + \mu_i(n_i, n_i + 1) \\ &\qquad\qquad + u_{i+1}(n_i, n_{i+1})\}p(n_i, n_{i+1}) \\ &= a_i(n_i - 1, n_{i+1})p(n_i - 1, n_{i+1}) \\ &\quad + \mu_i(n_i + 1, n_{i+1} - 1)p(n_i + 1, n_{i+1} - 1) \\ &\quad + u_{i+1}(n_i, n_{i+1} + 1)p(n_i, n_{i+1} + 1), \\ &\qquad n_i = 0, \ldots, M_i; \quad n_{i+1} = 0, \ldots, M_{i+1}, \end{aligned}$$

where, again, impossible terms are assumed to vanish and

$$a_i(n_i, n_{i+1})$$

$$= \begin{cases} \displaystyle\sum_{\substack{j \neq i \\ j \neq i+1}} \sum_{n_j = 0}^{M_j} \mu_{i-1}(n_{i-1}, n_i)\text{Prob}\{n_1, \ldots, n_K \mid n_i, n_{i+1}\}, \\ \qquad\qquad i = 2, \ldots, K - 1. \\ \lambda(n_i), \qquad i = 1. \end{cases}$$

and

$$u_{i+1}(n_i, n_{i+1}) = \begin{cases} \displaystyle\sum_{\substack{j\neq i\\ j\neq i+1}} \sum_{n_j=0}^{M_j} \mu_{i+1}(n_{i+1}, n_{i+2}) \\ \qquad\qquad \text{Prob}\{n_1, \ldots, n_K \mid n_i, n_{i+1}\}, \\ \qquad\qquad i = 1, \ldots, K-2, \\ \mu_K(n_{i+1}), \qquad i = K-1. \end{cases}$$

$\text{Prob}\{n_1, \ldots n_K \mid n_i, n_{i+1}\}$ is the conditional probability of having $(n_1, \ldots, n_{i-1}, n_{i+2}, \ldots, n_K)$ users at nodes outside the selected pair, given that there are (n_i, n_{i+1}) users at the selected nodes.

Rewriting this conditional probability as

$$\text{Prob}\{n_{i-1} \mid n_i, n_{i+1}\} \cdot \text{Prob}\{n_1, \ldots, n_K \mid n_{i-1}, n_i, n_{i+1}\}$$

and

$$\text{Prob}\{n_{i+2} \mid n_i, n_{i+1}\} \cdot \text{Prob}\{n_1, \ldots, n_K \mid n_i, n_{i+1}, n_{i+2}\}$$

in $a_i(n_i, n_{i+1})$ and $u_{i+1}(n_i, n_{i+1})$, respectively, and carrying out the summation, we readily obtain

$$a_i(n_i, n_{i+1}) = \begin{cases} \displaystyle\sum_{n_{i-1}=1}^{M_{i-1}} \mu_{i-1}(n_{i-1})\text{Prob}\{n_{i-1} \mid n_i, n_{i+1}\}, \\ \qquad\qquad i = 2, \ldots, K-1, \\ \lambda(n_i), \qquad i = 1. \end{cases} \tag{1}$$

$$n_i = 0, \ldots, M_i - 1; \quad n_{i+1} = 0, \ldots, M_{i+1},$$

and the rate of departures from the second node is

$$u_{i+1}(n_i, n_{i+1}) = \begin{cases} \displaystyle\sum_{n_{i+2}=0}^{M_{i+2}-1} \mu_{i+1}(n_{i+1})\text{Prob}\{n_{i+2} \mid n_i, n_{i+1}\}, \\ \qquad\qquad i = 1, \ldots, K-2, \\ \mu_K(n_{i+1}), \qquad i = K-1, \end{cases} \tag{2}$$

$$n_i = 0, \ldots, M_i; \quad n_{i+1} = 1, \ldots, M_{i+1}.$$

The equations for $p(n_i, n_{i+1})$ imply that, *with respect to* $p(n_i, n_{i+1})$, the singled-out pair behaves like the two-node tandem system of Figure 2 with a state-dependent arrival rate $a_i(n_i, n_{i+1})$ and the rate of departures from the second node $u_{i+1}(n_i, n_{i+1})$ given by (1) and (2), respectively.

In this equivalent system, the two nodes have the same limited capacity as in the original system. Interarrival times and service at the second node are exponentially distributed, and the service at the first node can start only when the second queue is not full.

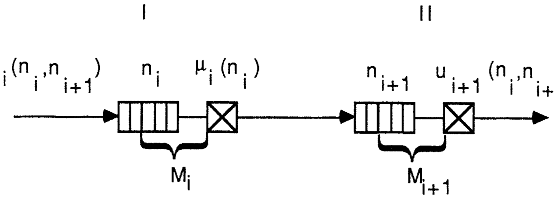

Figure 2. Equivalent system for a pair of queues.

It is important to the stress that the equivalence is nothing more than a representation of the marginal (with respect to the state of the whole network) probability $p(n_i, n_{i+1})$, and involves no approximation. The interested reader is referred to Brandwajn (1985) for a discussion of other applications of this approach.

Assume that we know how to solve (e.g., using a numerical method) such a two-node system. We would then obtain the joint probability distribution $p(n_i, n_{i+1})$ if we knew the "equivalent" rates of arrival and service, $a_i(n_i, n_{i+1})$ and $u_{i+1}(n_i, n_{i+1})$. These rates involve the probability of the number of users at node $i - 1$ and $i + 2$, respectively, given n_i, n_{i+1}. In order to provide an approximation of these probabilities, we assume that only the state of the immediate neighbor matters in the condition so that:

$$\text{Prob}\{n_{i-1} \mid n_i, n_{i+1}\} \simeq \text{Prob}\{n_{i-1} \mid n_i\}, \qquad i = 2, \ldots, K-1, \tag{3}$$

and

$$\text{Prob}\{n_{i+2} \mid n_i, n_{i+1}\} \simeq \text{Prob}\{n_{i+2} \mid n_{i+1}\}, \qquad i = 1, \ldots, K-2.$$

Hence,

$$a_i(n_i, n_{i+1}) \simeq a_i^*(n_i) \triangleq \begin{cases} \displaystyle\sum_{n_{i-1}=1}^{M_{i-1}} \mu_{i-1}(n_{i-1})\text{Prob}\{n_{i-1} \mid n_i\}, \\ \qquad\qquad i = 2, \ldots, K-1, \\ \lambda(n_1), \qquad i = 1. \end{cases} \tag{4}$$

and

$$u_{i+1}(n_i, n_{i+1}) \simeq u_{i+1}^*(n_{i+1}) \triangleq \begin{cases} \displaystyle\sum_{n_{i+2}=0}^{M_{i+2}-1} \mu_{i+1}(n_{i+1})\text{Prob}\{n_{i+2} \mid n_{i+1}\}, \\ \qquad\qquad i = 1, \ldots, K-2, \\ \mu_K(n_K), \qquad i = K-1. \end{cases}$$

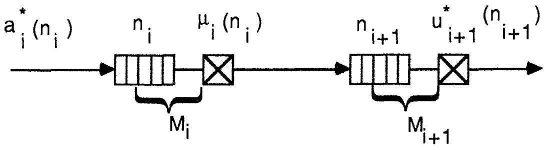

Figure 3. Two-server solution cell.

Intuitively, assumption (3) says that nearly all the influence of node $i + 1$ on node $i - 1$ is contained in the state of node i.

This approximation allows us to develop an iterative approach similar in spirit to those discussed in Gershwin, and in Labetoulle and Pujolle, that consider the nodes in the original tandem system of Figure 1 two-by-two. We use the $p(n_i, n_{i+1})$ obtained from the solution of the pair of nodes i and $i + 1$, now reduced to the system shown in Figure 3, to compute $u^*_i(n_i)$ and $a^*_{i+1}(n_{i+1})$ for the pairs $i - 1$, $i + 1$ and $i + 1$, $i + 2$, respectively. More precisely, the steps of the iterative procedure are as follows (we use a superscript to denote iteration number):

Step 1. Select an initial approximation for the conditional probabilities $\text{Prob}^0\{n_{i+1} \mid n_i\}$, $i = 2, \ldots, K - 1$.

Step 2. At iteration j, solve $K - 1$ pairs of neighboring nodes i, $i + 1$, for $i = 1, \ldots, K - 1$. For the solution of the pair i, $i + 1$, use $\text{Prob}^{j-1}\{n_{i+2} \mid n_{i+1}\}$ to obtain $u^*_{i+1}(n_{i+1})$, and $\text{Prob}^j\{n_{i-1} \mid n_i\}$ for $a^*_i(n_i)$. The solution of the pair yields $p^j(n_i, n_{i+1})$ used to compute $\text{Prob}^j\{n_i \mid n_{i+1}\}$ and $\text{Prob}^j\{n_{i+1} \mid n_i\}$ for other pairs, as well as any performance measures for nodes i and $i + 1$.

Step 3. If a convergence criterion, e.g., maximum absolute value between two iterates for each node less than a given value, has been met, stop. Otherwise, set j to $j + 1$ and return to Step 2.

To illustrate the solution procedure, let us apply it to a four node system.

Step 1. First, we select $p^0\{n_2 \mid n_1\}$, $p^0\{n_3 \mid n_2\}$, and $p^0\{n_4 \mid n_3\}$. We set j to 1.

Step 2. At iteration j, we solve 3 pairs of the two-server system of Figure 3 as follows:

(i) For the pair of nodes 1 and 2:
$a^*_1(n_1) = \lambda_1(n_1)$ is known, and we use $p^{j-1}\{n_3 \mid n_2\}$ to compute $u^*_2(n_2)$. Then we solve for $p^j(n_1, n_2)$.

(ii) For the pair of nodes 2 and 3:
We use $p^j\{n_1 \mid n_2\}$ to compute $a^*_2(n_2)$ and $p^{j-1}\{n_4 \mid n_3\}$ to compute $u^*_3(n_3)$. Then we solve for $p^j(n_2, n_3)$

(iii) For the pair of nodes 3 and 4:
We use $p^j\{n_2 \mid n_3\}$ to compute $a^*_3(n_3)$ and $u^*_4(n_4) = \mu_4(n_4)$ is known. We then solve for $p^j(n_3, n_4)$.

Step 3. If the convergence criterion has been satisfied, we stop. Otherwise, we set j to $j + 1$ and return to Step 2.

Each iteration involves $K - 1$ solutions of the two-server cell of Figure 3. To solve such a cell, one can apply a numerical method (e.g., Brandwajn 1979) to the balance equations. It is also possible to adapt the approximate solution in Boxma and Konheim (1981), which results in computation time savings, at the expense of accuracy in the evaluation of the joint probabilities.

We do not have a proof of convergence. In practice, in the many examples we have examined, the method has always converged within a reasonable number of iterations (low 10s), only moderately dependent on the number of nodes. As a result, the computational complexity of our approach grows relatively moderately (but more than linearly) with the number of queues in the tandem system.

It is worthwhile noting that the approximation that underlies our method (3) holds exactly if there is no blocking, so that an exact solution is then produced. As the next section illustrates, the accuracy of the approach is generally good, even in the presence of significant blocking.

2. Numerical Examples and Accuracy

We have tested our approach on a number of examples (since the method takes a two-node solution as a basis, only networks with three or more queues are of interest). The results obtained indicate that, in terms of expected queue lengths, marginal queue size distributions and also two-by-two node joint queue size distribution, the accuracy of our approximation is generally good except in a few specific and recognizable cases.

To illustrate this fact, we present in this section the results obtained for six examples chosen so as to cover a reasonable selection of parameters. As basis for comparison, we use either the "exact" solution obtained by solving numerically the global balance equations of the tandem network, or, for larger systems, the results of a discrete-event simulation.

We compare the joint queue length distribution for pairs of nodes for one example. We compare the marginal queue length distribution for individual nodes in the network for a small selection of parameters; all other examples compare two values of this distribution: the probability that the node is empty and the probability that the queue is full. This latter probability is an indication of blocking. We compare the average queue lengths, as well as the expected total number of customers in the system, the network throughput and the average sojourn time in system, for all examples.

Our examples utilized an exact numerical solution of the two-node cell. The system considered has state-independent service and arrival rates in all but the last example of this section.

First, we consider the three-queue tandem system studied in Example 1, p. 44 of Boxma and Konheim.

Example 1

$K = 3 \qquad M_1 = M_2 = M_3 = 2$

$\lambda = 1 \qquad \mu_1 = \mu_2 = \mu_3 = 1$

Table I shows the results obtained for Example 1. We observe that relative errors in average queue lengths and marginal queue size distributions are below 1%. Note that we report distributions rather than cumulative functions since the latter may somewhat artificially lower percent errors. We also compare the two-station joint probability distributions for which, in this example, the approximation has a maximum relative error below 5%.

Similar results have been obtained in many other

Table I
Results for Example 1

Average queue lengths				Expected total number in system, throughput, time in system			
Queue	Exact	Approximate	% error	Quantity	Exact	Approximate	% error
1	1.2538	1.2562	0.2%	Total in system	3.0000	3.0000	0.0%
2	1.0000	1.0000	0.0%	Throughput	0.5362	0.5342	0.4%
3	0.7462	0.7438	−0.3%	Time in system	5.5949	5.6159	0.4%

Marginal queue length distributions

	Queue i					
n_i	1		2		3	
	Exact	Approx.	Exact	Approx.	Exact	Approx.
0	0.2101	0.2096	0.3266	0.3275	0.4638	0.4658
1	0.3261	0.3247	0.3469	0.3451	0.3261	0.3247
2	0.4638	0.4658	0.3265	0.3275	0.2101	0.2096
Max % error	0.4%		0.5%		0.4%	

Two-node joint queue length distribution $p(n_1, n_2)$

	n_2					
n_1	0		1		2	
	Exact	Approx.	Exact	Approx.	Exact	Approx.
0	0.0585	0.0604	0.0788	0.0780	0.0728	0.0712
1	0.0795	0.0802	0.1306	0.1293	0.1160	0.1151
2	0.1886	0.1869	0.1375	0.1378	0.1378	0.1411
Max % error			3.3%			

Two-node joint queue length distribution $p(n_2, n_3)$

	n_3					
n_2	0		1		2	
	Exact	Approx.	Exact	Approx.	Exact	Approx.
0	0.1378	0.1411	0.1160	0.1151	0.0728	0.0712
1	0.1375	0.1378	0.1306	0.1293	0.0788	0.0780
2	0.1886	0.1869	0.0795	0.0802	0.0585	0.0604
Max % error			3.2%			

Table II
Results for Example 2

Average queue lengths				Expected total number in system, throughput, time in system			
Queue	Exact	Approximate	% error	Quantity	Exact	Approximate	% error
1	4.5204	3.9357	−12.9%	Total in system	8.1581	7.9353	2.7%
2	1.1510	1.1620	0.9%	Throughput	0.9327	0.9597	2.8%
3	2.4867	2.5823	3.8%	Time in system	8.7467	8.2685	−5.4%

Marginal queue length distribution						
	Queue i					
	1		2		3	
n_i	Exact	Approx.	Exact	Approx.	Exact	Approx.
0	0.1965	0.2030	0.3058	0.2914	0.0673	0.0403
1	0.1129	0.1310	0.2375	0.2553	0.0806	0.0692
2	0.0824	0.1023	0.4568	0.4534	0.1502	0.1585
3	0.0714	0.0884			0.7019	0.7321
4	0.0676	0.0795				
5	0.0666	0.0727				
6	0.0666	0.0669				
7	0.0669	0.0616				
8	0.0672	0.0568				
9	0.0674	0.0524				
10	0.0673	0.0452				
11	0.0673	0.0403				
Max % error	−40.1%		7.4%		−40.1%	

cases. There are, however, instances in which the approximation is much less accurate. Table II demonstrates such a case, using another three-node tandem system:

Example 2

$K = 3 \quad M_1 = 11 \quad M_2 = 2 \quad M_3 = 3$

$\lambda = 1 \quad \mu_1 = 2.5 \quad \mu_2 = 10 \quad \mu_3 = 1$

Here, the relative errors for expected queue lengths reach 13%, while the marginal queue size distributions are off by up to 40%. It is interesting to note that the expected queue length happens to be most distorted for the node with the largest buffer size, although the approximation that underlies our approach is asymptotically exact as the buffer size increases.

By examining the probability of a full buffer (i.e., blocking) in the marginal queue length distributions, we can see that this probability exhibits an order of magnitude "jump" between nodes 1 and 2. It has been our experience that poor accuracy is likely when there is such a large change in blocking probability within a pair of neighboring nodes and the first node has a relatively large buffer size (say, over 8). Note that, even though distorted, the approximation reproduces the large change in blocking probabilities so that it is usually apparent when poor accuracy can be suspected. It is worthwhile mentioning that in all such inaccurate cases we have encountered, the approximation underestimated the congestion at the affected node.

Note that a large buffer does not, per se, cause poor approximation. In Example 3, we consider a system with the same buffer lengths as in Example 2 and service rates chosen to remove the large jump in blocking probabilities.

Example 3

$K = 3 \quad M_1 = 11 \quad M_2 = 2 \quad M_3 = 3 \quad \lambda = 1$

$\mu_1 = 1.11\ldots \quad \mu_2 = 3.33\ldots \quad \mu_3 = 1.66\ldots$

Table III shows the results obtained for this example.

Our fourth example is also a three-node network. Its parameters were chosen to illustrate the performance of our method with moderate imbalance in queue sizes and service rates and with moderate blocking (probability of a full buffer around 0.25).

Example 4

$K = 3 \quad M_1 = 2 \quad M_2 = 4 \quad M_3 = 6 \quad \lambda = 1$

$\mu_1 = 2.5 \quad \mu_2 = 1.667 \quad \mu_3 = 0.833$

Table IV shows the results obtained for this example.

Table III
Results for Example 3

Average queue lengths				Expected total number in system, throughput, time in system			
Queue	Exact	Approximate	% error	Quantity	Exact	Approximate	% error
1	5.5876	5.5727	−0.3%	Total in system	6.9703	6.9530	−0.2%
2	0.4375	0.4365	−0.2%	Throughput	0.9217	0.9206	−0.1%
3	0.9452	0.9438	−0.2%	Time in system	7.5624	7.5526	−0.1%

Marginal queue length probabilities

Queue i	Prob$\{n_i = 0\}$		Prob$\{n_i = M_i\}$		Max % error
	Exact	Approx.	Exact	Approx.	
1	0.0783	0.0794	0.0782	0.0787	1.5%
2	0.6606	0.6612	0.0981	0.0977	0.1%
3	0.4469	0.4472	0.1124	0.1119	0.9%

We selected our next example to illustrate the performance of our method with a larger number of nodes, relatively unbalanced queue sizes and service rates, and moderate to high blocking (12%–60%). Here, we use simulation results are basis for comparison. Most results are reported as 95%-level confidence intervals obtained using the batch-means method for simulation output analysis. The simulation used 10 batches, each batch corresponding to 10,000 customers completing their transit through the network.

Example 5

$K = 10 \quad M_1 = M_3 = M_6 = M_7 = M_9 = 3$

$M_2 = M_5 = M_8 = 4 \quad M_4 = M_{10} = 2$

$\lambda = 1 \quad \mu_1 = 1.125$

$\mu_2 = \mu_4 = \mu_6 = 1 \quad \mu_3 = \mu_7 = 0.833\ldots$

$\mu_5 = \mu_8 = \mu_9 = 0.666\ldots \quad \mu_{10} = 1.111\ldots.$

Table V shows the results for this ten-node system. We observe that most approximation results in this example fall either within the confidence intervals or only slightly outside.

As is clear from its derivation, the proposed approach can be applied to networks with state-dependent arrival and/or service rates. Example 6 is a five-node system with state dependent service rates. Table VI compares our method with simulation results for this example. As in Table V, we obtained 95% level confidence intervals using the batch means method, with 10 batches of 10,000 completions each.

Example 6

$K = 5 \quad M_1 = 4 \quad M_2 = M_4 = 5 \quad M_3 = M_5 = 3$

$\lambda = 1 \quad \mu_i(n_i) = \mu_i\{1 - (n_i - 1)/(2M_i - 2)\},$

$n_i = 1, \ldots, M_i$

$\mu_i = \mu_2 = \mu_4 = \mu_5 = 2 \quad \mu_3 = 2.5$

Table IV
Results for Example 4

Average queue lengths				Expected total number in system, throughput, time in system			
Queue	Exact	Approximate	% error	Quantity	Exact	Approximate	% error
1	0.7220	0.7138	−1.1%	Total in system	6.8707	6.9499	1.2%
2	2.0373	2.0550	0.9%	Throughput	0.7822	0.7868	0.6%
3	4.1114	4.1811	1.7%	Time in system	8.7838	8.8331	0.6%

Marginal queue length probabilities

Queue i	Prob$\{n_i = 0\}$		Prob$\{n_i = M_i\}$		Max % error
	Exact	Approx.	Exact	Approx.	
1	0.4958	0.4994	0.2178	0.2132	2.1%
2	0.2315	0.2222	0.2551	0.2520	4.0%
3	0.0613	0.0558	0.3284	0.3368	8.9%

Table V
Comparisons for Example 5

Average queue lengths			Expected total number in system, throughput, time in system		
Queue	Simulation	Approximation	Quantity	Simulation middle point	Approximation
1	(2.3042, 2.3182)	2.3009	Total in system	19.9642	19.8392
2	(3.4005, 3.4206)	3.3933	Throughput	0.4476	0.4477
3	(2.3131, 2.3218)	2.3180	Time in system	44.6027	44.3136
4	(1.1728, 1.1938)	1.1848			
5	(2.6881, 2.7345)	2.7004			
6	(1.5299, 1.5734)	1.5347			
7	(1.7991, 1.8445)	1.8028			
8	(2.5773, 2.6233)	2.5553			
9	(1.5118, 1.5417)	1.5224			
10	(0.5248, 0.5351)	0.5272			

Marginal queue length probabilities

Queue i	Prob$\{n_i = 0\}$		Prob$\{n_i = M_i\}$	
	Simulation	Approx.	Simulation	Approx.
1	(0.0574, 0.0608)	0.0642	(0.5489, 0.5559)	0.5523
2	(0.0103, 0.0127)	0.0135	(0.6041, 0.6109)	0.6042
3	(0.0506, 0.0523)	0.0497	(0.5442, 0.5498)	0.5440
4	(0.2438, 0.2542)	0.2492	(0.4270, 0.4376)	0.4340
5	(0.0741, 0.0823)	0.0784	(0.3617, 0.3731)	0.3639
6	(0.2331, 0.2490)	0.2476	(0.2662, 0.2773)	0.2675
7	(0.1638, 0.1774)	0.1765	(0.3581, 0.3721)	0.3599
8	(0.0897, 0.0971)	0.1002	(0.3270, 0.3377)	0.3225
9	(0.2290, 0.2370)	0.2348	(0.2439, 0.2542)	0.2486
10	(0.5922, 0.5983)	0.5970	(0.1229, 0.1275)	0.1243

As a whole, the proposed method appears to produce fairly accurate results. The number of iterations needed for convergence is moderate: Table VII shows the number of iterations required in the examples just presented that are necessary to achieve a maximum difference of less than 10^{-5} between consecutive iterates.

The data of Table VII suggest that the number of iterations needed, for a network with a given number of nodes, depends on the particular set of the network

Table VI
Results for Example 6

Average queue lengths			Expected total number in system, throughput, time in system		
Queue	Simulation	Approximation	Quantity	Simulation middle point	Approximation
1	(1.5595, 1.6604)	1.5876	Total in system	7.4944	7.4218
2	(1.8349, 2.0180)	1.9283	Throughput	0.8275	0.8302
3	(0.9935, 1.0710)	1.0247	Time in system	9.0566	8.9398
4	(1.8398, 2.0031)	1.8715			
5	(0.9897, 1.0190)	1.0097			

Marginal queue length probabilities

Queue i	Prob$\{n_i = 0\}$		Prob$\{n_i = M_i\}$	
	Simulation	Approx.	Simulation	Approx.
1	(0.3239, 0.3498)	0.3447	(0.1642, 0.1808)	0.1699
2	(0.3029, 0.3354)	0.3172	(0.1270, 0.1515)	0.1374
3	(0.4502, 0.4763)	0.4680	(0.1610, 0.1841)	0.1725
4	(0.3096, 0.3365)	0.3292	(0.1306, 0.1528)	0.1340
5	(0.4628, 0.4730)	0.4660	(0.1574, 0.1652)	0.1633

Table VII
Number of Iterations Required

Example	1	2	3	4	5	6
Iterations	5	12	4	6	17	10
Network nodes	3	3	3	3	10	5

Table VIII
Number of Iterations vs. Number of Nodes

Nodes	5	7	9	11	13	15	17	19
Iterations	9	12	11	20	26	33	39	46

parameters. Intuitively, this number can be expected to grow as the number of nodes increases. Because the addition of a node in a tandem network modifies the behavior of other nodes, it seems difficult to study the effect of the number of nodes on convergence speed in isolation of the effect of other parameters. With this caveat, we present in Table VIII the number of iterations required to achieve convergence (less than 10^{-5} difference between iterates) in a network composed of K identical nodes; $M_i = 2$, $\mu_i = 1$, $i = 1, \ldots, K$, $\lambda = 1$, for several values of the number of nodes, K.

We observe that, although the number of iterations generally does increase with the number of nodes, it tends to remain in the low 10s even for relatively large numbers of nodes.

In the next section, we briefly consider a few extensions to our approach.

3. Extensions

As mentioned in the introduction, the proposed method can be extended to handle more general tandem queueing systems. One possible extension is to allow production type of blocking, as discussed in Altiok and Stidham and in Suri and Diehl.

To illustrate the performance of our method with this type of blocking, we use three examples taken from Perros and Altiok (1984) and, for comparison, we reproduce the results of the approximation method proposed by those researchers. All the comparisons include the average queue lengths, the expected total number of customers in the system, the throughput and the average sojourn time in the network. We also include two values of the marginal queue length distribution for individual nodes: the probability of an empty and full node.

In the comparisons, "exact" refers to a numerical solution of system equations, and "simulation" results are those reported in Perros and Altiok.

The first example in this section, Example 7, is a three-node system. Table IX shows the results for this example.

Example 7

$$K = 3 \quad M_1 = M_2 = M_3 = 2$$

$$\lambda = 3 \quad \mu_1 = 3 \quad \mu_2 = 4 \quad \mu_3 = 2$$

Examples 8 and 9 (Tables X and XI, respectively) both correspond to a five-node tandem network.

Table IX
Results for Example 7

Average queue lengths				Expected total number in system, throughput, time in system			
Queue	Exact	Our approximation	Perros-Altiok approximation	Quantity	Exact	Our approximation	Perros-Altiok approximation
1	1.2417	1.2443	1.2359	Total in system	3.9347	3.9126	3.8347
2	1.2609	1.2481	1.1991	Throughput	1.6476	1.6365	1.6242
3	1.4321	1.4202	1.3997	Time in system	2.3881	2.3908	2.3609
Max % error		1.0%	4.9%				

Marginal queue length probabilities

Queue i	Prob$\{n_1 = 0\}$			Prob$\{n_i = M_i\}$		
	Exact	Our approx.	P-A approx.	Exact	Our approx.	P-A approx.
1	0.2090	0.2102	0.2228	0.4508	0.4545	0.4586
2	0.2429	0.2492	0.2742	0.5038	0.4974	0.4733
3	0.1761	0.1818	0.1880	0.6082	0.6020	0.5877
Max % error		3.2%	12.9%		1.3%	6.1%

Table X
Results for Example 8

Average queue lengths				Expected total number in system, throughput, time in system			
Queue	Simulation	Our approximation	Perros-Altiok approximation	Quantity	Simulation	Our approximation	Perros-Altiok approximation
1	(1.3614, 1.3698)	1.3642	1.3612	Total in system	5.6443	5.5591	5.3256
2	(0.9319, 0.9725)	0.9450	0.8916	Throughput	1.4358	1.4307	1.4151
3	(0.9099, 0.9623)	0.9144	0.8218	Time in system	3.9311	3.8856	3.7634
4	(1.1713, 1.2187)	1.1598	1.0980				
5	(1.1770, 1.2144)	1.1757	1.1530				

Marginal queue length probabilities

Queue i	$\text{Prob}\{n_i = 0\}$			$\text{Prob}\{n_i = M_i\}$		
	Simulation	Our approx.	P-A approx.	Simulation	Our approx.	P-A approx.
1	0.1608	0.1589	0.1671	0.5214	0.5231	0.5283
2	0.3805	0.3825	0.4167	0.3325	0.3276	0.3083
3	0.3990	0.4100	0.4549	0.3351	0.3245	0.2767
4	0.2795	0.2957	0.3178	0.4745	0.4555	0.4158
5	0.2779	0.2847	0.2924	0.4736	0.4604	0.4455

Example 8

$K = 5 \quad M_1 = M_2 = \ldots = M_5 = 2$

$\lambda = 3 \quad \mu_1 = 2 \quad \mu_2 = 3 \quad \mu_3 = 4 \quad \mu_4 = 3 \quad \mu_5 = 2.$

Example 9

$K = 5 \quad M_1 = M_2 = \ldots = M_5 = 2$

$\lambda = 2 \quad \mu_1 = \mu_2 = \ldots = \mu_5 = 2.$

We observe that most of our approximation results in Examples 8 and 9 fall either within the confidence intervals or only slightly outside. The number of iterations needed to attain convergence (less that 10^{-5} difference between iterates) for the examples in this section is 5, 8 and 8, respectively.

Another extension that works well with the proposed method is to allow a more general state dependence of service rates; that is, it is easy to include rates that depend both on local congestion and on the number of customers at the next node. Denote by $\mu_i(n_i, n_{i+1})$, for $i = 1, \ldots, K - 1$, such a rate for node i. Note that this type of state dependency may arise if the service at node i actually involves some interven-

Table XI
Results for Example 9

Average queue lengths				Expected total number in system, throughput, time in system			
Queue	Simulation	Our approximation	Perros-Altiok approximation	Quantity	Simulation	Our approximation	Perros-Altiok approximation
1	(1.1983, 1.2240)	1.2338	1.2374	Total in system	5.5572	5.5084	5.4249
2	(1.2405, 1.2701)	1.2562	1.2307	Throughput	1.1378	1.1068	1.0862
3	(1.1688, 1.1966)	1.1609	1.1346	Time in system	4.8842	4.9768	4.9943
4	(1.0498, 1.0755)	1.0318	1.0119				
5	(0.8368, 0.8541)	0.8257	0.8103				

Marginal queue length probabilities

Queue i	$\text{Prob}\{n_i = 0\}$			$\text{Prob}\{n_i = M\}$		
	Simulation	Our approx.	P-A approx.	Simulation	Our approx.	P-A approx.
1	0.2199	0.2128	0.2195	0.4311	0.4466	0.4569
2	0.2365	0.2394	0.2571	0.4919	0.4956	0.4878
3	0.2737	0.2848	0.2995	0.4565	0.4457	0.4341
4	0.3277	0.3442	0.3563	0.3903	0.3760	0.3682
5	0.4362	0.4466	0.4569	0.2817	0.2723	0.2672

tion of node $i + 1$ (as, e.g., in a transmission/polling situation).

The only change required is the substitution of $\mu_i(n_i, n_{i+1})$ in lieu of the $\mu_i(n_i)$. All the steps of the iterative procedure remain the same as previously, with the proviso that it does not seem possible to extend the approximation of Boxma and Konheim to this case, so that a numerical solution (or some new approximation) must be used in the solution of the two-node cell of Figure 3.

As shown in the numerical examples, the proposed approach produces results that appear in many cases more accurate than existing approximations. Intuitively, this result seems attributable to the fact that we consider network stations in pairs (rather than individually). This tactic allows a better representation of the blocking, at least within each pair. These pairs are formally defined through two-station joint queue length distributions. The assumption that the state of immediate neighbors matters most—used to actually approximate parameters of such pairs—is simple, intuitively appealing, and judging by the methods performance, well satisfied in many cases. Our method is also applicable to tandem queues with state-dependent service rates, not covered by many existing approaches.

References

ALTIOK, T. 1982. Approximate Analysis of Exponential Tandem Queues with Blocking. *Eur. J. Opnl. Res.* **11,** 390–398.

ALTIOK, T. M., AND S. S. STIDHAM. 1982. A Note on Transfer Lines with Unreliable Machines, Random Processing Times, Finite Buffers. *IIE Trans.* **14,** 125–127.

BALSAMO, S., AND G. IAZEOLLA. 1983. Some Equivalence Properties for Queueing Networks with and without Blocking. In *Proceedings of Performance '83*, A. K. Argrawala and S. K. Tripathi (eds.). North Holland, pp. 351–360, Amsterdam.

BOXMA, O. J., AND A. G. KONHEIM. 1981. Approximate Analysis of Exponential Queueing Systems with Blocking. *Acta Inform.* **15,** 19–66.

BRANDWAJN, A. 1979. An Iterative Solution of Two-Dimensional Birth and Death Processes. *Opns. Res.* **27,** 595–605.

BRANDWAJN, A. 1985. Equivalence and Decomposition in Queueing Systems—A Unified Approach. *Perform. Eval.* **5,** 175–186.

CHANDY, K. M., V. HERZOG, AND L. WOO. 1975. Parametric Analysis of Queueing Networks. *IBM J. Res. Dev.* **19,** 36–42.

FOSTER, F. G., AND H. G. PERROS. 1980. On the blocking process in queue networks. *Eur. J. Opns. Res.* **5,** 276–283.

GERSHWIN, S. B. 1987. An Efficient Decomposition Method for Approximate Evaluation of Production Lines with Finite Storage Space. *Opns. Res.* **35,** 291–305.

GERSHWIN, S. B., AND I. C. SCHICK. 1983. Modeling and Analysis of Three Stage Transfer Lines with Unreliable Machines and Finite Buffers. *Opns. Res.* **31,** 354–380.

GORDON, W. J., AND G. F. NEWELL. 1967. Cyclic Queueing Systems with Restricted Length Queues. *Opns. Res.* **15,** 266–277.

HILLIER, F. S., AND R. W. BOLING. 1967. Finite Queues in Series with Exponential or Erlang Service Times—A Numerical Approach. *Opns. Res.* **15,** 286–303.

HUNT, G. C. 1956. Sequential Arrays of Waiting Lines. *Opns. Res.* **4,** 674–683.

KONHEIM, A. G., AND M. REISER. 1976. A Queueing Model with Finite Waiting Room and Blocking. *J. Assoc. Comput. Mach.* **23,** 328–341.

KONHEIM, A. G., AND M. REISER. 1978. Finite Capacity Queueing Systems with Applications in Computer Modelling. *SIAM J. Comput.* **7,** 210–229.

LABETOULLE, J., AND G. PUJOLLE. 1980. Isolation Methods in a Network of Queues. *IEEE Trans. Soft. Eng.* **SE-6,** 373–381.

PERROS, H. G., AND T. ALTIOK. 1984. Approximate Analysis of Open Networks of Queues with Blocking: Tandem Configurations. North Carolina State University and Rutgers University.

PITTEL, B. 1976. Closed Exponential Networks of Queues with Blocking. IBM Res. Rep. N. 26548, Yorktown Heights, NY.

SURI, R., AND G. W. DIEHL. 1984. A New "Building Block" for Performance Evaluation of Queueing Networks with Finite Buffers. *ACM Perf. Eval. Rev.* **12,** 134–142.

TAKAHASHI, Y., H. MIYAHARA, AND T. HASEGAWA. 1980. An Approximation Method for Open Restricted Queueing Networks. *Opns. Res.* **28,** 594–602.