Revised Jan. '88

A Survey of Two-node Queueing Networks with Blocking[1]

by

H.G. Perros

Department of Computer Science

and

Center for Communications and Signal Processing

North Carolina State University

Raleigh, NC 27695-8206

CCSP-TR-88/7
January 1988

[1]Supported in part by the National Science Foundation under Grants DCR-85-02540 and CCR-87-02258

Abstract

In recent years, queueing networks with blocking have been studied by researchers from various research communities such as Operations Research, Industrial Engineering, and Computer and Communications Performance Modelling. In view of this, related results are scattered throughout various journals. Furthermore, the lack of a comprehensive survey of these results has, in several instances, given rise to unnecessary duplication. In this paper, we give a survey of the literature related to two-node queueing networks with blocking. Only analytical and numerical results are discussed.

1. INTRODUCTION

Networks of queues with blocking have proved useful in modelling computer systems, distributed systems, telecommunication systems and flexible manufacturing systems. A queueing network with blocking can be thought of as a set of arbitrarily linked finite queues. Blocking occurs when the flow of units through one queue is momentarily stopped due to the fact that another queue has reached its capacity limitation.

Queueing networks with blocking are in general difficult to treat. Most of the techniques that are employed to analyze such queueing networks are in the form of approximations, numerical techniques, and simulation techniques. Results pertaining to queueing networks with blocking are scattered throughout various journals and conference proceedings. A bibliography of relevant papers was compiled by Perros [49]. Also, a survey of results related to closed queueing networks with blocking can be found in Onvural [44].

The objective of this paper is to survey analytical and numerical results related to open or closed two-node queueing networks with blocking. This is the simplest queueing network with blocking, and it consists of two nodes linked in tandem. A node consists of a queue served by one or more servers. The capacity of a node is equal to the total number of customers that can be accommodated in its queue plus those in service. In open networks, the first node may have finite or infinite capacity. The second node has always finite capacity. Blocking of a server in the first node occurs due to the finite capacity of the second node. In the case of closed queueing networks, the capacity of one node may be less or greater (or equal) to N, the total number of customers in the system, while the capacity of the other node is always less than N. Blocking of a server in a node will occur if the other node has a capacity less than N.

This paper is organized as follows. In section 2, we discuss various blocking mechanisms and their equivalence. Sections 3 to 5 deal with open two-node queueing networks with blocking. In particular, in section 3, we survey results related to exponential two-node queueing networks in which the second queue has a finite capacity. Section 4, contains results for the same queueing network assuming general service times. Section 5 deals with two-node queueing network assuming that each queue is served by unreliable servers. Finally, results for closed two-node queueing networks with blocking are presented in section 6.

2. BLOCKING MECHANISMS

Various blocking mechanisms have been considered in the literature so far. These blocking mechanisms arose out of various studies of real life systems, and they are distinct types of models for blocking, a fact that may be easily missed by a reader unfamiliar with the subject. Following Onvural and Perros [45], we classify the most commonly used blocking mechanisms for two-node queueing network as follows:

TYPE 1: A customer upon completion of its service at node i attempts to enter destination node j. If node j at that moment is full, the customer is forced to wait in front of its server at node i until there is a departure from node j. The server remains blocked for this period of time and it can not serve any other customers waiting in its queue.

TYPE 2: A customer in node i checks whether node j is full or not just before it starts its service. If node j is full, the server becomes blocked and it can not serve the customer. When a departure occurs from node j, the server becomes unblocked and the customer begins receiving service.

Depending on whether the blocked customer is allowed to occupy the position in front of the server when the server is blocked, we distinguish the following two subcategories: **TYPE 2.1:** Position in front of the server cannot be occupied when the server is blocked, and **TYPE 2.2:** Position in front of the server can be occupied when the server is blocked. The distinction is meaningful when modelling, for instance, production systems. Let us consider two queues in tandem with capacities m_1 and m_2. In Type 2.1, when there are m_2 customers in the second queue, there can be at the most $m_1 - 1$ customers in queue 1; but, in Type 2.2, there can be m_1 customers in queue 1.

TYPE 3: A customer upon service completion at node i attempts to join destination node j. If node j at that moment is full, the customer receives another service at node i. This is repeated until the customer completes a service at node i at a moment that the destination node is not full.

We note that in the above mentioned blocking mechanisms a server becomes unblocked when the number of customers in the destination node drops below its capacity. Latouche and Neuts [35] considered other extensions whereby unblocking of a server occurs when the number of units in the destination node drops below a predefined level, not necessarily equal to its capacity (see also Lavenberg [36]).

Equivalencies between these distinct types of blocking mechanisms have been obtained by Onvural and Perros [45], Onvural [43], Caseau and Pujolle [10], Altiok and Stidham [3] and Bocharov and Albores [7]. These equivalencies were obtained by assuming single server queues and exponential service times.

Now, let us consider an open exponential two-node queueing networks with blocking where each node is a single server queue (figure 1). The following equivalencies hold:

a) When the first queue is infinite, the open two-node queueing network has the same rate matrix under blocking mechanisms 2.1, 2.2 or 3. Furthermore, its rate matrix under type 1 blocking mechanism is identical to its rate matrix under type 2.1 (or 2.2, 3) blocking mechanism, if the capacity of its second queue is increased by one.

b) When the first queue is finite, the queueing network has the same rate matrix for blocking types 2.2 and 3. Also, its rate matrix under type 1 blocking is the same as its rate matrix under blocking type 2.1, if the capacity of the second queue is increased by one.

Similar equivalencies hold for two-node exponential closed queueing networks where each node is a single server queue (figure 3). In particular, the queueing network has the same rate matrix under blocking mechanisms 2.2 and 3. Furthermore, its rate matrix under type 1 blocking is the same as its rate matrix under type 2.1 blocking, if the capacities of both queues are increased by one.

3. EXPONENTIAL SERVICE TIMES

Let us consider two nodes in tandem as shown in figure 1. The first node has an infinite capacity and the second node has a finite capacity of size m. Service at each node is provided by a single exponential server. External arrivals join the first node at the rate λ with an exponentially distributed inter-arrival time. Let α, β be the mean service time at the first and second server respectively. Customers in each node are served in a FIFO manner.

Let us assume that the above queueing network operates under type 1 blocking mechanism. That is, the first server becomes blocked when a customer upon completion of its service finds the second node full. The server remains blocked, and it can not serve

any other customer waiting in its queue, until a service is completed at the second node.

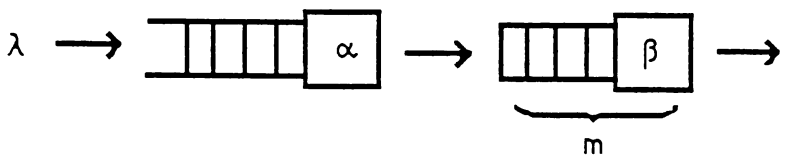


Figure 1: A two-node open queueing network with blocking

The state of the above queueing network can be described as (n_1, n_2), where n_1 indicates the number of units in the first node and n_2 the number of units in the second node. We have, $n_1 = 0.1,2,...,$ and $n_2 = 0,1, \cdots m, m+1$. The value $n_2 = m+1$ indicates that the second node is full and the first server is blocked. Thus, any state (i, m+1), i > 0, indicates that the first server is blocked and that there are i units in the first node. Of these i units, one has received its service but it is blocked from entering node 2, and the remaining i - 1 are waiting to be served.

For demonstration purposes we confine ourselves to the case of m=2. Let p_{ij} be the steady-state probability that the system is in state (i,j). The steady-state equations are as follows.

$$\begin{cases} \lambda p_{00} = (1/\beta)p_{01} \\ (\lambda+(1/\alpha))p_{i0} = (1/\beta)p_{i1} - \lambda p_{i-1,0} \quad , i \geq 1 \end{cases}$$

$$\begin{cases} (\lambda + (1/\beta))p_{01} = (1/\alpha)p_{10} + (1/\beta)p_{02} \\ (\lambda + (1/\alpha) + (1/\beta))p_{i1} = (1/\alpha)p_{i+1,0} + (1/\beta)p_{i2} + \lambda p_{i-1,1} \quad , i \geq 1 \end{cases}$$

$$\begin{cases} (\lambda + (1/\beta))p_{02} = (1/\alpha)p_{11} - (1/\beta)p_{13} \\ \lambda - (1/\alpha) - (1/\beta))p_{i2} = (1/\alpha)p_{i-1,1} - (1/\beta)p_{i-1,3} - \lambda p_{i-1,2} \quad , i \geq 1 \end{cases}$$

$$\begin{cases} \lambda - (1/\beta))p_{13} = (1/\alpha)p_{12} \\ (\lambda - (1/\beta))p_{i3} = (1/\alpha)p_{i2} + \lambda p_{i-1,3} \quad , i \geq 2 \end{cases}$$

We also have the normalizing equation $\Sigma p_{ij} = 1$. Now, define the following generating functions.

$$g_k(z) = \sum_{i=0}^{\infty} p_{ik} z^i, \quad k = 0,1,2$$

$$g_3(z) = \sum_{i=1}^{\infty} p_{i3} z^{i-1}$$

Then, the above set of equations can be written as follows

$$\begin{cases} g_0(z)[\lambda(1-z)+(1/\alpha)] = (1/\beta)g_1(z)+(1/\alpha)p_{00} \\ g_1(z)[\lambda(1-z)+(1/\alpha))+(1/\beta)] = (1/\alpha z)g_0(z)+(1/\beta)g_2(z)+(1/\alpha)p_{01}-(1/\alpha z)p_{00} \\ g_2(z)[\lambda(1-z)+(1/\alpha)+(1/\beta)] = (1/\alpha z)g_1(z)+(1/\beta)g_3(z)+(1/\alpha)p_{02}-(1/\alpha z)p_{01} \\ g_3(z)[\lambda(1-z)-(1/\beta)] = (1/\alpha z)g_2(z)-(1/\alpha z)p_{02} \end{cases} \tag{3.1}$$

The normalizing equation becomes $\sum_{k=1}^{3} g_k(1) = 1$.

The total number of independent equations is 6, namely the four equations given by (3.1), the normalizing condition, and equation $\lambda p_{00} = (1/\beta)p_{01}$ from the steady state equations. However, the total number of unknowns is 7, namely $g_0(z), g_1(z), g_2(z), g_3(z)$ and p_{00}, p_{01}, p_{02}. Thus, we are short of one equation. In the case of general m, the number of unknowns is 2m+3. This consists of m+2 unknown generating functions and m+1 unknown probabilities $p_{00}, p_{01}, \ldots, p_{0m}$. The number of available independent

equations is m+4, thus being m-1 equations short. In view of this, with the exception of the case of m=1, it has not been possible to obtain closed-form solutions for the generating functions $g_k(z)$, $k=0,1,...,m-1$.

Konheim and Reiser [28] (see also [29]) studied the above queueing network under blocking mechanism type 2 assuming that a customer departing from the second server may be fed back to the first queue. The authors obtained the additional equations necessary for the solution of this network using the fact that the generating function $g_m(z)$ is analytic within the unit circle. $g_m(z)$ is the generating function of the probabilities of the states in which the first server is blocked, and it is expressed as a rational polynomial, i.e. $g_m(z) = f(z)/h(z)$. In order for $g_m(z)$ to be analytic within the unit circle, all the zeroes of h(z) within the unit circle have to be zeroes of f(z) as well. This yields the additional necessary equations. The authors developed a lemma regarding the behavior of these roots, and described an algorithm for obtaining p_{ij}.

Pellaumail and Boyer [47] studied the same network assuming that the first queue is finite, and that departing customers from the network may be fed back to the first queue. The rate of external arrival as well as the service rate at the first and second server are dependent on the state (i,j) of the system. The authors devised a recursive procedure for obtaining the queue-length probability distributions.

3.1 Limiting Cases: Exact Analytical Results

In this section, we present some exact closed-form analytical results that have been obtained in certain limiting cases. Due to the equivalencies between blocking mechanisms 1,2 and 3, we will only consider the queueing network under type 1 blocking mechanism.

a) $\alpha \to 0$

Let us consider the queueing network shown in figure 1. In the case where $\alpha \to 0$, a unit arriving at the first node goes through it and joins the second node. If the second node already contains m units, the arriving (m+1st) unit will block the server of the first node. Now, if a unit departs from the second node the first server will become unblocked for an infinitely small time and then it will get blocked again, if there are units waiting in the first node. The queueing network is, therefore, reduced to an M/M/1 queue with traffic intensity $\rho = \lambda\beta$. Let p_i be the steady-state probability that there are i units in this M/M/1 queue. Then, the probability that there are i units in the second node of the queueing network is simply p_i, i=0,1...,m. The probability that there are i units in the first node of the queueing network is p_{m+1}.

This intuitive argument was given by Foster and Perros [16]. Reiser and Konheim [28] arrived at the same result using a rigorous mathematical approach.

b) Saturated first node

Let us consider the queueing network shown in figure 1, and let α_0 be the mean service time above which the first node is always saturated. Then, for $\alpha \geq \alpha_0$ the first server is either busy serving or blocked. A unit upon completion of its service at the first server gets blocked if at that moment the second node is full. The blocking unit remains in front of the first sever until a departure occurs from the second node. At that instance it moves to the mth position of the second node, and the first server becomes unblocked. During the blocking period, therefore, the first server can be seen as providing additional storage space to the second node. Furthermore, during the blocking period the first server does not serve any other units. That is, no more arrivals occur at the second node. Due to the exponential assumption, this is equivalent to saying that

arrivals do occur at the second node at the rate $1/\alpha$ but they are lost. In view of this, the second node becomes an M/M/1/m+1 queue with an overall arrival rate $1/\alpha$ and service rate $1/\beta$.

The above intuitive argument was given by Foster and Perros [16]. This result can be easily verified by studying the underlying Markov process (see Hatcher [22]). We note that this limiting case has been used to model two-stage production systems assuming that the two servers are unreliable. This model will be further examined in section 5.

Let us now consider the two-node queueing network with blocking assuming that each queue is served by multiple servers, as shown in figure 2. Let s_1 and s_2 be the number of servers at the first and second node respectively. Units arrive at the system in a Poisson fashion at the rate λ. Furthermore, let α and β be the mean service time of a server at the first and second node respectively. All service times are assumed to be exponentially distributed. Finally, let m ($m \geq s_2$) be the maximum capacity of the second node. That is, there can be s_2 units in service and $m - s_2$ units in its queue waiting for service.

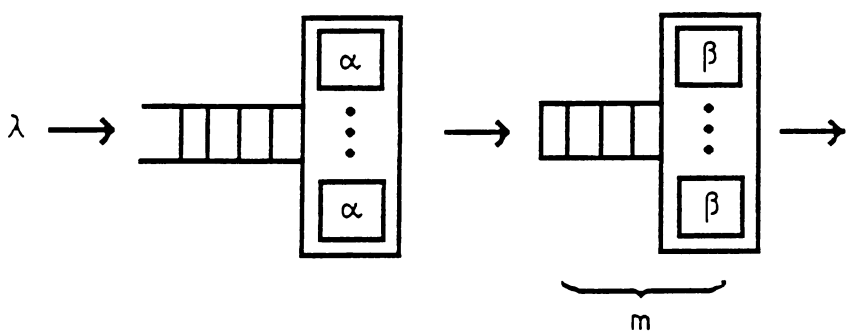


Figure 2: Two-node queueing network with blocking with multiple servers

A unit, upon completion of its service at the first node, will get blocked if at that moment the second node is full. Its server will also get blocked. When a departure occurs from the second node, one of the blocked customers will enter the second node and its associated server will become unblocked. It is not necessary to keep track in what order the servers at the first node get blocked, seeing that they are all identical. This model was first studied by Makino [37]. We now proceed to obtain a closed-form analytic solution of this queueing network, assuming that the first node is saturated.

Let p_n, $n = 0,1,\ldots,m, m-1,\ldots,m-s_1$, be the probability that there are n units in the second node including the units in the first node that are blocked. The steady-state equations associated with p_n are as follows:

$$s_1 \frac{1}{\alpha} p_0 = \frac{1}{\beta} p_1$$

$$\left(s_1 \frac{1}{\alpha} + i \frac{1}{\beta}\right) p_i = (i+1) \frac{1}{\beta} p_{i+1} + s_1 \frac{1}{\alpha} p_{i-1} \quad , \quad i = 1,2,\ldots,s_2 - 1$$

$$\left(s_1 \frac{1}{\alpha} + s_2 \frac{1}{\beta}\right) p_i = s_2 \frac{1}{\beta} p_{i+1} + s_1 \frac{1}{\alpha} p_{i-1} \quad , \quad i = s_2, \ldots, m$$

$$((s_1 - m - i) \frac{1}{\alpha} + s_2 \frac{1}{\beta}) p_i = s_2 \frac{1}{\beta} p_{i+1} + (s_1 - m - (i-1)) \frac{1}{\alpha} p_{i-1} \quad , i = m+1, \ldots m + s_1 - 1$$

$$s_2 \frac{1}{\beta} p_{m+s_1} = \frac{1}{\alpha} p_{m+s_1-1}$$

Using the above equations we can easily obtain

$$p_i = \begin{cases} \dfrac{s_1^i}{i!}\left(\dfrac{\beta}{\alpha}\right)^i p_0 & i = 1,2,\ldots,s_2 \\ \dfrac{s_1^i}{s_2!\,s_2^{i-s_2}}\left(\dfrac{\beta}{\alpha}\right)^i p_0 & i = s_2+1,\ldots,m \\ \dfrac{s_1^m (s_1)_{(i-m)}}{s_2!\,s_2^{i-s_2}}\left(\dfrac{\beta}{\alpha}\right)^i p_0 & i = m+1,\ldots,m+s_1 \end{cases} \tag{3.2}$$

where $(s_1)_k = s(s-1)(s-2)\cdots(s-(k-1))$. Using the normalizing condition we obtain

$$p_0 = 1/\left[\sum_{i=0}^{s_2} \frac{s_1^i}{i!}\left(\frac{\beta}{\alpha}\right)^i + \sum_{i=s_2+1}^{m} \frac{s_1^i}{s_2!\,s_2^{i-s_2}}\left(\frac{\beta}{\alpha}\right)^i + \sum_{i=m-1}^{m+s_1} \frac{s_1^m (s_1)_{(i-m)}}{s_2!\,s_2^{i-s_2}}\left(\frac{\beta}{\alpha}\right)^i\right] \tag{3.3}$$

We note that the departure rate from the second node is

$$\sum_{i=1}^{s_2-1} \left(\frac{i}{\beta}\right) p_i - \frac{s_2}{\beta} \sum_{i=s_2}^{m+s_1} p_i \tag{3.4}$$

which is equal to the departure rate from the first node

$$\sum_{i=0}^{m} \left(\frac{s_1}{\alpha}\right) p_i - \sum_{i=1}^{s_1-1} \left(\frac{s_1 - i}{\alpha}\right) p_{m+i} \tag{3.5}$$

where p_i , $i = 0,\ldots,m_1 + s_1$, is given by expressions (3.2) and (3.3). We can now use

expression (3.4) (or (3.5)) to obtain the condition for stability of this queueing network in the case where the first queue is assumed to be stable. In this case, expression (3.4) gives the maximum departure rate from the queueing network. Therefore, in order for the system to be stable, we should have

$$\lambda < \sum_{i=1}^{s_2-1} \left(\frac{i}{\beta}\right) p_i + \frac{s_2}{\beta} \sum_{i=s_2}^{m+s_1} p_i \quad .$$

The above expression agrees with the condition for stability derived by Latouche and Neuts [35] for the queueing network shown in figure 2 under one of the three blocking mechanisms that they considered. (In particular, for model B, $r^*=r$, $k^*=M-1+r-1$, and $\theta=0$). Finally, we note that Makino [37] also studied the output process of this queueing network.

c) The blocking probability

Let us consider the queueing network shown in figure 1. The blocking probability is the probability that a unit upon service completion at the first server will be blocked. That is, at that instant the second node is full. This probability (call it π) has been used in several approximation algorithms developed for the analysis of queueing networks with blocking (see for instance Perros and Snyder [48], Labetoulle and Pujolle [32], and Takahashi, Miyahara, and Hasegawa [56]). Below, we obtain the exact expression for this probability when $\alpha \to 0$ (call it π_1) and when $\alpha \geq \alpha_0$ (call it π_2) following arguments given in Foster and Perros [16]

As it was mentioned above, when $\alpha \to 0$ the queueing network behaves as an M/M/1 queue. Therefore,

$$\pi_1 = \sum_{i \geq m} p_i = (1-\rho) \sum_{i \geq m} \rho^i = \rho^m \quad ,$$

where $\rho = \lambda\beta$. Now, in the case where $\alpha \geq \alpha_0$ the second node behaves as an M/M/1/m+1 system. The effective arrival rate λ' at the second node is

$$\lambda' = (1/\alpha)(1-p_{m+1}) \tag{3.6}$$

where

$$p_{m+1} = \frac{(1-\sigma)\sigma^{m+1}}{1-\sigma^{m+2}} \quad ,$$

σ being equal to β/α. Using Little's relation we obtain

$$\lambda' \pi_2 \beta = p_{m+1}$$

or

$$(1/\alpha)(1-p_{m+1})\ \pi_2\ \beta = p_{m-1}$$

or

$$\sigma \left(1 - \frac{(1-\sigma)\sigma^{m+1}}{1-\sigma^{m+2}}\right) \pi_2 = \frac{(1-\sigma)\sigma^{m+1}}{1-\sigma^{m+2}}$$

or

$$\pi_2 = \frac{(1-\sigma)\sigma^m}{1-\sigma^{m+1}}$$

$$= \frac{\sigma^m}{1-\sigma+\cdots+\sigma^m} \tag{3.7}$$

The quantity α_0 is derived below in section 3.3 in connection with the condition for stability of the queueing network.

The quantity $1-\pi_2$ can be seen as the percentage of time that the first server is busy serving. This quantity was first derived by Hunt [23]. For the special case $\alpha = \beta$ we have that $1-\pi_2 = m/(m+1)$. We note that Tsiotras [57] reported on a recursive procedure for obtaining the blocking probability for m=2.

d) Processor-sharing discipline

We now analyze the queueing network shown in figure 1 under the processor-sharing discipline. In particular, we consider the case in which a unit cycles infinitely quickly between the two servers receiving an infinitesimal amount of service at each server. In this case, it is possible to obtain closed-form solutions (see Asare [5], Konheim and Reiser [28]).

Let us consider the above two-node network, and let us assume that a unit upon completion of its service in the second queue may either depart or it may join the first queue with probability 1-θ and θ respectively. Then, this processor-sharing discipline can be obtained as the limiting case $\theta \to 1$, $\alpha \to 0$, $\beta \to 0$ such that $\alpha/(1-\theta)$ and $\beta/(1-\theta)$ remain finite. The quantities $\alpha/(1-\theta)$ and $\beta/(1-\theta)$ can be seen as the mean service requirement of a unit at server one and two respectively. For presentation purposes, we obtain this limiting case for m=2. Let $\rho_1 = \lambda\alpha/(1-\theta)\rho_2 = \lambda\beta/(1-\theta)$. Then, the steady-state equations are as follows:

$$\begin{cases} \rho_2 p_{00} = p_{01} & (3.8) \\ (\rho_1\rho_2(1-\theta) - \rho_2)\, p_{i0} \quad \rho_1\rho_2(1-\theta)\, p_{i-1,0} - \rho_1(1-\theta)\, p_{i1}\rho_1\theta p_{i-1,1} \;,\; i \geq 1 & (3.9) \end{cases}$$

$$\begin{cases} (\rho_1\rho_2(1-\theta)(-\rho_1)\, p_{01} = \rho_2 p_{10} + \rho_1(1-\theta)\, p_{02} & (3.10) \\ (\rho_1\rho_2(1-\theta) - \rho_1 - \rho_2)\, p_{i1} = \rho_2 p_{i+1,0} + \rho_1(1-\theta)p_{i2} - \rho_1\theta p_{i-1,2} - \rho_1\rho_2(1-\theta)p_{i-1,1} \;,\; i \geq 1 & (3.11) \end{cases}$$

$$\begin{cases} (\rho_1\rho_2(1-\theta) + \rho_1)\, p_{02} = \rho_2 p_{11} - \rho_1(1-\theta)\, p_{13} & (3.12) \\ (\rho_1\rho_2(1-\theta) + \rho_1 + \rho_2)\, p_{i2} = \rho_2 p_{i+1,1} - \rho_1(1-\theta) p_{i-1,3} + \rho_1\theta p_{i3} - \rho_1\rho_2(1-\theta)\, p_{i-1,2}\ , \quad i \geq 1 & (3.13) \end{cases}$$

$$\begin{cases} (\rho_1\rho_2(1-\theta) + \rho_1)\, p_{13} = \rho_2 p_{12} & (3.14) \\ (\rho_1\rho_2(1-\theta) + \rho_1)\, p_{i3} = \rho_2 p_{i2} + \rho_1\rho_2(1-\theta)\, p_{i-1,3} \quad , i \geq 2 & (3.15) \end{cases}$$

We first observe that we can obtain equations which are independent of θ. In particular, by adding equations (3.8), (3.9) with i=1, we obtain:

$$\rho_2(p_{10} + p_{01}) = p_{11} + p_{02} \tag{3.16}$$

Likewise, adding equations (3.9) with i=2, (3.11) with i=1, and (3.12) gives

$$\rho_2(p_{20} + p_{11} + p_{02}) = p_{21} + p_{12} + p_{13} \quad . \tag{3.17}$$

Following this scheme we can obtain

$$\rho_2(p_{i0} + p_{i-1,1} + p_{i-2,2} + p_{i-2,3}) = p_{i1} + p_{i-1,2} + p_{i-1,3} \quad , i \geq 3 \tag{3.18}$$

We now substitute equations (3.9) by the above equations (3.16), (3.17), and (3.18). Thus, our original queueing network is described by equations (3.16) to (3.18), (3.8) and (3.10) to (3.15). We now consider the limiting case of these equations by letting $\alpha \to 0$, $\beta \to 0$, $\theta \to 1$ such that $\alpha/(1-\theta)$ and $\beta/(1-\theta)$ remain finite. In this case, equations (3.10) to (3.15) are significantly simplified so that we can obtain the following.

$$\begin{cases} p_{i3} = \left(\dfrac{\rho_2}{\rho_1}\right)^3 p_{i+2,0} & ,i \geq 1 \\ p_{i2} = \left(\dfrac{\rho_2}{\rho_1}\right)^2 p_{i+2,0} & ,i \geq 0 \\ p_{i1} = \left(\dfrac{\rho_2}{\rho_1}\right) p_{i+1,0} & ,i \geq 0 \end{cases} \tag{3.19}$$

Using (3.19), (3.8) and (3.16) to (3.18) we obtain

$$p_{i0} = \rho_1^i p_{00} \quad . \tag{3.20}$$

The normalizing equation gives

$$p_{00} = \frac{(1-\rho_1)(1-\rho_2)}{1-\rho_2^4} \tag{3.21}$$

Thus,

$$p_{ij} = \begin{cases} \left[\rho_1^i(1-\rho_1)\right] \cdot \left[\rho_2^j \dfrac{1-\rho_2}{1-\rho_2^4}\right] & , \quad j \leq 2 \ , \ i \geq 0 \\ \left[\rho_1^{i-1}(1-\rho_1)\right] \cdot \left[\rho_2^j \dfrac{1-\rho_2}{1-\rho_2^4}\right] & , \quad j = 3, \quad i \geq 1 \end{cases} \tag{3.22}$$

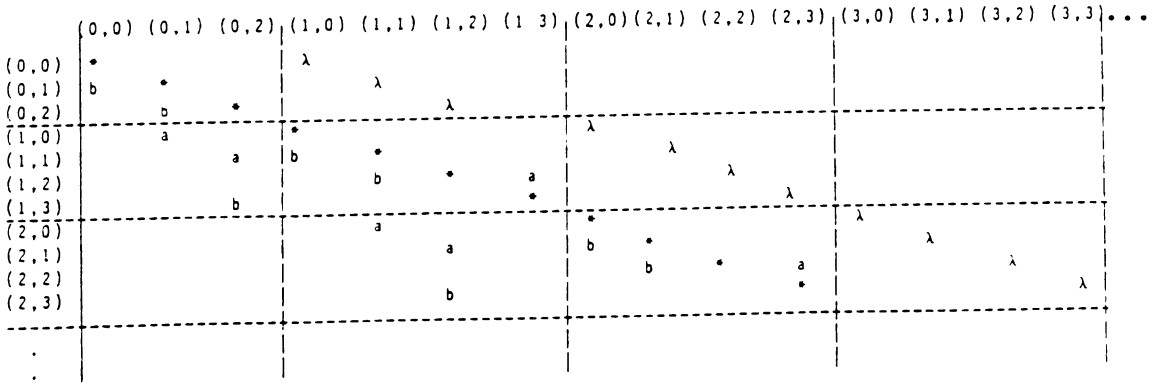
Using the above expression, it can be easily shown that the marginal queue-length probability distribution of the second node is that of an M/M/1/3 queue with an arrival rate $\lambda/(1-\theta)$ and mean service time β. Likewise, the marginal queue-length distribution of the first node is that of an M/M/1 queue with an arrival rate $\lambda/(1-\theta)$ and mean service time α, assuming that a blocking unit is not counted as being part of the first node.

3.2 Numerical Methods: Matrix-Geometric Solution

The numerical analysis of queueing networks with blocking (or, of any queueing system for that matter) involves the following three tasks: a) generation of the state space, b) generation of the rate matrix Q, and c) solution of the linear system $\underline{x}Q=0$ in order to calculate the stationary probability vector $\underline{x}$. Now, let us consider the queueing network shown in figure 1 under type 1 blocking mechanism. Let (n_1,n_2) be the state of the system as described at the beginning of section 3. If these states are enumerated lexicographically, the rate matrix Q will have the following block tri-diagonal form.

$$Q = \begin{bmatrix} A' & B' & & & & \\ C' & A & B & & & \\ & C & A & B & & \\ & & C & A & B & \\ & & & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot \end{bmatrix}$$

For instance, for m=2, the rate matrix Q is as follows



where $b=1/\beta$ and $a=1/\alpha$. For each row, the * on the diagonal indicates the negative

sum of all the non-zero elements in the row.

In view of the structure of the rate matrix, the stationary probability vector can be obtained using the matrix-geometric procedure (see Neuts [40]). In particular, let $\underline{x}$ be the vector of the steady-state probabilities associated with Q, i.e., $\underline{x}Q = 0$ and $\underline{x}\,\underline{e} = 1$. We partition $\underline{x}$ conformally with the blocks of matrix Q. That is, $\underline{x} = (\underline{x}_0, \underline{x}_1, \cdots \underline{x}_n, \cdots)$, where subvector $\underline{x}_0$ has m+1 components and subvector $\underline{x}_i$ (i>0) has m+2 components. We have

$$\underline{x}_0 A' + \underline{x}_1 C' = 0$$

$$\underline{x}_0 B' + \underline{x}_1 A + \underline{x}_2 C = 0$$

$$\underline{x}_i B + \underline{x}_{i-1} A + \underline{x}_{i+2} C = 0 \quad , \quad i \geq 1 \ .$$

Vectors $\underline{x}_i$, $i \geq 1$, can be calculated as follows

$$\underline{x}_i = \underline{x}_1 R^i \quad , i \geq 1$$

where R is the minimal non-negative solution of the matrix equation

$$B + RA + R^2 C = 0$$

Subvectors $\underline{x}_0$ and $\underline{x}_1$ can be obtained from the equations

$$\begin{cases} \underline{x}_0 A' + \underline{x}_1 C' = 0 \\ \\ \underline{x}_0 B' + \underline{x}_1 (A + RC) = 0 \end{cases}$$

and the normalizing condition.

Matrix-geometric solutions of queueing networks with blocking as shown in figure 2 were investigated by Latouche and Neuts [35]. They considered three different blocking

mechanisms based on the following two notions: a) The blocking of a server at the first node may or may not cause the remaining unblocked servers to become blocked. b) Unblocking of a server (or servers) may occur when the number of units in the second node drops below a predefined number. The authors studied the above queueing system by allowing a feedback loop of departures from the second node to the first node. They observed, that the rate matrix of the queueing network under any of the three blocking mechanisms, has a block tri-diagonal structure of the form

$$\begin{bmatrix} A_{01} & A_{02} & & & & & & \\ A_{10} & A_{11} & A_{12} & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & A_{s_1-2,0} & A_{s_1-2,1} & A_{s_1-2,2} & & & \\ & & & A_{s_1-1,0} & A_{s_1-1,1} & A_2 & & \\ & & & & A_0 & A_1 & A_2 & \\ & & & & & A_0 & A_1 & A_2 \\ & & & & & & \ddots & \ddots \end{bmatrix} \qquad (3.23)$$

The matrix-geometric procedure is an efficient procedure for analyzing two servers in tandem with a finite intermediate queue. This procedure has been employed by Onvural, Perros and Altiok [46] to analyze larger, configurations of exponential queueing networks with blocking. Time and space complexity problems arise when analyzing networks with more than four queues arbitrarily linked.

We note that other numerical methods, such as the power method, can be also used to efficiently calculate the stationary probability vector (see, for example Van Dijk and Lamond [58]). Finally, Wong, Giffin, and Disney [60] analyzed the two-node queueing network with blocking numerically. They observed that the solution can be expressed in terms of a particular matrix. The steady-state joint queue-length distribution was then

obtained in terms of the eigenvalues and eigenvectors of this matrix.

3.3 Condition for Stability

Let us consider the queueing network shown in figure 1. In section 3.1, the quantity α_0 was defined to be the critical mean service time of the first server above which the first queue unstable. Now if the first node is stable, then the effective rate into the second node is λ. However, if $\alpha \geq \alpha_0$ then the effective rate λ' is given by (3.6). Now, for $\alpha = \alpha_0$ we have $\lambda = \lambda'$ or

$$\lambda = \frac{1}{\alpha_0}\left(1 - \frac{(1-\sigma_0)\sigma_0^{m-1}}{1-\sigma_0^{m-2}}\right) \tag{3.24}$$

where $\sigma_0 = \beta/\alpha_0$. The quantity α_0 can be computed numerically using (3.24).

The condition for stability can be simply expressed as $\alpha < \alpha_0$. Now, if $\alpha \geq \alpha_0$, then $\lambda' \leq \lambda$ or using (3.6) we have $(1/\alpha)(1-p_{m+1}) \leq \lambda$ Thus, for $\alpha < \alpha_0$ we have $(1/\alpha)(1-p_{m+1}) > \lambda$ or

$$\lambda < \frac{1}{\alpha}\left(1 - \frac{(1-\sigma)\sigma^{m+1}}{1-\sigma^{m+2}}\right)$$

or

$$\lambda\alpha < \frac{1+\sigma+\cdots+\sigma^m}{1+\sigma+\cdots+\sigma^{m+1}} \tag{3.25}$$

This intuitive argument was given by Foster and Perros [16]. The above condition for stability has been derived by several other authors (see Asare [5], Bocharov and Albores [7], Konheim and Reiser [28], Latouche and Neuts [35], Pujolle and Potier [52], Hunt [23] and Lavenberg [36]). We note that Lavenberg [36] formally demonstrated that when the first queue is saturated, the rate of departures from the queueing network is equal to the supremum of the set of arrivals for which the network is stable. Furthermore, this

departure rate is equal to the supremum of the set of departure rates over all finite arrivals.

The stability condition of the queueing network shown in figure 2 can be obtained matrix-geometric procedure. The general form of this condition is given by (see Neuts [40])

$$\pi A_0 \underline{e} > \pi A_2 \underline{e} \tag{3.26}$$

where π is obtained solving $\pi (A_0 + A_1 + A_2) = 0$ and A_0, A_1, and A_2 are the block sub-matrices in the matrix given by (3.23).

3.4 Other references

We conclude this section by giving some additional relevant references. Clarke [11],[12],[13],[14] examined the following problem. Consider the two-node queueing network shown in figure 1, assuming for the moment that m=1, i.e. no intermediate queue. A unit receives a service either at the first server or at the second server, but not at both servers. A unit that arrives at the queueing network at a time that the network is empty, starts its service at the second server. The next arrival starts its service at the first server. Now if the unit at the first server completes its service first, the unit will get blocked seeing that it cannot exit through the second server, which is currently busy serving. During the time that the unit waits to exit from the queueing network, its server is also blocked. A busy rear server can also block access to a free server ahead. This type of blocking depicts situations, in which the units are forced by the physical layout to proceed in a single path, with servers located in tandem along the path, even though each unit requires only one service operation. Clarke analyzed the above model assuming that the inter-arrival times and service times are exponentially distributed. He

obtained the condition for stability for s (s≥2) servers in tandem with no intermediate queue [11], and for two servers in tandem with an arbitrary finite intermediate queue [12]. In [14] he obtained expressions related to the waiting time of a unit in the system consisting of two servers with no intermediate queue. Finally, in [13] he extended the results obtained in [11] to the case of general service times. The model analyzed in [12] was shown by Neuts [40] to have a matrix-geometric solution. Boxma and Konheim [8] analyzed approximately the queueing network shown in figure 1, assuming that the first queue is finite or infinite. Service times and inter-arrival times were assumed to be exponentially distributed. Type 2 blocking mechanism was assumed.

Asare [5] studied the queueing network shown in figure 2 assuming type 1 blocking mechanism. He obtained a closed-form solution of the queue-length probability distribution of the number of units in the second queue by ignoring some of the terms that appear in the steady-state equations. This approximate solution becomes exact in the case of the processor-sharing discipline outlined in section 3.1. The same model was also studied by Langaris and Conolly [33]. Their approach is similar to the method employed by Reiser and Konheim [28]. The waiting time process of this model was also analyzed by the same authors in [34]. Finally, Krishna and Shin [31] analyzed the queueing network shown in figure 2, assuming $s_1 = 1$ and $m = s_2$.

4. GENERAL SERVICE TIMES

Let us consider the queueing network shown in figure 1 assuming that the service times at the first and second server follow arbitrary distributions. Neuts [39] studied this model assuming general service times at the first server, and exponential service times at the second server under blocking mechanism type 1. The model was studied in

terms of a semi-Markov process imbedded at instances of service completion at the first server. Most of the results obtained are purely formal. A less general case was considered by Suzuki [55], Avi-Itzhak and Yadin [6] and Prabhu [51]. In particular, these authors analyzed the model studied by Neuts [39], assuming no intermediate waiting room, i.e., m=1. The transient behavior of this model was analyzed in [51], and its steady-state behavior was studied in [55] and in [6]. In the following section, we sketch a solution to this model following Avi-Itzhak and Yadin [6].

4.1 No Intermediate Waiting Room

Let us consider the queueing network shown in figure 1 assuming that the second node has a capacity equal to one, that is it can accommodate one unit, the unit in service. Units arrive at the first queue in a Poisson fashion with parameter λ. The service times at server 1 and 2, s_1 and s_2 respectively, are independent and arbitrarily distributed with probability density function (pdf) $f_1(\cdot)$ and $f_2(\cdot)$. Server 1 operates under blocking mechanism type 1. Define a random variable T as $T = \max(s_1, s_2)$. Let $f_T(\cdot)$ be its pdf. We have

$$f_T(t) = f_1(t)F_2(t) + f_2(t)F_1(t) \tag{4.1}$$

where $F_i(t) = \int_0^t f_i(s)ds$, $i = 1.2$. Expression (4.1) can be intuitively interpreted as follows. $f_T(t)$ equals $f_1(t)$ if the service at the second station is less than t (with probability $F_2(t)$) or vice versa. Any unit that arrives at the queueing network during the busy period of the first server, spends a period of time T = max (s_1, s_2) at server 1. (Here, the first server is considered busy if it is either busy serving or blocked.) However, the unit that begins the busy period of the first server will spend an amount of time, t_0, in front of the first server which is different than T. This is the time elapsing

from the arrival of the unit which starts the busy period until the unit enters the second server. Let $f_0(\cdot)$, be the pdf of the random variable t_0. Then, it can be shown that

$$f_0(t) = f_1(t)F_2(t) + e^{\lambda t}(\lambda F_1(t) - f_1(t)) \int_t^{\infty} e^{-\lambda s} f_2(s)ds \tag{4.2}$$

Having obtained $f_T(t)$ and $f_0(t)$, we can now analyze the first queue as an M/G/1 queue in which the first customer of each busy period receives an exceptional service. Using the known results (see references in [6] and also in [51]) we can obtain the Laplace transform $d^*(z)$ of the time a unit spends in the queue and also in front of the first server. Hence, we have

$$d^*(z) = \frac{1-b}{1-b+\lambda E(t_0)} \cdot \frac{(z+\lambda)f_0^*(z) - \lambda f_T^*(z)}{z+\lambda-\lambda f_T^*(z)} \tag{4.3}$$

where $f_0^*(z)$ and $f_T^*(z)$ are the Laplace transforms of the pdf $f_0(\cdot)$ and $f_T(\cdot)$, and b = λ E (T). The Laplace transform $h^*(z)$ of the time a unit spends in the whole queueing network is

$$h^*(z) = d^*(z)f_2^*(z) \tag{4.4}$$

where $f_2^*(z)$ is the Laplace transform of the pdf $f_2(\cdot)$. We note that the authors also obtain the generating function of the probability distribution of the numbers of units in front of server 2 (see [6] for further details).

4.2 A Matrix-Geometric Solution

In general, exact closed form analytical solutions of the two-node queueing network shown in figure 1 with arbitrary service time distributions are not easily attainable. However, if the service time at each server has a phase type distribution, then it is possible to analyze this model numerically using the matrix-geometric procedure outlined in

section 3.2. The efficiency of this procedure, in this case, has not been investigated. Gun and Makowski [21] obtained a closed form matrix-geometric solution for the network shown in figure 1 with phase type service distributions, assuming the first node is saturated. In particular, they obtained an expression for matrix R in terms of two known matrices. This matrix R is not the same as the R matrix mentioned in section 3.2.

4.3 Other References

We conclude this section by giving some additional references. Rao [53,54] obtained the mean effective service time at the first server assuming arbitrary service times, an unlimited supply of units in the first node (i.e. saturated node), type 1 blocking mechanism, and an intermediate waiting room of size $m \geq 0$. Pinedo and Wolf [50] considered two nodes in tandem with and without blocking. They compared the expected waiting time in the case where any given customer has independent service times at the two servers with the expected waiting time in the case where any given customer has equal service times at the two servers. Newell [41] analyzed two finite nodes in tandem using diffusion approximations. The service times at both nodes and the inter-arrival times were assumed to have the same mean but different variances. Each node was treated as if it were a stochastic continuous fluid. Finally, Van Dijk and Lamond [58] considered a two-node queueing network with blocking, where the first node has a finite capacity. For this system, they gave a lower and an upper bound of the probability that the first node is full using the job-balance property. This property guarantees a product-form solution. The bounds are insensitive to the service distributions for certain service disciplines.

5. UNRELIABLE SERVERS

We now consider the case of unreliable servers in connection with the queueing network studied in section 3 and shown in figure 1. In this case, a server in operation is allowed to fail. The server becomes operational again after it has been repaired. Below, we examine this model assuming a) exponentially distributed service times, and b) constant synchronized service times.

5.1. Exponential Service Times

In this section, we examine the case of unreliable servers assuming exponentially distributed service times when a) the first node is saturated, and b) when it is stable.

a) Saturated First Node

Let us consider the model shown in figure 1 assuming that the first node is saturated, i.e. there is always one unit in the node waiting to be served. We analyze this model following Gershwin and Berman [18]. A server is allowed to fail only during the time that it is busy serving. During the time a server is idle or blocked it cannot fail. A unit in service is not lost if during its service the server fails. Type 2 blocking is assumed.

The state of the system is described by the vector (n, s_1, s_2), where n is the number of units in the second node and s_1, s_2 is the status of the first and second server respectively. The quantity $s_i, i = 1,2$, takes the value 1 if the ith server is operational or 0 if it is broken down. A server is considered to be operational if it is busy serving, or idle, or blocked (first server only). Let r_1, p_1 and r_2, p_2 be the repair rate and failure rate of the first and second server respectively. Failure and repair times are assumed to be exponentially distributed. For presentation purposes, we analyze this system for the

case when m=3. The steady-state equations are as follows:

$$\begin{cases} (r_1+r_2)p_{000} = p_1 p_{010} \\ (r_1+r_2)p_{i00} = p_1 p_{i10} + p_2 p_{i01} \quad , \ i=1,2 \\ r_1+r_2)p_{300} = p_2 p_{301} \end{cases} \tag{5.1}$$

$$\begin{cases} (p_1-(1/\alpha)-r_2)p_{010} = r_1 p_{00} \\ (p_1-(1/\alpha)-r_2)p_{i10} = r_1 p_{i00} + p_2 p_{i11} + (1/\alpha)p_{i-1,1,0} \quad .i=1,2 \\ r_2 p_{310} = r_1 p_{300} + p_2 p_{311} + (1/\alpha)p_{210} \end{cases} \tag{5.2}$$

$$\begin{cases} r_1 p_{001} = r_2 p_{000} + p_1 p_{011} + (1/\beta)p_{101} \\ (r_1+(1/\beta)+p_2)p_{i01} = r_2 p_{i00} + p_1 p_{i11} + (1/\beta)p_{i-1,0,1} \quad , \ i=1,2 \\ (r_1-(1/\beta)+p_2)p_{301} = r_2 p_{300} \end{cases} \tag{5.3}$$

$$\begin{cases} (p_1-(1/\alpha))p_{011} = r_1 p_{001} - r_2 p_{010} + (1/\beta)p_{111} \\ (p_1+(1/\alpha)+p_2+(1/\beta))p_{i11} = r_1 p_{i01} - r_2 p_{i10} + (1/\beta)p_{i-1,1,1} - (1/\alpha)p_{i-1,1,1} \ .i=1,2 \\ (p_2+(1/\beta))p_{311} = r_1 p_{301} + r_2 p_{310} - (1/\alpha)p_{211} \end{cases} \tag{5.4}$$

The normalizing condition is: $\sum p(n,s_1,s_2) = 1$

Gershwin and Berman [18] observed that the above steady-state equations can be easily manipulated to obtain the following results.

a) Let us consider the equations involving only $p_{000}, p_{010}, p_{300}$ and p_{301}. We can easily establish that $p_{000} = p_{010} = p_{300} = p_{301} = 0$

b) Adding all the equations involving the probabilities $p(n,0,s_2)$ on the left hand side gives

$$r_1 \sum_{n=0}^{3} \sum_{s_2=0}^{1} p(n,0,s_2) = p_1 \sum_{n=0}^{2} \sum_{s_2=0}^{1} p(n,1,s_2) \quad .$$

Likewise, by adding all the equations involving the probabilities $p(n_1,s_1,0)$ on the left hand side we can establish that

$$r_2 \sum_{n=0}^{3} \sum_{s_1=0}^{1} p(n,s_1,0) = p_2 \sum_{n=1}^{3} \sum_{s_1=0}^{1} p(n,s_1,1)$$

The above two expressions imply that r_i * prob [server i is broken down]= p_i * prob [server i is busy serving] , i=1,2. That is, the rate at which a broken down server is repaired equals the rate at which it breaks down when it is busy serving, an intuitively obvious result.

c) Adding all the equations involving the probability $p(0,s_1,s_2)$ on the left hand side gives $(1/\alpha)[p(0,1,0) + p(0,1,1)] = (1/\beta)[p(1,0,1) + p(1,1,1)]$. Likewise, adding all the equations which involve $p(n,s_1,s_2)$, n=1,2,...,m, gives

$$(1/\alpha)[p(n,1,0) + p(n,1,1)] = (1/\beta)[p(n+1,0,1)+p(n+1,1,1)] \quad ,n=0,1,...,m-1 \quad .$$

This implies that rate of arrivals at the second node, given that there are n units in

the second node, equals the departure rate from the second node, given that there are n+1 units in the second node.

Gershwin and Berman hypothesized that the probability $p(n,s_1,s_2)$ is of the form

$$p(n,s_1,s_2) = c\ X^n\ Y^{s_1}\ Z^{s_2}$$

The quantities X,Y,Z and c were obtained by substituting the above expression into the steady-state equations. The computation of these quantities requires the numerical derivation of the zeroes of a fourth order polynomial.

b) Stable First Node

Now, let us consider the queueing network studied above assuming that the first node is stable. In particular, let us assume that the first queue is infinite and that arrivals occur at this queue in a poisson fashion at a constant rate. Let the vector (n_1,s_1,n_2,s_2) describe the state of the system, where n_1 and n_2 indicate the number of customers in the first and second node respectively, and s_1,s_2 indicate the status of the first and second server respectively. The quantity s_i, i=1,2, takes the value 1 if server i is operational or 0 if the server is broken down. Then, it can be easily verified that if we order the states of this system lexicographically, the non-zero elements of the resulting rate matrix have a block tri-diagonal structure. In view of this, the queueing network can be easily analyzed numerically.

5.2 Constant Synchronized Service Times: Saturated First Node

In this section, we examine the case of unreliable servers as in section 5.1 assuming constant service times. In particular, we consider the queueing network shown in figure 1 under a common fixed cycle time, whereby both servers start service at the beginning

of a cycle and end service at the end of the same cycle. Thus, both servers provide a constant service equal to the duration of a cycle. During a cycle, an operational server may break down with some fixed probability. A server that is broken down at the beginning of a cycle may be repaired during the cycle with some fixed probability. Thus, the only source of randomness in this model is due to the unreliability of the servers. This type of model has been motivated from studies of synchronized assembly machines.

Below, we examine this model assuming that there is an unlimited supply of units in the first node (i.e., the first queue is always saturated). We also make the following assumptions.

a) If server i, i=1,2, is operational at the beginning of a cycle, then at the end of the same cycle it may break down with probability b_i, or it may be still operational with probability $\bar{b}_i = 1 - b_i$.

b) If server i , i=1,2, is broken down at the beginning of a cycle, then at the end of the same cycle it may be fixed (i.e. become operational) with probability c_i, or it may be still broken down with probability $\bar{c}_i = 1 - c_i$.

c) A server may break down at any time, i.e., when it is busy serving, idle, or blocked (first server only).

d) When a server breaks down, the unit under service is assumed to be fully served.

e) The maximum number of units that can be accommodated in the second node is m. The second queue can accommodate at the most m-1 units, and the server can accommodate a unit in front of it whether it is operational or broken down.

f) The first server operates under blocking mechanism type 1.

Under the above assumptions, the state of the system can be described by the vector (s_1, s_2, n), where n is the number in the second queue and s_i is the state of the ith server, i=1,2. We have $0 \le n \le m$, and $s_i = 0,1,2,3$. The first server, at any instance, may be busy serving ($s_1 = 0$), broken down ($s_1 = 1$), blocked but not broken down ($s_1 = 2$), and, blocked and broken down ($s_1 = 3$). Similarly, the second server, at any instance, may be busy serving ($s_2 = 0$), broken down and having a unit in front of it ($s_2 = 1$), operational but without having a unit in front of it ($s_2 = 2$), and, broken down without having a unit in front of it ($s_2 = 3$).

Now, let us group the states of the system so that they all have the same number of units in the second node. The states within each group are arranged lexicographically. Groups are arranged in increasing order of the number of units in the second node. For instance, for m=5 we obtain the following arrangement of states.

(0)	020,	030,	120,	130
(1)	000,	010,	100,	110
(2)	001,	011,	101,	111
(3)	002,	012,	102,	112
(4)	003,	013,	103,	113
(5)	004,	014,	104,	114
(6)	204,	214,	304,	314

Now, let us construct the one-step transition probability matrix P, assuming that a cycle is equal to a unit time. Then, the non-zero elements of P have the following familiar block tri-diagonal structure.

$$P = \begin{bmatrix} A' & B' & & & & & \\ C & A & B & & & & \\ & C & A & B & & & \\ & & C & A & B & & \\ & & & C & A & B & \\ & & & & C & A & B \\ & & & & & C'' & A'' \end{bmatrix}$$

where

$$A' = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ c_1\bar{b}_2 & c_1 b_2 & \bar{c}_1\bar{b}_2 & \bar{c}_1 b_2 \\ c_1 c_2 & c_1\bar{c}_2 & \bar{c}_1 c_2 & \bar{c}_1\bar{c}_2 \end{bmatrix} \quad B' = \begin{bmatrix} \bar{b}_1\bar{b}_2 & \bar{b}_1 b_2 & b_1\bar{b}_2 & b_1 b_2 \\ \bar{b}_1 c_2 & \bar{b}_1\bar{c}_2 & b_1 c_2 & b_1\bar{c}_2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$C = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ c_1\bar{b}_2 & c_1 b_2 & \bar{c}_1\bar{b}_2 & \bar{c}_1 b_2 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad A = \begin{bmatrix} \bar{b}_1\bar{b}_2 & \bar{b}_1 b_2 & b_1\bar{b}_2 & b_1 b_2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ c_1 c_2 & c_1\bar{c}_2 & \bar{c}_1 c_2 & \bar{c}_1\bar{c}_2 \end{bmatrix}$$

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \bar{b}_1 c_2 & \bar{b}_1\bar{c}_2 & b_1 c_2 & b_1\bar{c}_2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad C'' = \begin{bmatrix} \bar{b}_1\bar{b}_2 & \bar{b}_1 b_2 & b_1\bar{b}_2 & b_1 b_2 \\ 0 & 0 & 0 & 0 \\ c_1\bar{b}_2 & c_1 b_2 & \bar{c}_1\bar{b}_2 & \bar{c}_1 b_2 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

and

$$A'' = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \bar{b}_1 c_2 & \bar{b}_1\bar{c}_2 & b_1 c_2 & b_1\bar{c}_2 \\ 0 & 0 & 0 & 0 \\ c_1 c_2 & c_1\bar{c}_2 & \bar{c}_1 c_2 & \bar{c}_1\bar{c}_2 \end{bmatrix}$$

The stationary probability vector can now be easily obtained (see for instance Neuts [40]).

We note that this model was first analyzed by Artamonov [4] who obtained the closed-form analytic solution by manipulating the steady-state equations algebraically.

5.3 Other References

The queueing network analyzed in section 5.1 with a saturated first node was also studied by Buzacott [9] assuming a constant failure probability per service. Koenigsberg [27] studied the same system assuming that the time between two successive failure of a server is exponentially distributed. Ignall and Silver [24] considered the case where each node is served by multiple servers. They gave a heuristic procedure for estimating the hourly output rate. Altiok [2] analyzed numerically the system studied in section 5.1 assuming that the service and repair times have a phase type distribution. The time between two successive failures was assumed to be exponentially distributed. Upon completion of a repair, service resumes from the phase where the server was when it broke down. He devised an efficient way of constructing the rate matrix. The stationary probability vector was obtained using the power method. For the same system, Gun [20] obtained a closed form matrix-geometric solution assuming phase-type operational and repair distributions. Finally, Wijngaard [59] considered the case of continuous flow. Failure and repair times were exponentially distributed. The first server was assumed to operate under blocking mechanism type 2. Wijngaard obtained the expected cost per cycle, which is the time between two regeneration points. A regeneration point is the entrance into the state: server 1 is down, server 2 is operative and the intermediate waiting room is empty (see also De Koster and Wijngaard [15]).

Two-node queueing networks similar to the one analyzed in section 5.2 have also been studied by Okamura and Yamashina [42], Muth and Yeralan [38], Gershwin and Schick [17] and Jafari and Shanthikumar [25]. In particular, Muth and Yeralan [38] analyzed the same system studied in section 5.2 assuming that a server cannot break down when it is idle or blocked, and that a broken down server cannot accommodate a unit in front of it (These two assumptions are different to assumptions (c) and (e) mentioned above in section 5.2). The states of the system were enumerated in such a way so

that the non-zero elements of the one-step transition probability matrix P have the following structure

$$P = \begin{bmatrix} A' & B' & & & & & \\ & C & D & & & & \\ & & C & D & & & \\ & & & \cdot & & & \\ & & & & \cdot & & \\ & & & & & C & D \\ & & & & & E & F \end{bmatrix}$$

Utilizing the above structure, they gave a numerical procedure for deriving the stationary probability vector. We note, that the non-zero elements of this matrix P have a block tri-diagonal structure, if the states of the system are enumerated as described in connection with the model studied in section 5.2. Gershwin and Schick [17] analyzed the system studied by Muth and Yeralan [38], assuming that the first server operates under blocking mechanism type 2, and, that when a server fails the unit under service has to repeat its service. They obtained the solution numerically using a method described in Gershwin and Berman [18] and outlined in section 5. Finally, Jafari and Shanthikumar [25] analyzed numerically the case where the operational and repair times have a phase type distribution. When a breakdown occurs, the customer may be scrapped with a fixed probability. They observed that the rate matrix has a block tri-diagonal structure. The stationary probability vector was obtained recursively using an algorithm reported in Jafari and Shanthikumar [26].

6. TWO-NODE CLOSED QUEUEING NETWORKS WITH BLOCKING

In this last section, we examine a two-node closed exponential queueing network with blocking as shown in figure 3. Let N be the number of units in the network.

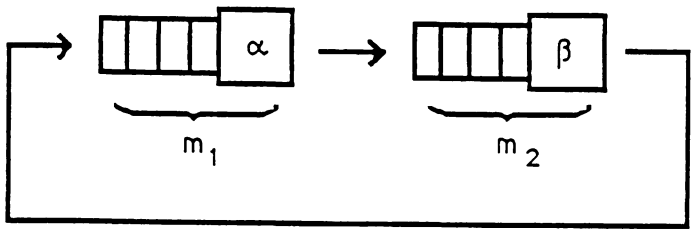


Figure 3: A two-node closed queueing network with blocking

Let m_1 and m_2 be the capacity of the first and second node respectively. We have, $N \leq m_1 + m_2$. In the following section, we examine the case where the service times are exponentially distributed. Results pertaining to the case of general service times are given in section 6.2.

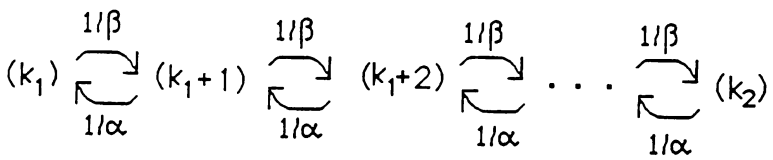
6.1 Exponential Service Times

We examine this queueing network under type 2 and 1 blocking mechanisms. Service times at the first and second node are exponentially distributed with mean α and β respectively. For presentation purposes, the analysis is restricted to the case where each queue is served by a single server. The results obtained here can be easily extended to the case where each queue is served by multiple servers.

a) Blocking Mechanism Type 2.2

We consider the queueing network described above assuming that a server cannot start a new service if the other queue is full (type 2.2 blocking). We note that for $N = m_1 + m_2$ we have a deadlock. This is because both nodes are full and therefore both

servers are blocked. This deadlock can not be resolved without violating the rules of this blocking mechanism. Therefore, we examine this model when $N < m_1 + m_2$. This model was first studied by Gordon and Newell [19]. Following their argument, let $p(n_1,n_2)$ be the steady-state probability that there are n_1 and n_2 units in the first and second node respectively. Seeing that $n_1 + n_2 = N$, we have $p(n_1,n_2) = p_1(n_1)$, where $p_1(n_1)$ is the marginal probability that there are n_1 units in the first node. Let $m_1, m_2 < N$. Define $k_1 = \max(0, N - m_2)$ and $k_2 = \min(N, m_1)$. Then, $k_1 \le n_1 \le k_2$. The rate diagram associated with node 1 has the following simple form



Thus, we can easily obtain that

$$p_1(n_1) = \left(\frac{\alpha}{\beta}\right)^{n_1 - k_1} p_1(k_1) \quad , \quad k_1 \le n \le k_2 \quad , \tag{6.1}$$

and

$$p_1(k_1) = \begin{cases} \dfrac{1 - (\alpha/\beta)}{1 - (\alpha/\beta)^{k_2 - k_1 + 1}} & , \alpha \neq \beta \\ \\ 1/(k_2 - k_1 + 1) & , \alpha = \beta \quad . \end{cases} \tag{6.2}$$

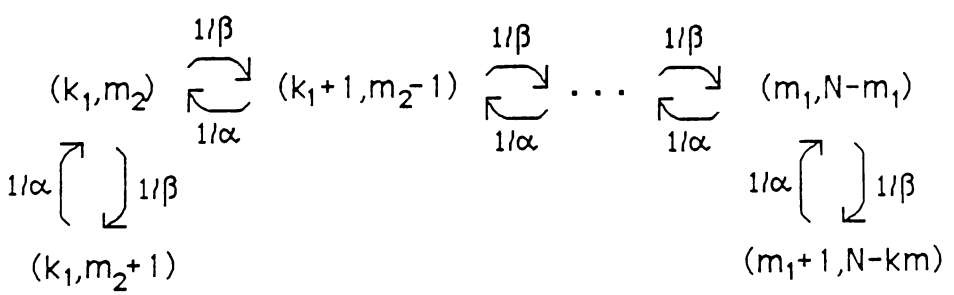
If $m_1, m_2 \ge N$, then $k_1 = 0$ and $k_2 = N$. In this case, the queue-length distribution of the first queue becomes identical to that of an M/M/1/N queue with an arrival and ser-

vice rate equal to $1/\beta$ and $1/\alpha$ respectively. If $m_1 \leq N \leq m_2$, then $k_1=0$ and $k_2=m_1$. In this case, the queue-length distribution of the first queue becomes identical to that of an $M/M/1/m_1$ queue with an arrival and service rate equal to $1/\beta$ and $1/\alpha$ respectively.

Finally, we note that the steady-state solutions (6.1) and (6.2) also hold if the above queueing network is considered under type 3 blocking mechanism.

b) Blocking Mechanism Type 1

We now analyze the queueing network shown in figure 3 under blocking mechanism type 1. In this case, $N \leq m_1 - m_2$. Let k_1 and k_2 be the minimum and maximum occupancy of node 1 as defined above. If $m_1, m_2 < N$, we have that $k_1 > 0$, $N - k_1 = m_2$, and $k_2 = m_1$. Let (n_1, n_2) indicate the state of the queueing network, where n_1 and n_2 is the number of units in the first and second node respectively. Then, the rate diagram associated with this system is:



The states $(k_1, m_2 - 1)$ and $(m_1 - 1, N - m_1)$ are associated with the first server respectively, the second server being blocked. We can easily obtain that

$$p(k_1 + i, m_2 - i) = \left(\frac{\alpha}{\beta}\right)^{i-1} p(k_1, m_2 + 1) \;,\; 0 \leq i \leq m_1 - k_1 + 1 \qquad (6.3)$$

and

$$p(k_1, m_2 - 1) = \begin{cases} \dfrac{1 - (\alpha/\beta)}{1 - (\alpha/\beta)^{m_1 - k_1 + 3}} & ,\alpha \neq \beta \\ \\ 1/(m_1 - k_1 + 3) & ,\alpha = \beta \end{cases} \tag{6.4}$$

If $m_1 \leq N \leq m_2$, then the first server can never get blocked seeing that the capacity of the second node is large enough to accommodate all N units. The rate diagram of the system can be obtained from the one above by ignoring the blocking state (K_1, m_2+1) and setting $K_1=0$ and $K_2 = m_1$. The solution has a similar form as (6.3) and (6.4). We note that the first node becomes identical to an $M/M/1/m_1+1$ queue with arrival and service rate equal to $1/\beta$ and $1/\alpha$ respectively. We contrast this to a similar result obtained in section 6.1, where for $m_1 \leq N \leq m_2$ the first node becomes identical to an $M/M/1/m_1$ queue. If $m_1, m_2 \geq N$, then neither server can get blocked. Therefore, the first queue becomes identical to an M/M/1/N queue with an arrival and service rate equal to $1/\beta$ and $1/\alpha$ respectively.

Akyildiz [1] studied this queueing network under type 1 blocking mechanism and assuming multiple servers. He demonstrated that the marginal queue-length distributions of the system under study are identical to those of a two-node closed queueing network without blocking. This queueing network without blocking is identical to the system under study but assuming infinite queues and a fixed number of customers equal to z-1, where $z = \min\{N, m_1 + s_2\} + \min\{N, m_2 + s_1\} - N + 1$.

6.2 General Service Times

Gun and Makowski [21] obtained a closed-form matrix-geometric solution for a two-node closed queueing network assuming phase type distributions. This model has

also been studied approximately by Kouvatsos and Almond [30] under type 2 blocking using the principle of maximum entropy. Each node i is served by multiple homogenous servers. General service times are assumed with state dependent service rates and state independent coefficient of variation. The authors showed that for a special service distribution (which they called generalized exponential) the results become exact.

Acknowledgement

I would like to thank Dr. Raif Onvural for his help in revising this paper. Without his assistance this paper would have still be sitting gathering dust!

REFERENCES

[1] Akyildiz, I.F., "Exact Product Form Solution for Queueing Networks with Blocking", IEEE Trans. Comp., 36 (1987), 122-125

[2] Altiok, T., "Production Lines with Phase Type Operations and Repair Times and Finite Buffers", Int. J. Prod. Res., 28 (1985), 489-498

[3] Altiok, T. and Stidham, S., A Note on Transfer Lines with Unreliable Machines, Random Processing Times and Finite Buffers", IIE Trans. (1982), 125-127

[4] Artamonov, G.T., "Productivity of a Two-Instrument Discrete Processing Line in the Presence of Failures", Kibernetica 3, (1976) 126-130 (Engl. tr. Cybernetics, 12 (1977), 464-468

[5] Asare, B., "Queue Networks with Blocking", Ph.D. Thesis (1978), Trinity College Dublin, Ireland

[6] Avi-Itzhak, B. and Yadin, M., "A Sequence of Two Servers with no Intermediate Queue", Mgmt Sci., 11 (1965), 565-571

[7] Bocharov, P.P. and Albores, F.K., "On Two-Stage Exponential Queueing System with Internal Losses or Blocking", Problems of Control and Information Theory, 9 (1980), 365-379

[8] Boxma, O. and Konheim, A., "Approximate Analysis of Exponential Queueing Systems with Blocking", Acta Informatica, 15 (1981), 19-66

[9] Buzacott, J.A., "The Effect of Station Breakdowns and Random Processing Times on the Capacity of Flow Lines with In-Process Storage", AIIE Trans., 4 (1972), 308-312

[10] Caseau, P. and Pujolle, G., "Throughput Capacity of a Sequence of Queues with Blocking Due to Finite Waiting Room", IEEE Trans. Soft Eng. SE-5 (1979), 631-642

[11] Clarke, A.B., "Markovian Queues with Servers in Tandem", Tech. Rep. 49, Math. Dept., Western Michigan Univ., (1977)

[12] Clarke, A.B., "A Two-Server Tandem Queueing System with Storage Between Servers", Tech. Rep. 50, Math. Dept., Western Michigan Univ., (1977)

[13] Clarke, A.B., "A Multiserver General Service Time Queue with Servers in Series", Tech. Rep. 51, Math. Dept., Western Michigan Univ., (1978)

[14] Clarke, A.B., "Waiting Times for Markovian Queue with Servers in Series", Tech. Rep. 52, Math. Dept., Western Michigan Univ., (1978)

[15] De Koster, M.B.M. and Wijngaard, J., "Approximations of Two-Stage Production Lines with Intermediate Buffer", Tech. Rep., Dept. of Ind. Eng., Eindhoven Univ., (1985)

[16] Foster, F.G. and Perros, H.G., "On the Blocking Process in Queue Networks", European J. Oper. Res., 5 (1980), 276-283

[17] Gershwin, S.B. and Shick, I.C., "Modeling and Analysis of Three-Stage Transfer Lines with Unreliable Machines and Finite Buffers", Oper. Res., 31 (1983), 354-380

[18] Gershwin, S.B. and Berman, O., "Analysis of Transfer Lines Consisting of Two Unreliable Machines with Random Processing Times and Finite Storage Buffers", AIIE Trans., 13 (1981), 2-11

[19] Gordon, W.J. and Newell, G.F., "Cyclic Queueing Systems with Restricted Length Queues", Oper. Res., 15 (1967), 266-278

[20] Gun, L., "Tandem Queueing Systems Subject to Blocking with Phase-Type Servers-Analytical Solutions and Approximations", SRC Tech. Rep., 87-002, Univ. of Maryland, (1987)

[21] Gun, L and Makowski, A.M., "Matrix-Geometric Solution for Finite Queues with Phase Type Distribution", Performance'87, (Eds: Courtois and Latouche), North Holland (1988), 269-282

[22] Hatcher, J.M., "The Effect of Internal Storage on the Production Rate of a Series of Stages Having Exponential Service Times", AIIE Trans., 1 (1969), 150-156 (See also: Letter to the above Article by A.D. Knott, AIIE Trans., (1970), 273)

[23] Hunt, G.C., "Sequential Arrays of Waiting Lines", Oper. Res., 4 (1956), 674-683

[24] Ignall, E. and Silver, A., "The Output of a Two-Stage System with Unreliable Machines and Limited Storage", AIIE Trans., 9 (1977), 183-188

[25] Jafari, M.A. and Shanthikumar, J.G., "A Model of Two-Stage Flow Lines with Non-Geometric Uptimes and Downtimes and Possible Scrapping of Workpieces", Tech. Rep. 85-008, Ind. Eng. and Oper. Res. Dept., Syracuse University, (1985)

[26] Jafari, M.A. and Shanthikumar, J.G., "Finite State Spatially Non-Homogeneous Quasi-Birth-And-Death-Processes", Tech. Rep., 85-009, Ind. Eng. and Oper. Res. Dept., Syracuse University, (1985)

[27] Koenigsberg, E., "Production Lines and Internal Storage- A Review", Mgmt. Sce., 5 (1959), 410-433

[28] Konheim, A.G. and Reiser, M., "A Queueing Model with Finite Waiting Room and Blocking", J. ACM, 23 (1976), 328-341

[29] Konheim, A.G. and Reiser, M., "Finite Capacity Queueing Systems with Applications in Computer Modeling", SIAM J. Computing, 7 (1978), 210-229

[30] Kouvatsos, D. and Almond, J., "Maximum Entrophy Two-Station Cyclic Queues with Blocking", Tech. Rep., Comp. Sci. Dept., Univ. of Bradford, (1987)

[31] Krishna, C. and Shin, K., "Queueing Analysis of A Canonical Model of Real-Time Multiprocessors", Proc. ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems, Bruell and Dowdy (Eds), (1983), 175-189

[32] Labetoulle, J. and Pujolle, G., "Modeling of Packet Switching Communication Networks with Finite Buffer Size at Each Node", Computer Performance, Chandy and Reiser (Eds), North Holland (1977), 515-535

[33] Langaris, C. and Conally, B., "Two queues in Tandem with Multiple Facilities and Restricted Waiting Space", METRON, Vol. XLIII (1985)

[34] Langaris, C. and Conally, B., "On the Waiting Time of a Two-Stage Queueing System with Blocking", J. Appl. Prob., 21 (1984), 628-638

[35] Latouche, G. and Neuts, M.F., "Efficient Algorithmic Solutions to Exponential Tandem Queues with Blocking", SIAM J. of Alg. Disc. Meth., 1 (1980), 93-106

[36] Lavenberg, S.S., "Stability and Maximum Departure Rate of Certain Open Queueing Networks Having Finite Capacity Constraints", RAIRO Informatique/ Computer Science, 12 (1978), 353-370

[37] Makino, T., "On the Mean Passage Time Concerning Some Queueing Problems of the Tandem Type", J. Oper. Res. Soc. Japan, 7 (1964), 17-47

[38] Muth, E.J. and Yeralan, S., "Effect of Buffer Size on Productivity of Work Stations that are Subject to Breakdowns", Proc. 20th IEEE Conf. on Decision and Control, (1981), 643-648

[39] Neuts, M.F., "Two-Queues in Series with a Finite, Intermediate Waiting Room", J. Appl. Prob., 5 (1968), 123-142

[40] Neuts, M.F., "Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach", The John Hopkins Univ. Press, (1981), 232-245

[41] Newell, G.F., "Approximate Behavior of Tandem Queues", Lecture Notes in Economics and Mathematical Systems, 171, Springer-Verlag, (1979)

[42] Okamura, K. and Yamashina, H., "Analysis of the Effect of Buffer Storage Capacity in Transfer Line Systems", AIIE Trans, 9 (1977), 127-135

[43] Onvural, R.O., "Closed Queueing Networks with Finite Buffers", Ph.D. Thesis, (1987), North Carolina State University

[44] Onvural, R.O., "A Survey of Closed Queueing Networks with Finite Buffers", Tech. Rep., Comp. Sci. Dept., NC State University, (1987)

[45] Onvural, R.O. and Perros, H.G., "On Equivalencies of Blocking Mechanisms in Queueing Networks with Blocking", Oper. Res. Letters, 5 (1986), 293-297

[46] Onvural, R.O., Perros, H.G. and Altiok, T., "On the Complexity of the Matrix Geometric Solution of Exponential Open Queueing Networks with Blocking", Proc. Int. Workshop on Modeling Techniques and Performance Evaluation, Paris, March 9-10, 1987, 3-12

[47] Pellaumail, J. and Boyer, P., "Deux Files D'Attente A Capacite Limitee en Tandem", Tech. Rep. 147, CNRS/INRIA-Rennes, France, (1981)

[48] Perros, H.G. and Snyder, P.M., "A Computationally Efficient Approximation Algorithm for Analyzing Queueing Networks with Blocking", Tech. Rep., Comp. Sci. Dept., NC State University, (1986)

[49] Perros, H.G., "Queueing Networks with Blocking: A Bibliography", Performance Evaluation Review, 12 (1984), 8-12

[50] Pinedo, M. and Wolff, R.W., "A Comparison Between Tandem Queues with Dependent and Independent Service Times", Oper. Res., 30 (1982), 464-479

[51] Prabhu, N.U., "Transient Behavior of a Tandem Queue", Mgmt. Sci., 13 (1967), 631-639

[52] Pujolle, G. and Potier, D., "Reseaux de Files D'Attente A Capacite Limitee Avec des Applications aux Systemes Informatiques", RAIRO Informatique/Computer Sci., 13 (1979), 175-197

[53] Rao, N.P., "On the Mean Production Rate of a Two-Stage Production System of the Tandem Type", Int. J. Prod. Res., 13 (1975), 207-217

[54] Rao, N.P., "Two-Stage Production Systems with Intermediate Storage", AIIE Trans., 7 (1975), 414-421

[55] Suzuki, T., "On a Tandem Queue with Blocking", J. Oper. Res. Soc. Japan, 6 (1964), 137-157

[56] Takahashi, T., Miyahara, H. and Hasegawa, T., "An Approximation Method for Open Restricted Queueing Networks", Oper. Res., 28 (1980), 594-602

[57] Tsiotras, G.D., "Exact and Approximate Analysis of Queueing Network Models with Blocking", Ph.D. Thesis, (1987), State University of New York at Stony Brook

[58] Van Dijk, N. and Lamond, B.F., "Bounds for the Call Congestion in Tandem Queues", Tech. Rep. 1078, Faculty of Commerce and Business Administration, British Columbia University, (1984)

[59] Wijngaard, J., "The Effect of Interstage Buffer Storage on the Output of Two Unreliable Production Units in Series, with Different Production Rates", AIIE Trans., 11 (1979), 42-47

[60] Wong, B., Giffin, W. and Disney, R., "Two-Finite M/M/1 Queues in Tandem: A Matrix Solution for the Steady State", Opsearch, 14 (1977), 1-18