A Survey of Queueing Networks with Blocking
Part I

By

H.G. Perros

Center for Communications and Signal Processing
Department of Computer Science
North Carolina State University

January 1986

CCSP-TR-86/4

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, it is attempted to give a systematic presentation of the literature related to two-node queueing networks with blocking. Only papers which contain analytical investigations or numerical investigations of queueing networks with blocking were considered. Results related to queueing networks with blocking consisting of more than two nodes will be surveyed in a companion paper.

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 arises because of the limitations imposed on the size of these queues. In particular, 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. Some exact closed-form results have been reported in the literature. However, closed-form solutions are not generally available. Most of the techniques, therefore, that are employed to analyze such queueing networks are in the form of approximations, numerical techniques, and simulation techniques.

Various configurations of queueing networks with blocking have been considered in the literature so far. The simplest configuration is the queueing network with blocking consisting of two queues in tandem with a finite intermediate queue. This model has been studied under a multiplicity of different assumptions regarding service times, feedback and blocking mechanisms in about 43% of the papers listed in the reference section. Other specific configurations that have been considered in the literature are the tandem, split, and merge configurations. A tandem configuration consists of $m(m>1)$ queues in series, a split configuration consists of a queue linked to $m(m>1)$ parallel queues, and a merge configuration consists of $m(m>1)$ parallel queues linked to a single queue. Of these three configurations, the tandem configuration is the one that has been studied most. In particular, it has been analyzed under different assumptions in about 42% of the references given in this paper. Finally, arbitrary configurations of queueing networks with blocking have also been considered. These configurations are, in general, more difficult to analyze than tandem, split and merge configurations.

This paper gives a survey of exact, approximate and numerical results related to a two-node queueing network with blocking. Results related to queueing network with blocking consisting of more than two nodes will be surveyed in a companion paper. Sections 3 to 7 deal with open two-node queueing networks with blocking. In particular, in section 3, we survey results related to two queues in tandem, where the second queue has a finite capacity, assuming exponentially distributed service times. Section 4, contains results for the same queueing network assuming arbitrary service times. In section 5, we consider the two-queue network assuming that each queue is served by multiple servers. Sections 6 and 7, deal with the queueing network analyzed in section 3, assuming unreliable servers. Finally closed two-node queueing networks with blocking are dealt with in section 8.

We note that the references given at the end of this paper, is a bibliography of all papers which contain exact, approximate and numerical results related to queueing networks with blocking. It is an updated version of a bibliography compiled by Perros [75]. Not all papers given in this bibliography are referenced in this paper.

2. BLOCKING MECHANISMS

Before we proceed with the survey, we review briefly various blocking mechanisms that have been considered in the literature so far. These blocking mechanisms arose out of various studies of real life systems. They are distinct types of models for blocking, a fact that may be easily misissed by a reader unfamiliar with the subject. Onvural and Perros [69] have classified the most commonly used blocking methods as follows:

TYPE 1: A customer upon completion of its service at queue i attempts to enter destination queue j. If queue j at that moment is full, the customer is forced to wait in front of server i until it enters destination queue j. The server remains blocked for this period of time and it can not serve any other customers waiting in its queue. This blocking mechanism has been used to model systems such as production systems and disk I/O subsystems (cf. Altiok [3], Perros [72]).

TYPE 2: A customer in queue i declares its destination queue j before it starts its service. If queue j is full, the ith server becomes blocked, i.e. it can not serve customers. When a departure occurs from destination queue j, the ith server becomes unblocked and the customer begins receiving service. This blocking mechanism has been used to model systems such as production systems and telecommunication systems. (cf. Boxma and Konheim [17], Gershwin and Berman [33].)

Depending on whether the blocking 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.

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 c_1 and c_2. In Type 2.1, when there are c_2 customers in the second queue, there can be at the most c_1-1 customers in queue 1, but, in Type 2.2, there can be c_1 customers in queue 1.

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

Within this category of blocking mechanisms, we distinguish the following two sub-categories.

TYPE 3.1: Once the customer's destination is determined it cannot be altered. This blocking mechanism arose in modelling telecommunication systems (cf. Caseau and Pujolle [89]).

TYPE 3.2: A destination node is chosen at each service completion independently of the destination node chosen the previous time. This type of blocking is associated with reversible queueing networks with blocking (cf. Yao and Buzacott [90]).

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 maximum capacity. Latouche and Neuts [54[considered other extensions whereby unblocking of a server occurs when the number of units in the destination queue drops below a predefined level, not necessarily equal to its maximum capacity (see also Lavenberg [55]). Also, there are several blocking mechanisms which are also

associated with product-form solutions (c.f. Hordijk and Van Dijk [41], Yao and Buzacott [90], and Le Ny [57], [58]). These blocking mechanisms are not considered here.

Comparisons between these distinct types of blocking mechanisms have been carried out to a limited extent by Caseau and Pujolle [22], Altiok and Stidhan [6] and Bocharov and Albores [15]. More recently, Onvural and Perros [69] made extensive comparisons between the above mentioned blocking mechanisms for tandem, split and merge configurations. The objective of these comparisons is to obtain an equivalency between a queueing network (call it A) with type i blocking mechanism and another queueing network (call it B) with type j blocking mechanism. Queueing network B may be identical to A, or it may be obtained from A by simply changing the queue capacities of some or all of the nodes of A. The two queueing networks are said to be equivalent if they have identical marginal queue-length probabilities.

The equivalencies identified for two-node queueing networks with blocking are as follows.

a) When the first queue is infinite, we have 1=2.1=2.2=3.1=3.2.

The notation i=j means that the queueing network under type i blocking mechanism is equivalent to the same queueing network under type j blocking mechanism. All equivalencies are valid without any adjustments to the queue sizes except the equivalencies 1=2.1, 1=2.2, 1=3.1 and 1=3.2. In this case, a two-node queueing network with type 1 blocking mechanism is equivalent to an identical configuration with type 2.1 (or 2.2, 3.1, 3.2) blocking mechanism, if the maximum capacity of its second queue is increased by one.

b) When the first queue is finite, we have $1=2.1$ and $2.2=3.1=3.2$.

We note that equivalency $1=2.1$ is valid when the maximum capacity of the second queue is increased by one, as mentioned in (a) above.

We now proceed with the survey by first examining the open two-node queueing network under exponentially distributed service times.

3. EXPONENTIAL SERVICE TIMES

Let us consider two queues in tandem as shown in figure 1. The first queue is infinite, and the second queue has a finite capacity of size m (including the one in service). Each queue is served by a single exponential server. External arrivals join the first queue

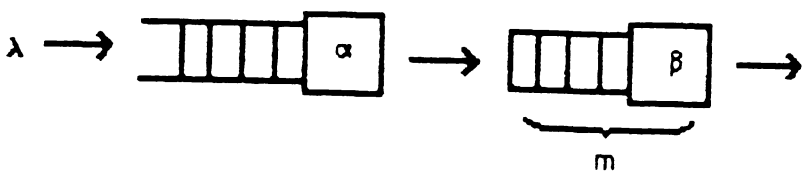


Figure 1: A two-node open queueing network with blocking

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.

A unit arrives at the first queue, gets served by the first server in a FIFO manner, and then it proceeds to the second queue. A unit in the second queue gets served in a FIFO manner and then it departs from the queueing network. Due to finite capacity of the second queue, the first server occasionally becomes blocked.

3.1 Blocking Mechanism Type 1

Let us assume that the above queueing network operates under the type 1 blocking mechanism. That is, server 1 can start a new service even when the second queue is full (i.e., it contains m units). The first server will become blocked if it completes its service before a departure occurs from the second queue. During the time the server is blocked it cannot serve any units that might be waiting in its queue. The server becomes unblocked the moment a service is completed at the second server.

The state of the above queueing network can be described by the variables (n_1,n_2), where n_1 indicates the number of units in the first queue and n_2 the number of units in the second queue. We have, n_1=0,1,2, $\cdots$ and n_2=0,1, $\cdots$ m,m+1. The value n_2=m+1 indicates that the second queue 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 queue. Of these i units, one has received its service but it is blocked from entering queue 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.

$$\begin{cases} \lambda p_{00} = (1/\beta)p_{01} & (3.1) \\ (\lambda+(1/\alpha))p_{i0} = (1/\beta)p_{i1} + \lambda p_{i-1,0} \quad , i \geq 1 & (3.2) \end{cases}$$

$$\begin{cases} (\lambda + (1/\beta))p_{01} = (1/\alpha)p_{10} + (1/\beta)p_{02} & (3.3) \\ (\lambda + (1/\alpha) + (1/\beta))p_{i1} = (1/\alpha)p_{i+1,0} + (1/\beta)p_{i2} + \lambda p_{i-1,1} \quad ,i \geq 1 & (3.4) \end{cases}$$

$$\begin{cases} (\lambda + (1/\beta))p_{02} = (1/\alpha)p_{11} + (1/\beta)p_{13} & (3.5) \\ \lambda + (1/\alpha) + (1/\beta))p_{i2} = (1/\alpha)p_{i+1,1} + (1/\beta)p_{i+1,3} + \lambda p_{i-1,2} \quad ,i \geq 1 & (3.6) \end{cases}$$

$$\begin{cases} \lambda + (1/\beta))p_{13} = (1/\alpha)p_{12} \end{cases} \tag{3.7}$$

$$\begin{cases} (\lambda + (1/\beta))p_{i3} = (1/\alpha)p_{i2} + \lambda p_{i-1,3} \quad ,i \geq 2 \end{cases} \tag{3.8}$$

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, equations (3.1) to (3.8) can be written as follows

$$\begin{cases} g_0(z)[\lambda(1-z)+(1/\alpha)] = (1/\beta)g_1(z)+(1/\alpha)p_{00} \\ g_1(z)[\lambda(1-z)+(1/\alpha))+(1/\beta)] = (1/\alpha z)g_0(z)+(1/\beta)g_2(z)+(1/\alpha)p_{01}-(1/\alpha z)p_{00} \\ g_2(z)[\lambda(1-z)+(1/\alpha)+(1/\beta)] = (1/\alpha z)g_1(z)+(1/\beta)g_3(z)+(1/\alpha)p_{02}-(1/\alpha z)p_{01} \\ g_3(z)[\lambda(1-z)+(1/\beta)] = (1/\alpha z)g_2(z)-(1/\alpha z)p_{02} \end{cases} \tag{3.9}$$

The normalizing equation becomes $\sum_{k=1}^{4} g_k(1) = 1$.

The total number of independent equations is 6, namely the four equations given by (3.9), the normalizing condition, and an additional equation ($\lambda p_{00} = (1/\beta)p_{01}$) that can be obtained from the first equation of (3.9) by setting z=0. 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

closed-form solutions for the generating functions $g_k(z)$,k=0,1,...,m+1.

3.2 Blocking Mechanism Type 2

Let us now consider the queueing network shown in figure 1 with type 2 blocking mechanism. Due to the fact that the first queue is infinite, we do not distinguish between types 2.1 and 2.2. Under blocking mechanism type 2, a unit in the first queue cannot start its service if the second queue is full.

Let $p_{ij,}$ i=0,1,..., and j=0,1,...m, be the probability that there are i units in the first queue and j units in the second queue. States (i,m), $i \geq 0$, are associated with the first server being blocked. For presentation purposes, let us consider the queueing network when m=3. We note that if we associate $p_{i3,} i \geq 1$, of the queueing network under blocking mechanism type 1 with probability $p_{i-1,3}$ of the queueing network under blocking mechanism type 2, the steady-state equations of the type 2 model are identical to those of type 1 given by equations (3.1) to (3.8). That is, the queueing network with type 1 blocking mechanism is equivalent to an identical configuration with type 2 blocking mechanism if the capacity of its second queue is increased by one.

In view of this equivalency, the comment regarding the lack of equations made in relation to the queueing network under blocking mechanism type 1 also applies here. Konheim and Reiser [48] (see also [49]) studied the queueing network shown in figure 1 under blocking mechanism type 2 assuming that a customer departing from the second server may be feedback to the first queue. The authors obtained the additional equations necessary for the solution of this network by requiring that the generating function $g_m(z)$ be analytic within the unit circle $g_m(z)$ can be written in the form $g_m(z) = f(z)/h(z)$. 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. The authors developed a lemma regarding the behaviour of these roots, and described an algorithm for obtaining p_{ij}.

3.3 Blocking Mechanism Type 3

Consider the queueing network shown in figure 1 with type 3 blocking mechanism. Due to the fact that there is only one destination queue (queue 2) we do not distinguish between type 3.1 and type 3.2. Let p_{ij} be the steady-state probability that there are i units in the first queue and j units in the second queue, where i=0, 1,..., and j=0,1,...,m. We note that the steady-state equations of the system are identical to those of the queueing network under type 2 blocking mechanism. This system was first studied by Pujolle and Potier [80]. The authors recognized the above equivalencies between blocking mechanisms 1.2 and 3 for this particular queueing network. They obtained an approximate closed-form solution for p_{ij} by decomposing the queueing network to two individual queues and studying each of these queues separately.

Pellaumail and Boyer [71] 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.

3.4 Limiting Cases: Exact Results

Exact closed-form results have been obtained in some limiting cases. Due to the above mentioned equivalencies between blocking mechanisms 1,2 and 3, we will only consider the queueing network under type 1 blocking mechanism.

a) $\alpha \to 0$

In this case, a unit arriving at the first queue goes through it and joins the second queue. If the second queue already contains m units, the arriving (m+1st) unit will block the server of the first queue. Now, if a unit departs from the second queue the first server will become unblocked for an infinitely small time and then it will get blocked again, if

thee are units in the first queue. 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 there are i units in the second queue of the queueing network is simply p_i. The probability there are i units in the first queue of the queueing network is p_{m+i}.

This intuitive argument was given by Foster and Perros [29]. Reiser and Konheim [48] arrived at the same result using a rigorous mathematical approach.

b) Saturated first queue

Let α_0 be the mean service time above which the first queue is always saturated. Then, for $\alpha \geq \alpha_0$ the first server is either busy serving or blocked. In particular, a unit upon completion of its service at the first server gets blocked if at that moment the second queue is full. The blocking unit remains in front of the first sever until a departure occurs from the second queue. At that instance it moves to the mth position of the second queue, and the first server becomes unblocked. In view of this, during the blocking period the first server acts as an additional space that belongs to the second queue. Furthermore, due to the blocking period the first server does not serve any other units. That is, no more arrivals occur at the second queue. Due to the exponential assumption, this is equivalent to saying that arrivals do occur at the second queue at the rate $1/\alpha$ but they are lost. In view of this, the second queue 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 [29]. This result can be easily verified by studying the underlying Markov process (see Hatcher [36]). 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 examined below.

c) The blocking probability

This is the probability that a unit upon service completion at the first server will be blocked. That is, at that instant the second queue is full. This probability (call it π) has been used in several approximation algorithms developed for the analysis of queueing networks with blocking (cf. Altiok and Perros [5], Labetoulle and Pujolle [52], and Takahashi, Miyahara, and Hasegawa [89]). Below, we obtain the exact expression for this probability when $\alpha \to 0$ (call it π_1) and when $\alpha \geq \alpha$ (call it π_2) following arguments given in Foster and Perros [29]

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

$$\pi_1 = \sum_{i \geq m} p_i = (1-\rho) \sum_{i \geq m} \rho^i = \rho^m \quad , \tag{3.10}$$

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

$$\lambda' = (1/\alpha)(1-p_{m+1}) \tag{3.11}$$

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

$$\pi_2 = \frac{(1-\sigma)\sigma^m}{1-\sigma^{m+1}} \qquad (3.12)$$

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

The quantity α_0 is derived below 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 [40]. For the special case $\alpha = \beta$ we have that $1-\pi_2 = m/(m+1)$.

d) Processor-sharing discipline

We now analyze the queueing network shown in figure 4 under the discipline of processor-sharing. 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 [10], Konheim and Reiser [48] and also Perros [72]).

Let us consider the queueing network shown in figure 1, and let us assume that a unit upon completion of its second service 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. Below, we obtain this limiting case for $m=2$.

The steady-state equations are as follows:

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

$$\begin{cases} (\lambda + \dfrac{1}{\beta})\, p_{01} = \dfrac{1}{\alpha}\, p_{10} + \dfrac{1-\theta}{\beta}\, p_{02} \\ \\ (\lambda + \dfrac{1}{\alpha} + \dfrac{1}{\beta})\, p_{i1} = \dfrac{1}{\alpha}\, p_{i+1,0} + \dfrac{1-\theta}{\beta}\, p_{i2} + \dfrac{\theta}{\beta}\, p_{i-1,2} + \lambda p_{i-1,1} \quad , \quad i \geq 1 \end{cases}$$

$$\begin{cases} (\lambda + \dfrac{1}{\beta})\, p_{02} = \dfrac{1}{\alpha}\, p_{11} + \dfrac{1-\theta}{\beta}\, p_{13} \\ \\ (\lambda + \dfrac{1}{\alpha} + \dfrac{1}{\beta})\, p_{i2} = \dfrac{1}{\alpha}\, p_{i+1,1} + \dfrac{1-\theta}{\beta}\, p_{i+1,3} + \dfrac{\theta}{\beta}\, p_{i3} + \lambda p_{i-1,2} \quad , \ i \geq 1 \end{cases}$$

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

Let $\rho_1 = \lambda\alpha/(1-\theta)\rho_2 = \lambda\beta/(1-\theta)$. Then, the above equations can be put into the following form:

$$\begin{cases} \rho_2 p_{00} = p_{01} \end{cases} \tag{3.13}$$

$$\begin{cases} (\rho_1\rho_2(1-\theta) + \rho_2)\, p_{i0} = \rho_1\rho_2(1-\theta)\, p_{i-1,0} + \rho_1(1-\theta)\, p_{i1}\rho_1\theta p_{i-1,1} \ , i \geq 1,1 \end{cases} \tag{3.14}$$

$$\begin{cases} (\rho_1\rho_2(1-\theta) + \rho_1)\, p_{01} = \rho_2 p_{10} + \rho_1(1-\theta)\, p_{02} & (3.15) \\ (\rho_1\rho_2(1-\theta) + \rho_1 + \rho_2)\, p_{i1} = \rho_2 p_{i+1,0} + \rho_1(1-\theta) p_{i2} + \rho_1\theta p_{i-1,2} + \rho_1\rho_2(1-\theta) p_{i-1,1} \;, \; i \geq 1 & (3.16) \end{cases}$$

$$\begin{cases} (\rho_1\rho_2(1-\theta) + \rho_1)\, p_{02} = \rho_2 p_{11} + \rho_1(1-\theta)\, p_{13} & (3.17) \\ (\rho_1\rho_2(1-\theta) + \rho_1 + \rho_2)\, p_{i2} = \rho_2 p_{i+1,1} + \rho_1(1-\theta) p_{i+1,3} + \rho_1\theta p_{i3} + \rho_1\rho_2(1-\theta)\, p_{i-1,2} \;, \; i \geq 1 & (3.18) \end{cases}$$

$$\begin{cases} (\rho_1\rho_2(1-\theta) + \rho_1)\, p_{13} = \rho_2 p_{12} & (3.19) \\ (\rho_1\rho_2(1-\theta) + \rho_1)\, p_{i3} = \rho_2 p_{i2} + \rho_1\rho_2(1-\theta)\, p_{i-1,3} \quad, i \geq 2 & (3.20) \end{cases}$$

We first observe that we can obtain equations which are independent of θ. In particular, by adding equations (3.13), (3.14) with i=1, and (3.14) we obtain:

$$\rho_2(p_{10} + p_{01}) = p_{11} + p_{02} \tag{3.21}$$

Likewise, adding equations (3.14) with i=2, (3.16) with i=1, and (3.17) gives

$$\rho_2(p_{20} + p_{11} + p_{02}) = p_{21} + p_{12} + p_{13} \quad . \tag{3.22}$$

Following this scheme we can obtain

$$\rho_2(p_{i0} + p_{i-1,1} + p_{i-2,2} + p_{i-2,3}) = p_{i1} + p_{i-1,2} + p_{i-1,3} \quad, i \geq 3 \tag{3.23}$$

We now substitute equations (3.14) by the above equations (3.21), (3.22), (3.23). Thus, our original queueing network is described by equations (3.21) to (3.23), (3.13) and

(3.15) to (3.20). 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 (3.15) to (3.20) are significantly simplified so that we can obtain the following.

$$\begin{cases} p_{i3} = \left(\dfrac{\rho_2}{\rho_1}\right)^3 p_{i+2,0} & ,i \geq 1 \\ p_{i2} = \left(\dfrac{\rho_2}{\rho_1}\right)^2 p_{i+2,0} & ,i \geq 0 \\ p_{i1} = \left(\dfrac{\rho_2}{\rho_1}\right) p_{i+1,0} & ,i \geq 0 \end{cases} \tag{3.24}$$

Using (3.24), (3.13) and (3.21) to (3.23) we obtain

$$p_{i0} = \rho_1^i \, p_{00} \quad . \tag{3.25}$$

The normalizing equation gives

$$p_{00} = \frac{(1-\rho_1)(1-\rho_2)}{1-\rho_2^4} \tag{3.26}$$

Thus,

$$p_{ij} = \begin{cases} \left[\rho_1^i(1-\rho_1)\right] \cdot \left[\rho_2^j \dfrac{1-\rho_2}{1-\rho_2^4}\right] & , \ j \leq 2 \ , \ i \geq 0 \\ \left[\rho_1^{i-1}(1-\rho_1)\right] \cdot \left[\rho_2^j \dfrac{1-\rho_2}{1-\rho_2^4}\right] & , \ j = 3, \ i \geq 1 \end{cases} \tag{3.27}$$

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

3.5 Condition for Stability

As it was mentioned above the first queue is stable if $\alpha<\alpha_0$, where α_0 is the critical mean service time of the first server above which the first queue saturates (i.e., it becomes unstable). Now if the first queue is stable, then the effective rate into the second queue is λ. However, if $\alpha\geq\alpha_0$ then the effective rate λ' is given by (3.11). Now, for $\alpha = \alpha_0$ we have $\lambda = \lambda'$ or

$$\lambda = \frac{1}{\alpha_0}\left[1 - \frac{(1-\sigma_0)\sigma_0^{m+1}}{1-\sigma_0^{m+2}}\right] \tag{3.27}$$

where $\sigma_0=\beta/\alpha_0$. The quantity α_0 can be computed numerically using (3.27).

The condition for stability can be simply expressed as $\alpha<\alpha_0$. Now, if $\alpha\geq\alpha_0$, then $\lambda'\leq\lambda$ or using (3.11) 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

$$\lambda\alpha < \frac{1+\sigma+\cdots+\sigma^m}{1+\sigma+\cdots+\sigma^{m+1}} \quad . \tag{3.28}$$

This intuitive argument was given by Foster and Perros [29]. The above condition for stability has been derived by several other authors (see Asare [10], Bocharov and Albores [15], Konheim and Reiser [48], Latouche and Neuts [54], Pujolle and Potier [80], Hunt [40] and Lavenberg [55]). We note that Lavenberg [55] formally demonstrated that when the first queue is saturated, the rate of departures from the queueing network equals the supremum of the set of arrivals for which the network is stable. Furthermore, this departure rate equals the supremum of the set of departure rates over all finite arrivals.

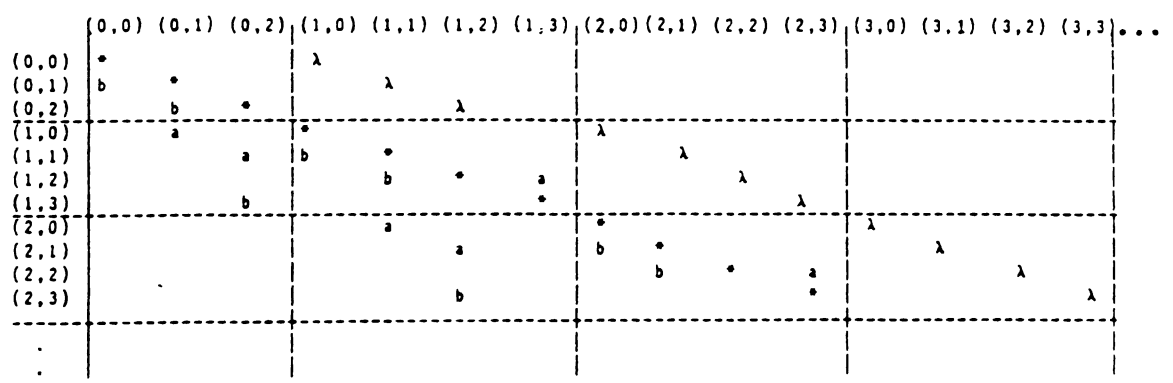
3.6 Numerical Methods: Matrix-Geometric Solution

Let us consider the queueing network shown in figure 1 with type 1 blocking

mechanism. Let (n_1,n_2) be the state of the system as described earlier on in connection with the type 1 blocking mechanism. If these states are enumerated lexicographically, the rate matrix Q will have the following block-triangular form.

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

for instance, for m=2 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, this Markovian process can be analyzed using the matrix-geometric procedure (see Neuts [64]). 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

$$(\underline{x}_0,\underline{x}_1,\underline{x}_2,\underline{x}_3,\ldots)\begin{bmatrix} A' & B' & & & & \\ C' & A & B & & & \\ & C & A & B & & \\ & & C & A & B & \\ & & & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot \end{bmatrix} = 0$$

or

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

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

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

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 involving two queues in tandem were investigated by Latouche and Neuts [54]. 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 Altiok and Perros [5] to analyze larger, configurations of queueing networks with blocking. Time and space complexity problems arise when analyzing networks with more than four queues arbitrarily linked.

Finally, we note that Wong, Giffin, and Disney [93] 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.

3.7 Other references

We conclude this section by giving some additional relevant references. Clarke [23],[24],[25],[26] 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 ($s \geq 2$)

servers in tandem with no intermediate queue [23], and for two servers in tandem with an arbitrary finite intermediate queue [24]. In [26] he obtained expressions related to the waiting time of a unit in the system consisting of two servers with no intermediate queue. Finally, in [25] he extended the results obtained in [23] to the case of general service times. The model analyzed in [24] was shown by Neuts [64] to have a matrix-geometric solution.

Finally, Boxma and Konheim [17] analyzed approximately the queueing network shown in figure 1, assuming that the first queue may be finite or infinite. Service times and inter-arrival times were assumed to be exponentially distributed. Type 2 blocking mechanism was assumed.

4. GENERAL SERVICE TIMES

Let us consider the queueing network, studied in section 3 and shown in figure 1, assuming that the service times at the first and second server follow arbitrary distributions. Neuts [63] studied this model assuming general service times at the first server, and exponential service times at the second server. The first server was assumed to operate 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 [88], Avi-Itzhak and Yadin [12] and Prabhu [79]. In particular, these authors analyzed the model studied by Neuts [63], assuming no intermediate waitingroom, i.e., m=1. The transient behavior of this model was analyzed in [79], whereas its steady-state behaviour was studied in [88] and [12]. Below we obtain some results pertaining to this model following Avi-Itzhak and Yadin [12].

4.1 No Intermediate Waiting Room

The queueing network under investigation is shown in figure 2. Units arrive at the first queue in a Poisson fashion until parameter λ. The service times at server 1 and 2,

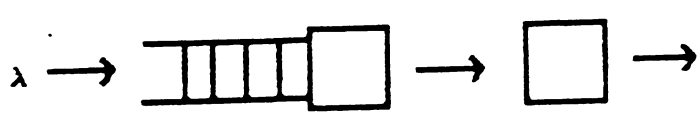


Figure 2. The case of m=1

s_1 and s_2 respectively, are independent and arbitrarily distributed with probability den-

sity 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. In view of this, the first queue can be seen as an M/G/1 queue in which the first customer of each busy period receives an exceptional service. This fact is used in the analysis of the queueing network.

We now proceed to obtain $f_0(\cdot)$, the pdf of the random variable t_0. As indicated above, this is the time elapsing from the arrival of the unit which starts the busy period until the unit enters the second server. We note that an idle period of the first server always begins with the first queue being empty and the second server just starting a new service. In view of this, we have

$$t_0 = s_1 + \max(s_2 - \tau_0 - s_1,0) \quad , \tag{4.2}$$

where τ_0 is the length of the idle period. This pdf $f_0(t)$ can be obtained by conditioning on s_2. For any given value of s_2 we have:

$$f_0(t \mid s_2) = f_1(t) \quad , \quad \text{if} \quad s_2 \leq t < +\infty \quad . \tag{4.3}$$

That is, if $t \geq s_2$ then the unit simply receives a service s_1 greater than s_2, and hence $t_0 = s_2$.

Let us now consider the case where $0 \leq t < s_2$. We have

$$f_0(t \mid s_2) = e^{-\lambda(s_2-t)} f_1(t) + F_1(t)\, \lambda e^{-\lambda(s_2-t)} \quad . \tag{4.4}$$

If $t<s_2$ then two different situations may arise: a) $\tau_0 + t_0 > s_2$ (or $\tau_0 > s_2 - t_0$) and b) $\tau_0 + t_0 = s_2$ (or $\tau_0 = s_2 - t_0$) with probability $\text{prob}[\tau_0 > s_2 - t_0] = e^{-\lambda(s_2-t)}$ respectively $\text{prob}[s_1 < t] = F_1(t)$. In the first case $f_0(t \mid s_2)$ equals $f_1(t)$ and in the second case it equals $\lambda e^{-\lambda\tau_0} = e^{-\lambda(s_2-t)}$ Now, combining (4.3) and (4.4) we have

$$f_0(t) = \int_0^t f_1(t) f_2(s) ds + \int_t^\infty e^{-\lambda(s-t)} (\lambda F_1(t) + f_1(t)) f_2(s) ds$$

or

$$f_0(t) = f_1(t) F_2(t) + e^{\lambda t} (\lambda F_1(t) + f_1(t)) \int_t^\infty e^{-\lambda s} f_2(s) ds \tag{4.5}$$

The first server undergoes a two-phased cycle consisting of an idle phase and a busy phase (busy serving or blocked). Let τ be the length of this cycle. We have

$$\tau = \tau_0 + \tau_b \quad ,$$

where τ_0 is the length of the idle period and τ_b is the length of the busy period. $E(\tau_0) = 1/\lambda$, seeing that the idle period is exponentially distributed with parameter λ. The quantity $E(\tau_b)$ can be obtained as follows. Let

$$\tau_b = t_0 + t_1 + t_2 + \ldots \tag{4.6}$$

where t_0 is defined as above, and t_i, $i \geq 1$, is the period elapsing from the termination of t_{i-1} until the customer who arrived last during t_{i-1} enters service at the second server. We have

$$E(t_i) = \lambda E(t_{i-1}) E(T) \quad , \quad i \geq 1 \, , \tag{4.7}$$

seeing that the mean number of customers who arrived during the period t_{i-1} is $\lambda E(t_{i-1})$ and each of them spent at the first server a mean time equal to E(T). Recursively, from (4.7) we obtain

$$E(t_i) = b^i E(t_0)$$

where $b = \lambda E(T)$. If $b \leq 1$ then $E(\tau_b)$ is finite and is equal to

$$E(\tau_b) = E(t_0)/(1-b) \quad . \tag{4.8}$$

Thus

$$\begin{aligned} E(\tau) &= E(\tau_0) + E(\tau_b) \\ &= \frac{1}{\lambda} + E(t_0)\,\frac{1}{1-b} \\ &= \frac{1-b+\lambda E(t_0)}{1-b} \end{aligned} \tag{4.9}$$

We now have defined $f_T(t)$ and $f_0(t)$ and therefore, we can analyze the first queue as an M/G/1 queue in which the first customer of each busy period receives an exceptional service using known results (see references in [12] and also in [79]) we can obtain the Laplace transform $g^*(z)$ of the time a unit spends in the queue and also in front of the first server. We have

$$g^*(z) = \frac{1-b}{1-b+\lambda E(t_0)} \cdot \frac{(z+\lambda)f_0^*(z) - \lambda f_\tau^*(z)}{z+\lambda-\lambda f_T^*(z)} \tag{4.10}$$

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

$$h^*(z) = g^*(z) f_2^*(z) \tag{4.11}$$

where $f_2^*(z)$ is the Laplace transform of the pdf $f_2(\cdot)$.

We now proceed to obtain the generating function of the probability distribution of the numbers of units in front of server 2. Let π_n be the steady-state probability that a unit leaves n units behind upon its entering server 2. We have

$$\pi_n = \int_0^\infty \pi_0\, e^{-\lambda t}\, \frac{(\lambda t)^n}{n!}\, f_0(t)dt + \int_0^\infty \left(\sum_{k=1}^{n+1} \pi_k\, e^{-\lambda t}\, \frac{(\lambda t)^{n-k+1}}{(n-k+1)!} \right) f_T(t)dt \tag{4.12}$$

The first term is related to the case where the unit was the one that started a new busy period. The second term is related to a unit arriving at the network during the busy period of the first server.

Let $g_n(z)$ be the generating function of π_n. We have,

$$g_n(z) = \sum_{n=0}^{\infty} \pi_n z^n$$

$$= \sum_{n=0}^{\infty} z^n \int_0^{\infty} \pi e^{-\lambda t} \frac{(\lambda t)^n}{n!} f_0(t)dt + \sum_{n=0}^{\infty} z^n \int_0^{\infty} \left(\sum_{k=1}^{n+1} \pi_k e^{-\lambda t} \frac{(\lambda t)^{n-k+1}}{(n-k+1)!} \right) f_T(t)dt$$

$$= \pi_0 \int_0^{\infty} e^{-\lambda(1-z)t} f_0(t)dt + \left(\sum_{k=1}^{\infty} \pi_k z^{k-1} \right) \int_0^{\infty} e^{-\lambda(1-z)t} f_T(t)dtL \quad .$$

After some calculations we can obtain

$$g_\pi(z) = \frac{\pi_0[z\beta_0\lambda(1-z) - \beta_T\lambda(1-z)]}{z - \beta_T\lambda(1-z)} \tag{4.13}$$

where $\beta_T(x) = \int_0^{\infty} f_T(t)e^{-xt}dt$ and $\beta_0(x) = \int_0^{\infty} f_0(t)e^{-x}dt$. The quantity π_0 is the probability of a unit being the one to initiate a busy period. We have

$$\pi_0 = \frac{E(\tau_0)}{E(\tau)} = \frac{1-b}{1-b+\lambda E(t_0)}$$

Avi-Itzhak and Yadin [12] demonstrate that π_n equals the probability that there are n units in front of server 2. For further details see [12].

4.2 Other References

We conclude this section by giving some additional references. Rao [80,81] obtained the mean effective service time at the first server assuming arbitrary service times, an unlimited supply of units in the first queue (i.e., saturated queue), type 1 blocking mechanism, and an intermediate waiting room of size $c \geq 0$.

Pinedo and Wolf [75] considered two queues 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.

Wijngaard [91] analyzed two queues in tandem with blocking and unreliable servers. The first queue was assumed to be saturated. General service times were assumed at both servers. 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 waitingroom is empty.

Finally, Newell [65] analyzed two finite queues in tandem using diffusion approximations. The service times at both queues and the inter-arrival times were assumed to have the same mean but arbitrary variances. In this type of approximation, each queue is treated as if it were a stochastic continuous fluid. The interested reader is referred to [65] for further details.

5. MULTIPLE SERVERS

We consider the two-node queueing network with blocking assuming that each queue is served by multiple servers, as shown in figure 3. Let s_1 and s_2 be the number of servers at the first and second queue. Units arrive at the system in a Poisson fashion

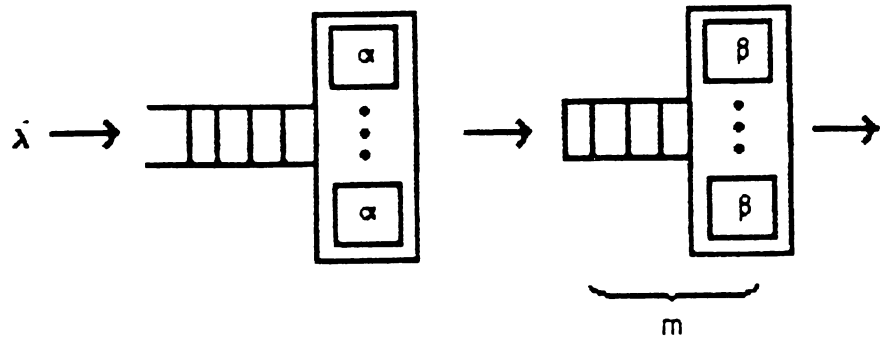


Figure 3: Two-node queueing network with blocking

at the rate λ. Let α and β be the mean time a server at the first and second queue respectively. Service times are assumed exponentially distributed. Finally, let m be the maximum capacity of the second queue. That is, it can accommodate s_2 units being served and $m-s_2(\geq 0)$ units waiting in its queue for service.

5.1 Saturated First Queue

We now proceed to obtain a closed-form analytic solution of the queueing network shown in figure 3, assuming that the first queue is saturated. Service times are assumed exponentially distributed. A unit, upon completion of its service at the first queue, will get blocked if at that moment the second queue is full. Its associated server will also get blocked. When a departure occurs from the second queue, one of the blocked servers at the first queue will become unblocked. It is not necessary to keep track in what order servers at the first queue get blocked, seeing they are identical. This model was first studied by Makino [59].

Let p_n, n=0,1,...,m,m+1,...,m+s_i, be the probability that there are n units in the second queue including the units in the first queue that are blocked. The steady-state equations associated with p_n are as follows:

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

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

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

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

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

Using the above equations we can easily obtain

$$p_i = \begin{cases} \frac{s_1^i}{i!}\left(\frac{\beta}{\alpha}\right)^i p_0 & i=1,2,\ldots,s_2 \\ \frac{s_1^i}{s_2! s_2^{i-s_2}}\left(\frac{\beta}{\alpha}\right)^i p_0 & , \quad i=s_2+1,\ldots,m \\ \frac{s_1^m (s_1)_{(i-m)}}{s_2! s_2^{i-s_2}}\left(\frac{\beta}{\alpha}\right)^i p_0 & , \quad i=m+1,\ldots,m+s_1 \end{cases} \tag{5.1}$$

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

$$p_0 = 1/\left[\sum_{i=0}^{s_2} \frac{s_1^i}{i!}\left(\frac{\beta}{\alpha}\right)^i + \sum_{i=s_2+1} \frac{s_1^i}{s_2! s_2^{i-s_2}}\left(\frac{\beta}{\alpha}\right)^i + \sum_{i=m+1}^{m+s_1} \frac{s_1^m (s_1)_{(i-m)}}{s_2! s_2^{i-s_2}}\left(\frac{\beta}{\alpha}\right)^i\right] \tag{5.2}$$

We note that the departure rate from the second queue is

$$\sum_{i=1}^{s_2-1} \left(\frac{i}{\beta}\right) p_i + \frac{s_2}{\beta} \sum_{i=s_2}^{m+s_1} p_i \tag{5.3}$$

which is equal to the departure rate from the first queue

$$\sum_{i=0}^{m} \left(\frac{s_1}{\alpha}\right) p_i + \sum_{i=1}^{s_1-1} \left(\frac{s_1-i}{\alpha}\right) p_{m+i} \tag{5.4}$$

where p_i ,i=0,...,$m_1 + s_1$, is given by expressions (5.1) and (5.2). We can now use expression (5.3) (or (5.4)) to obtain the condition for stability of this queueing network in the case where the first queue is assumed to be stable. For, in this case, expression (5.3) gives the maximum departure rate from the queueing network. Therefore, in order for the system to be stable, we should have

$$\lambda < \sum_{i=1}^{s_2-1} \left(\frac{i}{\beta}\right) p_i + \frac{s_2}{\beta} \sum_{i=s_2}^{m+s_1} p_i \quad . \tag{5.5}$$

The above expression agrees with the condition for stability derived by Latouche and Neuts [54] for the queueing network shown in figure 3 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 [59] also studied the output process of the queueing network considered in this section.

5.2 Stable First Queue

We now examine the queueing network shown in figure 3 assuming that the first queue is stable. Service times and inter-arrival times are exponentially distributed. This queueing network was analyzed by Latouche and Neuts [54] using the matrix-geometric procedure outlined in section 3.6. They considered three different blocking mechanisms based on the following two notions: a) The blocking of a server at the first queue 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 queue drops below a predefined number. The authors studied the above queueing system by allowing a feedback loop of departures from second queue to the first queue. 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

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

The above rate matrix has a structure which is more general than the one considered in section 3.6, and it is associated with quasi birth-death processes. Its stationary probability vector can be obtained using the matrix-geometric procedure outlined in section 3.6. The authors also obtained explicitly the stability condition of the queueing network for each of the three different blocking mechanism. The general form of this condition is given by

$$\underline{\pi}\, A_0\, \underline{e} > \underline{\pi}\, A_2\, \underline{E} \tag{5.6}$$

where $\underline{\pi}$ is obtained solving $\underline{\pi}\,(A_0 + A_1 + A_2) = 0$.

5.3 Other References

Asare [10] studied the queueing network shown in figure 3 assuming the blocking mechanism used in connection with the model studied in section 5.1. Departing units 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 3.1.

Krishna and Shin [50] analyzed the queueing network shown in figure 3, assuming $s_1=1$ and $m=s_2$.

6. UNRELIABLE SERVERS: EXPONENTIAL SERVICE TIMES

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.

6.1 Saturated First Queue

We first examine this model assuming that the first queue is always saturated, i.e. there is always one unit waiting to be served. In particular, we consider the following model analyzed by Gershwin and Berman [33]. 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. Below, we examine this queueing network under type 2 blocking mechanism.

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 queue including the one in service, and s_1,s_2 is the status of the first and second server respectively. The quantity s_i,i$-$1,2, takes the values of 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 at the first and second server respectively. Failure and repair times are assumed exponentially distributed. Then, the steady-state equations of the system under study are as follows, assuming that m=3.

$$\begin{cases} (r_1+r_2)p_{000} = p_1 p_{010} \\ (r_1+r_2)p_{i00} = p_1 p_{i10} + p_2 p_{i01} \quad , \ i=1,2 \\ r_1+r_2)p_{300} = p_2 p_{301} \end{cases} \tag{6.1}$$

$$\begin{cases} (p_1+(1/\alpha)+r_2)p_{010} = r_1 p_{00} \\ (p_1+(1/\alpha)+r_2)p_{i10} = r_1 p_{i00} + p_2 p_{i11} + (1/\alpha)p_{i-1,1,0} \quad ,i=1,2 \\ r_2 p_{310} = r_1 p_{300} + p_2 p_{311} + (1/\alpha)p_{210} \end{cases} \tag{6.2}$$

$$\begin{cases} r_1 p_{001} = r_2 p_{000} + p_1 p_{011} + (1/\beta)p_{101} \\ (r_1+(1/\beta)+p_2)p_{i01} = r_2 p_{i00} + p_1 p_{i11} + (1/\beta)p_{i+1,0,1} \quad , \quad i=1,2 \\ (r_1+(1/\beta)+p_2)p_{301} = r_2 p_{300} \end{cases} \tag{6.3}$$

$$\begin{cases} (p_1+(1/\alpha))p_{011} = r_1 p_{001} + r_2 p_{010} + (1/\beta)p_{111} \\ (p_1+(1/\alpha)+p_2+(1/\beta))p_{i11} = r_1 p_{i01} + r_2 p_{i10} + (1/\beta)p_{i+1,1,1} + (1/\alpha)p_{i-1,1,1} \quad ,i=1,2 \\ (p_2+(1/\beta))p_{311} = r_1 p_{301} + r_2 p_{310} + (1/\alpha)p_{211} \end{cases} \tag{6.4}$$

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

Gershwin and Berman [33] 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 very 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, we can very easily establish that

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

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 x prob [server i is broken down]= p_i x 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 $\mu_1[p(0,1,0) + p(0,1,1)] = \mu_2[p(1,0,1) + p(1,1,1)]$. Repeating the above process but using these equations which involve $p(n,s_1,s_2)$ for n=1,2,...,m we can obtain the following expression

$$\mu_1[p(n,1,0) + p(n,1,1)] = \mu_2[p(n+1,0,1)+p(n+1,1,1)] \quad ,n=0.1,...,m-1 \quad .$$

This implies that the arrival rate at the second queue, given that there are n units in the second queue, equals the departure rate from the second queue, given that there are n+1 units in the second queue.

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.

6.2 Stable First queue

Now, let us consider the queueing network studied above assuming that the first queue is stable (as opposed to being saturated as it was assumed in the analysis above). 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, including the one in service, in the first and second queue 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.

It can be easily verified that if we order the rates of this system lexicographically, the non-zero elements of the resulting rate matrix has the following block tri-diagonal structure.

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

In view of this, we note that the queueing network can be analyzed numerically using the matrix-geometric procedure (see Neuts [64]).

6.3 Other References

The queueing network analyzed above was also studied by Buzacott [20] assuming a

constant failure probability per service. Koenigsberg [47] also studied the same system assuming that the time between two successive failure of a server is exponentially distributed. Ignall and Silver [42] studied the above queueing network assuming that each queue is served by multiple servers. They gave a heuristic procedure for estimating the hourly output rate.

7. UNRELIABLE SERVERS: CONSTANT SYNCHRONIZED SERVICE TIMES

In this section, we examine the case of unreliable servers as in section 6, but assuming constant service times. In particular, let us consider the queueing network shown in figure 1 under a common fixed cycle time. Both servers start service at the beginning of a cycle and end service at the end of the same cycle. Thus, the service at both servers is constant and it is 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 fixed during the cycle with some 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 assembly lines.

7.1 Saturated First Queue

We first examine this model assuming that there is an unlimited supply of units in the first queue waiting to be served (i.e., the first queue is always saturated). In particular, we consider the model examined by Artamonov [9]. We make the folowing 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 breakdown 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 $0 \leq n \leq m$, 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 arrange the states of the system as follows. States are grouped together so that they all have the same number of units in the second node. The states within each group are sorted out 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

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

We note that the stationary probability vector can be obtained using Neuts' matrix-geometric procedure (see Neuts [64]). Artomonov [9] obtained the closed-form analytic solution by first setting-up the steady-state equations and subsequently manipulating these equations algebraically. The interested reader is referred to [9] for the expression of this closed-form solution.

7.2 Other References

Two-node queueing networks similar to the one analyzed above have also been studied by Okamura and Yamashina [68], Muth and Yeralan [62] and Gershwin and Schick [31]. In particular, Muth and Yeralan [62] analyzed the same system studied above 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.) The states of the system were enumerated in such a way so that the non-zero elements of the one-step transition probability pile P have the following structure

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

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 above. In this case, the stationary probability vector can then be obtained using Neuts' matrix-geometric procedure.

Finally, Gershwin and Schick [31] analyzed the system studied by Muth and Yeralan [62], 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 [33] and outlined in section 6.

8. TWO-NODE CLOSED QUEUEING NETWORKS WITH BLOCKING

We now proceed to examine a two-node closed exponential queueing network with blocking as shown in figure 4. Let N be the number of units constantly circulating

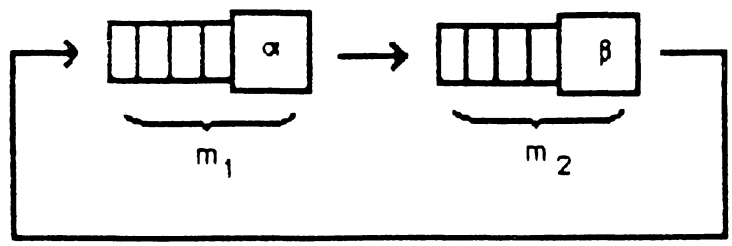


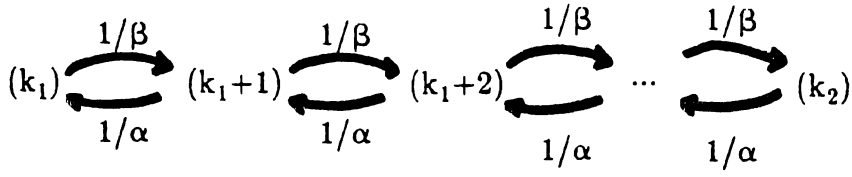
Figure 4: A two-node closed queueing network with blocking

in the queueing network. Let m_1 and m_2 be the size of the capacity of the first and second queue respectively (including the unit in service). We have, $N \leq m_1+m_2$. Service times at the first and second queue that are exponentially distributed with mean α and β respectively. We now proceed to examine the above queueing network under different types of blocking mechanisms. For presentation purposes, the analysis is restricted to the case where each queue is served by a single server. Clearly, the results obtained here can be easily extended to the case where each queue is served by multiple servers.

8.1 Blocking Mechanism Type 2.1

We consider the queueing network described above assuming that a server cannot start a new service if the other queue is full (type 2.1 blocking). This model was first studied by Gordon and Newell [34]. 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 queue 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 queue. Define $k_1= \max(0,N-m_2)$ and

$k_2 = \min(n, m_1)$. Then, $k_1 \leq n_1 \leq k_2$. The rate diagram associated with queue 1 has the following simple form



Thus, we can easily obtain that

$$p_1(n_1) = \left(\frac{\alpha}{\beta}\right)^{n_1 - k_1} p_1(k_1) \quad , \; k_1 \leq n \leq k_2 \quad , \tag{8.1}$$

and

$$p_1(k_1) = \begin{cases} \dfrac{1 - (\alpha/\beta)}{1 - (\alpha/\beta)^{k_2 - k_1 + 1}} & , \alpha \neq \beta \\ \\ 1/(k_2 - k_1 + 1) & , \alpha = \beta \end{cases} . \tag{8.2}$$

If $m_1, m_2 \geq N$, 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 service 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.

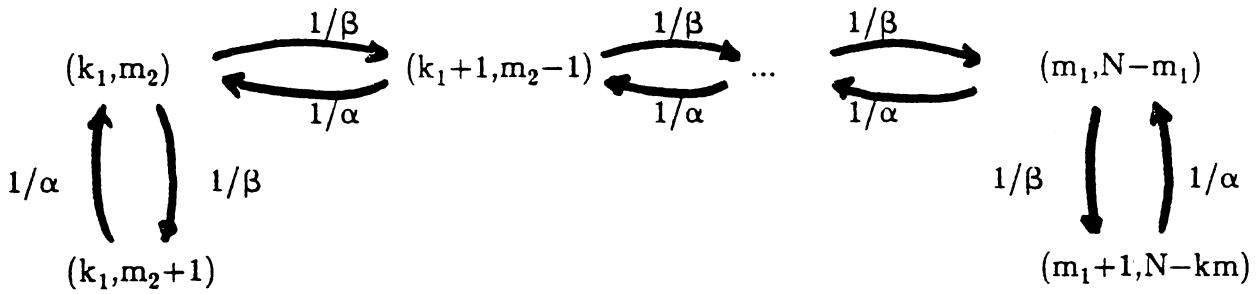
Finally, we note that the steady-state solutions (8.1) and (8.2) also hold if the above queueing network is considered under blocking mechanism type 3.

8.2 Blocking Mechanism Type 1

We analyze the queueing network shown in figure 4 under blocking mechanism type 1, by considering three different cases.

a) $m_1, m_2 < N$

Let k_1 and k_2 be the minimum and maximum occupancy of queue 1 as defined in section 8.1. Seeing that $m_1, m_2 < N$ we have $k_1 > 0$, $N - k_1 = m_2$. 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 queue 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

$$p(k_1+i, m_2-i) = \left(\frac{\alpha}{\beta}\right)^{i+1} p(k_1, m_2+1)\ ,\ 0 \le i \le m_1 - k_1 + 1 \tag{8.3}$$

and

$$p(k_1, m_2+1) = \begin{cases} \dfrac{1-(\alpha/\beta)}{1-(\alpha/\beta)^{m_1-k_1+3}} & ,\alpha \neq \beta \\ 1/(m_1-k_1+3) & ,\alpha = \beta \end{cases} \tag{8.4}$$

b) $m_1 \le N \le m_2$

In this case, the first server can never get blocked seeing that the capacity of the second queue is large enough to accommodate all N units. The rate diagram of this system can be obtained from the one above by ignoring the blocking state (k_1, m_2+1) and setting $k_1 = 0$ and $k_2 = m_1$. We can easily obtain that

$$p(n_1,N-n_1) = \left(\frac{\alpha}{\beta}\right)^{n_1} p(0,N)$$

and

$$p(0,N) = \begin{cases} \dfrac{1-(\alpha/\beta)}{1-(\alpha/\beta)^{m_1+2}} & ,\alpha \neq \beta \\ 1/(m_1+2) & ,\alpha = \beta \end{cases}$$

We note that the first queue is 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 under type 2.1 blocking mechanism in section 8.1, where for $m_1 \leq N \leq m_2$ the first queue becomes identical to an $M/M/1/m_1$ queue.

c) $m_1,m_2 \geq N$

In this case, there is no blocking at either queues. 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. This is the same result obtained for $m_1,m_2 \geq N$ in section 8.1.

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

$$z = \min\{N,m_1+s_2\} + \min\{N,m_2+s_1\} - N+1 \ .$$

REFERENCES

1. Akyildiz, I.F., "Exact Analysis of Queueing Networks with Multiple Servers and with Blocking", CS Rep., Louisiana State Univ., (1985)
2. Altiok, T., "Approximate Analysis of Production Lines with General Service and Repair Times and with Finite Buffers" ,IE Rep. 84-4, Rutgers Univ., (1984)
3. Altiok, T., "Approximate Analysis of Exponential Tandem Queues with Blocking", European J. Oper. Res., 11 (1982), 390-398
4. Altiok, T., and Perros, H.G., "Open Networks of Queues with Blocking: Split and Merge Configurations", CS Rep. 83-10, NC State Univ., (1983)
5. Altiok, T., and Perros, H.G., " Approximate Analysis of Arbitrary Configurations of Open Queueing Networks with Blocking", CS Rep. 85-03, NC State Univ., (1985)
6. Altiok, T., and Stidham, S., A Note on Transfer Lines with Unreliable Machines, Random Processing Times, and Finite Buffers", IIE Trans. (1982), 125-127
7. Altiok, T., and Stidham, S.," The Allocation of Interstage Buffer Capacities in Production Lines", IIE Trans., 15 (1983), 292-299.
8. Ammar, M.H., and Gershwin, S.B., " Equivalence Relation in Queueing Models of Manufacturing Networks", Proc. IEEE Conf. on Decision and Control (Albuquerque, 1980)
9. Artamonov, G.T., "Productivity of a Two-Instrument Discrete Processing Line in the Presence of Failures", Kibernetika 3, (1976) 126-130 (Engl. tr. Cybernetics, 12 (1977), 464-468)
10. Asare, B., "Queue Networks with Blocking", Ph.D. Thesis (1978), Trinity College Dublin, Ireland
11. Avi-Itzhak, B., "A Sequence of Service Station with Arbitrary Input and Regular Service Times", Mgmt. Sci., 11 (1965), 565-571
12. Avi-Itzhak, B., and Yadin, M., "A Sequence of Two Servers with no Intermediate Queue", Mgmt Sci., 11 (1965), 553-564
13. Balsamo, S., and Iazeolla, G., "Some Equivalence Properties for Queueing Networks With and Without Blocking", Performance'83, Agrawala and Tripathi(Eds.), North-Holland (1983), 351-360
14. Bell, P.C., "The Use of Decomposition Techniques for the Analysis of Open Restricted Queueing Networks", Oper. Res. Letters, 1 (1982) 230-235
15. 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

16. Bocharov, P.P., and Rokhas, G.P., "On an Exponential Queueing System in Series with Blocking", Problems of Control and Information Theory, 9 (1980), 441-455

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

18. Brandwajn, A., and Jow, Y.L, "An Approximation Method for Tandem Queues with Blocking", Amdahl Corp., (1985)

19. Buzacott, J.A., "Automatic Transfer Lines with Buffer Stocks", Int. Jnl. Prod. Res., 5 (1967), 183-200

20. 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

21. Buzacott, J.A., and Hanifin, L.E., "Models of Automatic Transfer Lines with Inventory Banks- A Review and Comparison", AIIE Trans., 10 (1978), 197-207

22. 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

23. Clarke, A.B., "Markovian Queues with Servers in Tandem", Math. Rep. 49, Western Michigan Univ., (1977)

24. Clarke, A.B., "A Two-Server Tandem Queueing System with Storage between Servers", Math. Rep. 50, Western Michigan Univ., (1977)

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

26. Clarke, A.B., "Waiting Times for Markorvian Queue with Servers in Series", Math. Rep. 52, Western Michigan Univ., (1978)

27. Evans, R.V., "Capacity of Queueing Networks", Oper. Res.,15 (1967), 530-536

28. Evans, R.V., "Geometric Distributions in Some Two-Dimensional Queueing Systems", Oper. Res., 15 (1967) 830-846

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

30. Foster, F.G., and Perros, H.G., "Hierarchical Queue Networks with Partially Shared Servicing", J. Opl. Res. Soc., 30 (1979), 157-166

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

32. Gershwin, S.B., "An Efficient Decomposition Method for the Approximate Evaluation of Tandem Queues with Finite Storage Space and Blocking", Manuscript, MIT, Lab. for Information and Decision Sciences, (1983)

33. 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

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

35. Goto, K., Takahashi, Y., and Hasegawa, T., "An Approximate Analysis of Controlled Tandem Queues", Proc. Modelling Techniques and Tools for Performance Analysis, June 1985, Cannes France

36. 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), p. 273)

37. Hildebrand, D.K., "Stability of Finite Queue, Tandem Server Systems", J. Appl. Prob., 4(1967), 571-583

38. Hildebrand, D.K., "On the Capacity of Tandem Server, Finite Queue, Service Systems", Oper. Res.,16 (1968), 72-82

39. Hillier, F.S., and Boling, R.W., "Finite Queues in Series with Exponential or Erlang Service Times - A Numerical Approach", Oper. Res., 15 (1967), 286-303

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

41. Hordijk, A., and Van Dijk, N., "Networks of Queues with Blocking", Performance'81, Kylstra(Ed.),North-Holland (1981), 51-65

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

43. Jafari, M.A., and Sharthikumar, J.G., "Allocation of Buffer Storages Along a Multi-Stage Automatic Transfer Line", IE Rep. 84-003, Arizona Univ., (1980)

44. Kelly, F.P., "The Throughput of a Series of Buffers", Adv. Appl. Prob., 14 (1982), 633-653

45. Kelly, F.P., "Blocking, Reordering, and the Throughput of a Series of Servers", Stochastic Processes and their Appl., 17 (1984) 327-336

46 Knott, A.D., "The Inefficiency of a Series of Work Stations - A Simple Formula", Int. Jnl. Prod. Res., 8 (1970), 109-119

47. Koenigsberg, E., "Production Lines and Internal Storage - A Review", Mgmt. Sci. 5 (1959) 410-433

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

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

50. 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

51. Labetoulle, J., and Pujolle, G., "A Study of Queueing Networks with Deterministic Service and Application to Computer Networks", Proc. on Computer Performance Modeling, Measurement and Evaluation, Chen and Franklin(Eds.), ACM (1976), 230-240

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

53. Labetoulle, J., and Pujolle, G., "Isolation Method in a Network of Queues", IEEE Trans. Soft. Eng., SE-6 (1980), 373-381

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

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

56. Lazar, A.A., and Robertazzi, T.G., "The Geometry of Lattices for Markovian Queueing Networks", Manuscript, Columbia Univ., Electrical Engineering Dept., (1984)

57. Le Ny, L.-M., "Forme Produit Pour des Reseaux Multiclasses a Routage Dynamiques", Annales Scientifiques de l'Universite de Clermont-Ferrand II, 76 (1983) 17-34

58. Le Ny, L.-M., "Etude Analytique de Reseaux de Files d'Attende Multiclasses a Routage Variable", RAIRO Recherche Operationnelle, 14 (1980) 331-347

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

60. Muth, E.J., "The Production Rate of a Series of Work Stations with Variable Service Times", Int. J. Prod. Res., 11(1973), 155-169

61. Muth, E.J., "The Reversibility Property of Production Lines", Mgmt. Sci., 25 (1979), 152-158

62. 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

63. Neuts, M.F., "Two Queues in Series with A Finite, Intermediate Waiting-room", J. Appl. Prob., 5 (1968), 123-142

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

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

66. Nilsson, A.A., and Altiok T., "Open Queueing Networks with Finite Capacity Queues", Proc. 1981 Int'l Conf. Parallel Processing, (1981) 87-91

67. Ohmi, T. "An Approximation for the Production Efficiency of Automated Transfer Lines with In-Process Storage", AIIE Trans., 13 (1981), 22-28

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

69. Onvural, R.O., and Perros, H.G., "On Equivalences of Blocking Mechanisms in Queueing Networks with Blocking", CCSP 85/9, NC State Univ., (1985)

70. Patterson, R.L., "Markov Processes Occurring in the Theory of Traffic Flow Through an N-Stage Stochastic Service System", J. Ind. Eng. 14 (1964) 188-193

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

72. Perros, H.G., "A Symmetrical Exponential Open Queue Network with Blocking and Feedback", IEEE Trans. Soft. Eng., SE-7 (1981), 395-402

73. Perros, H.G., "A Two-Level Open Queue Network with Blocking and Feedback", RAIRO Recherche operationnelle/Oper. Res.,15 (1981),27-38

74. Perros, H.G., and Altiok, T., "Approximate Analysis of Open Networks of Queues with Blocking : Tandem Configurations", CS Rep. 83-11, NC State Univ., (1984)

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

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

77. Pittel, B., "Closed Exponential Networks of Queues with Saturation: The Jackson-Type Stationary Distribution and its Asymptotic Analysis", Math. Oper. Res., 4 (1979), 367-378

78. Pollock, S.M., and Birge, J.R., "Parallel Iteration for Multiple Servers", Ind. and OR 83-16, Michigan Univ., (1983)

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

80. 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

81. 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

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

83. Sevast'yanov, B.A., "Influence of Storage Bin Capacity on the Average Standstill Time of a Production Line", Teoriya Veroyatnostey i ee Primeneniya, 7 (1962), 429-438 (Engl. tr. Theory of Probability and its Applications, 7 (1962), 429-438

84. Sheskin, T.J., "Allocation of Interstage Storage along an Automatic Production Line", AIIE Trans., 8 (1976), 146-152

85. Soyster, A.L., and Toof, D.I., "Some Comparative & Design Aspects of Fixed Cycle Production Systems", Naval Research Logistics Quarterly, 23 (1976), 437-454

86. Soyster, A.L., Schmidt, J.W., and Rohrer, M.W., "Allocation of Buffer Capacities for a Class of Fixed Cycle Production Lines", AIIE Trans., 11 (1979), 140-146

87. Suri, R., and Diehl, G.W., "A Variable Buffer-Size Model and its Use in Analytic Closed Queueing Networks with Blocking", Proc. ACM SIGMETRICS on Measurement and Modelling of Computer Systems, (1984) 134-142

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

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

90. Yao, D., and Buzacott, J.A., "Modeling a Class of Flexible Manufacturing Systems with Reversible Routing", IE Rep. 83-02, Univ. of Toronto, (1983)

91. Yao, D., and Buzacott, J.A., "Modelling the Performance of Flexible Manufacturing Systems", Manuscript, Columbia Univ., I.E. Dept., (1983)

92. 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

93. 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