

## Theory and Methodology

---

# Routing and buffer allocation models for a telecommunication system with heterogeneous devices

Behnam Pourbabai

*Department of Mechanical Engineering, University of Maryland, College Park, MD 20742, USA*

Abraham Seidmann

*W.E. Simon Graduate School of Business Administration, University of Rochester, Rochester, NY 14627, USA*

Received February 1990; revised December 1990

**Abstract:** A branching system which consists of several parallel nonidentical communication devices with finite input buffers is considered in this paper. The system is designed and operated under a stochastic congestion control policy. Several job routing and buffer allocation decision models are analyzed in order to maximize the aggregate system throughput, or to minimize the storage space costs. The polynomial and pseudo-polynomial optimization algorithms which are presented here take advantage of the structural properties of this system. The results obtained are also of particular relevance in the design of load balancing policies for distributed computer systems.

**Keywords:** Telecommunication; system design; queueing network; optimization

### Introduction

Queueing network models are used extensively for the design and analysis of numerous communication, computer and manufacturing systems (Gelenbe and Pujolle, 1987; Kleinrock, 1975; Shalev-Oren, et al., 1985; Tenenbaum, 1981). Typical performance measures derived from these models include node congestion, device utilization, response times, and system throughput. In general, a queueing network can be viewed as a set of arbitrarily linked processing stations. Blocking occurs in these networks when the flow of units is interrupted momentarily due to the fact

that some other queue in the network has reached its capacity limitations. Queueing networks with blocking are difficult to analyze as closed-form solutions for mean sojourn times, or the queue length distributions are attainable only under very limited assumptions (Kingman, 1969; Onvural, 1988). Studies of such systems reported today use mainly heuristic decompositions, numerical approximations, and digital simulations (Perros, 1984).

This paper examines a simple multi-way branching telecommunication network with finite queueing buffers. Arrivals of new messages to the system form a Poisson process of rate  $\gamma$ . Upon

arrival, these messages are routed to be processed by one of  $L$  parallel heterogeneous devices (Figure 1). These devices may be communication controllers operating with distinct Baud rates, memory banks in mainframe computers, or other generic parallel service centers (i.e., production lines, toll gates, or drive-in bank facilities). New messages are routed to device  $i$  ( $i = 1, 2, \dots, L$ ) with probability  $p_i$  and hence the arrival process to the input buffer of device  $i$  is also Poisson, with rate  $\gamma_i = p_i \gamma$  (Genedenko and Kovalenko, 1968). Each device and its input buffer have room for  $K_i$  ( $\geq 1$ ) messages: the one being processed plus up to  $K_i - 1$  in the input buffer. A new message routed to a full buffer is lost. A first come, first served discipline is followed when each device admits messages from its own input buffer. Device  $i$  has a processing rate which is exponentially distributed with mean  $1/\mu_i$ . Messages leave the system after being processed by their respective devices.

This model, which is a generalization of the classical M/M/C structure was first posed by Larsen (1981), who also presented an extensive motivation in the context of computer systems design. Assuming a shared input buffer, Larsen (1981) studied the optimal message routing policy for a two-device network; Lin and Kumar (1984) later presented an elegant proof showing that the optimal policy which minimizes the mean sojourn time of messages in this two-device network is of threshold type. Other studies of shared-memory

systems include Tijms and Eikeboom (1986). The problem of optimal part routing in the case of dedicated parallel storage buffers was first studied by Hahne (1981) and later extended by Seidmann and Schweitzer (1984), Seidmann and Tenenbaum (1987), and by Daskalaki and Smith (1988).

The system performance depends on the job routing policy, as well as on the (controllable) arrival rate and on the way the limited memory resources are allocated among the various devices. Although several studies have recognized the impact of the local input buffers on the performance of such branching networks, no analytical formulation has been presented to assure the optimality of any buffer allocation policy. Several aspects of memory allocation decisions and their interaction with the job routing scheme are examined in this paper.

The organization of this paper is as follows. In Section 2, the main components of the models are presented. More specifically, in Section 2.3.1, for a given topology and specified buffer sizes, the net throughput rate is maximized while controlling the bottleneck; in Section 2.3.2, for a given topology and specified arrival rates, the optimal buffer sizes are selected while controlling the bottleneck; in Section 2.3.3, an extension of the previous model is presented and solved; and in section 2.3.4, generalized versions of model 1 and 2 are provided. There, for a given topology, the optimal buffer sizes are selected to maximize

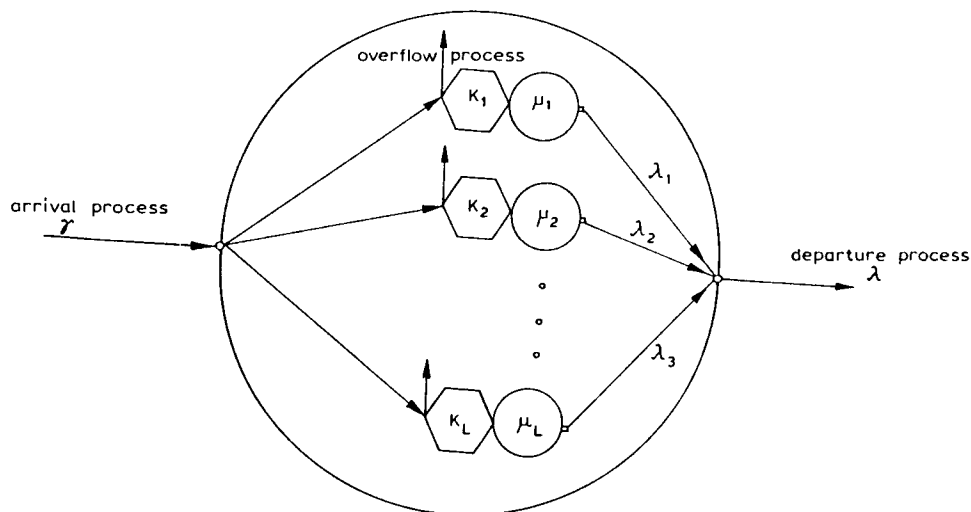


Figure 1. A telecommunication system with parallel finite-capacity devices

the throughput rate while controlling the bottleneck. In Section 3, the concluding remarks are discussed.

## 2. Main results

In this section, several models are proposed for designing the telecommunication system detailed above, see Figure 1.

### 2.1. Assumptions

The models presented in this paper assume that the acceptable probability of finding congestion at each device (in excess of its input storage capacity) is sufficiently close to zero. That is, the performance of each finite-capacity M/M/1/ $K_i$  Markovian queue (Kleinrock, 1967) device with Poisson input, exponential service times, first come, first served queueing discipline, and an input buffer of size  $K_i - 1$  will be approximated with the performance of a compatible infinite-capacity M/M/1 queueing model. This approximation is based on the congestion control requirement that the system be designed and operated in such a way that the resulted congestion probability is bounded by a very small acceptable threshold. For instance, given that the processing rate is fixed, the arrival rate may be set such that the congestion probability is less than 0.001 at each device. This value also establishes the maximal error associated with the use of the above approximation for computing the expected throughput. It is noted that this perspective in control of the blocking and the bottleneck is motivated and evaluated in Pourbabai (1989a, 1990). There, extensive simulation results are provided to investigate the accuracy of the proposed approximation technique. Hence, in this paper, for brevity, numerical results are not provided, and the interested reader is referred to Pourbabai (1989a,b, 1990).

### 2.2. Notation

$L$ : Total number of parallel devices.  
 $i$ : Device index,  $i = 1, 2, \dots, L$ .  
 $\alpha_i$ : Acceptable probability of finding device  $i$  blocked.

$K_i, \hat{K}_i$  and  $K_i^*$ : Input buffer of device  $i$ , its maximum allowable capacity, and its optimal capacity, respectively [messages].

$\mu_i$ : Processing rate of device  $i$  [messages/time unit].

$P_i$ : Fraction of units routed to device  $i$ .

$\gamma_i$  and  $\gamma_i^*$ : Arrival rate and the maximum arrival rate at device  $i$ , respectively [messages/time unit].

$\lambda_i$  and  $\lambda_i^*$ : Throughput rate and the maximum throughput rate of device  $i$ , respectively [messages/time unit].

$c_i$ : Marginal allocation cost of increasing the buffer size of device  $i$  by one unit [\$/message].

$n_i, P(n_i = \cdot)$ : Number of units found at device  $i$ , and its probability mass function.

$K$ : Total available capacity [messages].

$\gamma$  and  $\gamma^*$ : Arrival rate and the maximum arrival rate at the system, respectively [messages/time unit].

$\lambda$  and  $\lambda^*$ : Throughput rate and the maximum throughput rate from the system, respectively [messages/time unit].

It is well-known that in an M/M/1 system

$$P(n_i = k_i) = \left(1 - \frac{\gamma_i}{\mu_i}\right) \left(\frac{\gamma_i}{\mu_i}\right)^{k_i},$$

$$k_i = 0, 1, 2, \dots \text{ and } i = 1, 2, \dots, L. \quad (1)$$

This expression is now used to approximate the overflow probability of an M/M/1/ $K_i$  queueing system:

$$P(n_i \geq k_i) = \left(\frac{\gamma_i}{\mu_i}\right)^{k_i} \equiv \alpha_i, \quad i = 1, 2, \dots, L. \quad (2)$$

Note that (2) means that

$$\lambda_i = \lim_{\alpha_i \rightarrow 0} \gamma_i, \quad i = 1, 2, \dots, L. \quad (3)$$

The four network design models developed next assume a finite value of  $\lambda_i, i = 1, 2, \dots, L$ .

### 2.3. Network design models

#### 2.3.1. Model 1

In this model, we treat the case of a given topology and buffer allocation. The message routes must be determined in order to maximize the overall system throughput. It means that the values of  $\alpha_i, \mu_i$ , and  $K_i$  ( $i = 1, 2, \dots, L$ ) are given. Finding the values of  $P_i$  also fixes  $\gamma_i, \lambda_i, \gamma$ ,

and  $\lambda$ . For this purpose, the following chance-constrained optimization model is introduced.

$$\text{Max } \{\lambda\} \quad (4)$$

$$\text{s.t. } P(n_i \geq K_i) \leq \alpha_i, \quad i = 1, 2, \dots, L, \quad (5)$$

$$\sum_{i=1}^L \lambda_i = \lambda, \quad (6)$$

$$0 \leq \lambda_i < \mu_i, \quad i = 1, 2, \dots, L. \quad (7)$$

In the above model, expression (5) prevents the resulted blocking probability (i.e., the left-hand side of the inequality) of each device to exceed the acceptable blocking probability (i.e., the right-hand side of the inequality), expression (6) quantifies the throughput rate from the system, and expression (7) is provided for the stability of each device. Following (1)–(3), the model stated in (4)–(7) is reformulated as follows:

$$\text{Max } \{\lambda\} \quad (8)$$

$$\text{s.t. } \left( \frac{\gamma_i}{\mu_i} \right)^{K_i} \leq \alpha_i, \quad i = 1, 2, \dots, L, \quad (9)$$

$$\sum_{i=1}^L \lambda_i = \lambda, \quad (10)$$

$$0 \leq \gamma_i < \mu_i, \quad i = 1, 2, \dots, L. \quad (11)$$

Hence, the maximum throughput rate for the above problem based on expression (9) is

$$\gamma^* = \lambda^* = \sum_{i=1}^L \gamma_i^* = \sum_{i=1}^L \lambda_i^* \quad (12)$$

where

$$\gamma_i^* = \lambda_i^* = \mu_i (\alpha_i)^{1/K_i}, \quad i = 1, 2, \dots, L. \quad (13)$$

Moreover, the optimal routing probabilities are

$$P_i = \frac{\lambda_i^*}{\lambda^*}, \quad i = 1, 2, \dots, L.$$

### 2.3.2. Model 2

The model handles the case of a given set of devices along with the routing probabilities and the total available storage capacity. The desired economical allocation of storage capacity for each device is now determined subject to technological and congestion control constraints. It means that now the values of  $\lambda_i$ ,  $\mu_i$ ,  $\alpha_i$ ,  $\hat{K}_i$ ,  $c_i$  ( $i =$

$1, 2, \dots, L$ ), and  $K$  are given while the values of  $K_i$  are sought. For this purpose, the following chance-constrained optimization model is presented.

$$\text{Min } \left\{ \sum_{i=1}^L c_i K_i \right\} \quad (14)$$

$$\text{s.t. } P(n_i \geq K_i) \leq \alpha_i, \quad i = 1, 2, \dots, L, \quad (15)$$

$$\sum_{i=1}^L K_i = K, \quad (16)$$

$$\hat{K}_i \geq K_i \geq 0 \quad (17)$$

where  $K_i$  is integral, for  $i = 1, 2, \dots, L$ .

Expression (14) in the above model minimizes the total buffer allocation cost, expression (15) is similar to expression (5), and expression (16) guarantees that the total number of allocated buffers is equal to  $K$ . Specializing the model according to (1)–(3), it is restated as follows:

$$\text{Min } \left\{ \sum_{i=1}^L c_i K_i \right\} \quad (18)$$

$$\text{s.t. } (\lambda_i / \mu_i)^{K_i} \leq \alpha_i, \quad i = 1, 2, \dots, L, \quad (19)$$

$$\sum_{i=1}^L K_i = K, \quad (20)$$

$$\hat{K}_i \geq K_i \geq 0, \quad i = 1, 2, \dots, L. \quad (21)$$

However, expressions (19) and (21) can be restated as follows:

$$\hat{K}_i \geq K_i \geq \frac{\log \alpha_i}{\log(\lambda_i / \mu_i)}, \quad i = 1, 2, \dots, L. \quad (22)$$

Notice that the right-hand side of (22) identifies the *minimum* required buffer capacity for device  $i$ . This quantity is later denoted as  $\Delta_i$  in (27). Analysis of (9)–(11) leads to the following properties:

**Property 1.** *The throughput rate of each device is an increasing concave function of its input buffer capacity.*

**Proof.** Given  $\alpha_i$  and  $\mu_i$  values, let  $\gamma_i(K)$  denote the throughput rate of device  $i$  as a function of its input buffer capacity.

(i) *Increasing:* It is required to show that

$$\gamma_i(K+1) > \gamma_i(K), \quad K = 1, 2, \dots \quad (23)$$

or equivalently, from (9), for all  $0 < \alpha_i < 1$ :

$$\alpha_i^{1/(K+1)} > \alpha_i^{1/K}. \tag{24}$$

(ii) *Concave*: Since  $K$  is integer, the following should be true:

$$\gamma_i(K) + \gamma_i(K + 2) < 2\gamma_i(K + 1), \quad K = 1, 2, \dots \tag{25}$$

This obviously holds for all  $0 < \alpha_i < 1$ , as

$$\alpha_i^{1/K} + \alpha_i^{1/(K+2)} < 2\alpha_i^{1/(K+1)}. \tag{26}$$

**Property 2.** *The blocking probability of each device is a decreasing convex function of its input buffer capacity.*

The proof of Property 2 follows similar lines to the previous proof and it is omitted for brevity.

Exploiting these properties, the optimal solution can be obtained using the marginal allocation method (Fox, 1966).

*Step 1.* Set

$$\Delta_i = \frac{\log \alpha_i}{\log(\lambda_i/\mu_i)} \tag{27}$$

if

$$\Delta_i > \hat{K}_i \quad \text{for one } i \text{ or more, } \quad i = 1, 2, \dots, L, \tag{28}$$

or

$$\sum_{i=1}^L \Delta_i > K, \tag{29}$$

or

$$\sum_{i=1}^L \hat{K}_i < K, \tag{30}$$

stop, no feasible solution exists. Otherwise, set  $K_i = \Delta_i$ , or  $i = 1, 2, \dots, L$ , and go to Step 2.

*Step 2.* Set

$$K = K - \sum_{i=1}^L K_i, \tag{31}$$

$$\theta_i = \hat{K}_i - K_i, \tag{32}$$

$$\phi_i = \text{Min}(\theta_i, K) \quad \text{for } i = 1, 2, \dots, L. \tag{33}$$

If  $\phi_i = 0$ , for  $i = 1, 2, \dots, L$ , stop. An optimal feasible solution is approximated. Otherwise, go to Step 3.

*Step 3.* Choose index  $s$  ( $1 \leq s \leq L$ ) such that

$$\{s: \phi_s > 0 \text{ and } c_s = \text{Min}(c_1, \dots, c_L)\}. \tag{34}$$

Then, set  $K_s = K_s + \phi_s$ , and go to Step 2.

The algorithm discussed above can be augmented to handle the general case of nonlinear storage allocation costs. It can be shown that by following Steps 1–3, one will reach an optimal solution in  $O(L + K \log L)$  time for any convex objective function (Knuth, 1973).

### 2.3.3. Model 3

This subsection presents the following extension to model 2 above. This model is a dynamic programming formulation and has been accordingly solved. For a review of the related dynamic programming perspectives, see Yamashita and Suzuki (1987) and Jafari and Shanthikumar (1989). Here, we consider the case of a nonlinear cost function  $g_i(K_i)$  where

$$\frac{\partial g_i(K_i)}{\partial K_i} > 0, \quad i = 1, 2, \dots, L. \tag{35}$$

(Otherwise, the solution is trivial.)

The optimization problem now becomes:

$$\text{Min} \left\{ \sum_{i=1}^L g_i(K_i) \right\} \tag{36}$$

s.t.

$$\sum_{i=1}^L K_i = K, \tag{37}$$

$$\hat{K}_i \geq K_i \geq \bar{K}_i \equiv \frac{\log \alpha_i}{\log(\lambda_i/\mu_i)}, \quad i = 1, 2, \dots, L. \tag{38}$$

For brevity, denote

$$\sum_{i=1}^L \bar{K}_i = \bar{K}. \tag{39}$$

Since  $g_i(K_i)$  may not be all convex, the following pseudo-polynomial dynamic programming

scheme is proposed. Start by defining the state value as:

$$f_i^*(S) = \text{Min} \left\{ \sum_{j=1}^i g_j(K_j) \mid \sum_{j=1}^i K_j = S, \right. \\ \left. \hat{K}_i > K_j > \bar{K}_i, j = 1, 2, \dots, i \right\}$$

$$\forall i = 1, 2, \dots, L \text{ and } S = K - \bar{K}, \dots, K. \quad (40)$$

The optimal problem solution is given by  $f_L^*(K)$ .

The recursive relationship among the state variables is given by:

$$f_i^*(S) = \text{Min} \{ f_{i-1}^*(S - q) + g_i(q) \mid \\ q = 0, 1, 2, \dots, S \}$$

$$\forall S = K - \bar{K}, \dots, K, \text{ and } \forall i = 2, 3, \dots, L. \quad (41)$$

The boundary conditions are:

$$f_1^*(S) = g_1(S), \quad S = K - \bar{K}, \dots, K. \quad (42)$$

Hence, the optimal solution can be approximated by the following procedure:

*Step 1.* If  $\bar{K}_i > \hat{K}_i$   
for one  $i$  ( $i = 1, 2, \dots, L$ ) or more,  
or if

$$\sum_{i=1}^L \bar{K}_i > K,$$

stop, no possible solution exists.

*Step 2.* Set

$$f_1^*(S) = g_1(S), \quad S = K - \bar{K}, \dots, K. \quad (43)$$

Let  $i = 2$ , and go to Step 3.

*Step 3.* If  $i = L$ , go to Step 4.

Otherwise, set

$$f_i^*(S) = \text{Min} \{ f_{i-1}^*(S - q) + f_i(q) \mid \\ q = 0, 1, \dots, S \}$$

$$\text{for } S = K - \bar{K}, \dots, K. \quad (44)$$

Let  $i = i + 1$  and repeat Step 3.

*Step 4.* If  $i = L$ , let

$$f_L^*(K) = \text{Min} \{ f_{L-1}^*(K - q) + g_L(q) \mid \\ q = K - \bar{K}, \dots, K \}. \quad (45)$$

and stop with the optimal solution being  $f_L^*(K)$ .

**Lemma.** *The dynamic programming procedure correctly solves the optimization problem in  $O(L(K - \bar{K})^2)$  time (i.e., it is pseudo-polynomial), if the evaluation of each cost function  $g_i(K_i)$  is done in constant time.*

**Proof.** *Correctness* is proved by contradiction. Assume a vector

$$h^T = (h_1, h_2, \dots, h_i) \in \mathbb{R}I^+ \quad \text{with } \sum_{j=1}^i h_j = S,$$

$$\hat{K}_j > k_j > \bar{K}_j$$

$$\forall j = 1, 2, \dots, i \text{ and } \sum_{j=1}^i g_j(h_j) = f_i^*(S). \quad (46)$$

We want to show that (46) is sufficient for the following recursion to hold:

$$\sum_{j=1}^{i-1} g_j(h_j) = f_{i-1}^*(S - h_i). \quad (47)$$

First note that

$$\sum_{j=1}^{i-1} h_j = S - h_i, \quad (48)$$

and that the minimum condition in (40) requires

$$\sum_{j=1}^{i-1} g_j(h_j) \geq f_{i-1}^*(S - h_i). \quad (49)$$

But, if

$$\sum_{j=1}^{i-1} g_j(h_j) > f_{i-1}^*(S - h_i), \quad (50)$$

then we can find another vector  $H = (H_1, H_2, \dots, H_i) \in \mathbb{R}I^+$  which satisfies both

$$\sum_{j=1}^{i-1} H_j = S - h_j, \quad \hat{K}_j \geq H_j \geq \bar{K}_j \quad \forall j = 1, 2, \dots, i, \quad (51)$$

and

$$\sum_{j=1}^{i-1} g_j(H_j) > f_{i-1}^*(S - h_i). \quad (52)$$

Moreover, the vector

$$V = (H_1, H_2, \dots, H_{i-1}, h_i) \quad (53)$$

clearly satisfies

$$\sum_{j=1}^{i-1} H_j + h_i = S, \quad (54)$$

and (from (42) and (44)):

$$\begin{aligned} \sum_{j=1}^i g_j(V_j) &= f_{i-1}^*(S - h_i) + g_i(h_i) \\ &< \sum_{j=1}^i g_j(h_j) = f_i^*(S). \end{aligned} \quad (55)$$

However, the inequality in (55) contradicts the definition of (40) and hence (49) holds as equality. The recursive relation (41) is parametrized over all  $q$ -values from 0 to  $S$  in order to compute the desired values of  $h_i$ .

*Time complexity:* A time of  $O(K - \bar{K})$  is required to compute each  $f_i^*(S)$  and  $O(L(K - \bar{K})^2)$  is required to get  $f_i^*(S)$  for all  $i = 2, 3, \dots, L - 1$  and  $S = 0, 1, 2, \dots, K - \bar{K}$ . Step  $f_i^*(S)$  requires  $O(K - \bar{K})$  time. The overall time complexity is bounded by  $O(L(K - \bar{K})^2)$ .

#### 2.3.4. Model 4

The model discussed here addresses both message routing and buffer allocation decisions aimed at maximizing the aggregate system throughput. The given system is characterized by the values of  $\mu_i$ ,  $\alpha_i$ ,  $\hat{K}_i$  ( $i = 1, 2, \dots, L$ ), and  $K$ . The model obtains the desired values of  $K_i$  and  $P_i$ , along with  $\lambda_i^*$ ,  $\lambda$ ,  $\gamma_i^*$ , and  $\gamma$ . Notice that model 1 is a special case of model 4. To find the optimal buffer allocation and job routes, the following chance-constrained optimization model is formulated:

$$\text{Max } \{\lambda\} \quad (56)$$

$$\text{s.t. } P(n_i \geq K_i) \leq \alpha_i, \quad i = 1, 2, \dots, L, \quad (57)$$

$$\sum_{i=1}^L \lambda_i = \lambda, \quad (58)$$

$$\sum_{i=1}^L K_i = K, \quad (59)$$

$$0 \leq \lambda_i < \mu_i, \quad i = 1, 2, \dots, L, \quad (60)$$

$$\hat{K}_i \geq K_i \geq 0 \quad (61)$$

where  $K_i$  is integral, for  $i = 1, 2, \dots, L$ .

In the above model, expressions (57), (58), (60), and (50) are similar to expressions (5), (6), (7), and (16), respectively. Assuming  $\alpha_i \rightarrow 0$  for all  $i$ , the model can be restated as follows:

$$\text{Max } \{\lambda\} \quad (62)$$

$$\text{s.t. } (\lambda_i/\mu_i)^{K_i} \leq \alpha_i, \quad i = 1, 2, \dots, L, \quad (63)$$

$$\sum_{i=1}^L \gamma_i = \lambda, \quad (64)$$

$$\sum_{i=1}^L K_i = K, \quad (65)$$

$$0 \leq \lambda_i < \mu_i, \quad i = 1, 2, \dots, L, \quad (66)$$

$$\hat{K}_i \geq K_i > 0, \quad i = 1, 2, \dots, L. \quad (67)$$

In the above model, expression (63) is replaced by

$$\gamma_i \leq \mu_i(\alpha_i)^{1/K_i}. \quad (68)$$

Recall that for  $\alpha_i < 1$ ,  $\gamma_i$  is a stepwise increasing function of  $K_i$ .

Therefore, the optimal solution is approximated using the greedy method as follows:

*Step 1.* Set  $K_i = 1$ ,  $\lambda_i = 0$ , for  $i = 1, 2, \dots, L$ , if

$$\sum_{i=1}^L K_i \leq K,$$

go to Step 2. Otherwise, stop because no feasible solution exists.

*Step 2.* Set

$$D_i = \begin{cases} 0 & \text{if } K_i = \hat{K}_i, \\ \mu_i(\alpha_i^{1/K_{i+1}} - \alpha_i^{1/K_i}), & \text{otherwise.} \end{cases} \quad (69)$$

Choose the index  $s$  such that:

$$D_s = \text{Max}\{D_j : j \text{ such that } D_j > 0 \text{ for } j = 1, 2, \dots, L\}. \quad (70)$$

Then, set

$$K_s = K_s + 1, \quad (71)$$

$$\lambda_s = D_s + \lambda_s, \quad (72)$$

$$K_i = K_i \text{ for } i \neq s \text{ and } i = 1, 2, \dots, L, \quad (73)$$

$$K = K - 1, \quad (74)$$

and go to Step 3.

Step 3. If  $K > 0$ , got to Step 2.

Otherwise, compute the optimal results:

$$\lambda_i^* = \gamma_i^* = \mu_i(\alpha_i)^{1/K_i}, \quad i = 1, 2, \dots, L, \quad (75)$$

$$\lambda^* = \gamma^* = \sum_{i=1}^L \lambda_i, \quad (76)$$

$$P_i = \frac{\lambda_i^*}{\lambda^*}, \quad i = 1, 2, \dots, L. \quad (77)$$

### 3. Concluding remarks

An efficient design methodology is presented to optimally allocate storage buffers among a set of heterogeneous devices in a branching network topology. This methodology consists of several optimization models, all sharing a common set of basic operational and cost assumptions. The first model presented derives the throughput maximizing message routes in the case of a given buffer allocation. The second model assumes that these routes are already given. It then determines the minimal-cost buffer allocation policy subject to congestion control constraints. The third model extends the analysis to handle nonconvex cost functions. These cost functions result in a pseudo-polynomial algorithm, assuming that the evaluation of each cost function can be done in a constant amount of time. Finally, the fourth model simultaneously determines both message routes and buffer allocation decisions aimed at maximizing the aggregate system throughput. An appropriate upper bound on the external arrival rate is also suggested by each model in order to assume a minimally controlled loss of messages due to buffer congestion. It is noted that in this paper, because the solutions were presented in closed forms, no numerical results were provided. Moreover, as was noted earlier, because the accuracy of the proposed blocking control technique which is used in this paper has been investigated for a similar network topology in Pourbabai (1989a,b, 1990), there was no need to replicate these simulation results.

The optimization models presented in this paper can also be used to evaluate the performance of generalized telecommunication systems with finite-capacity buffers. Ensuring minimal congestion effects, the acceptable blocking probability

of each device should be sufficiently close to zero. Then, the performance of each finite-capacity device (e.g., a G/M/1/ $K_i$  queueing system) can be approximated with the performance of a compatible infinite-capacity device (e.g., a G/M/1 queueing system) conditioned in such a way that the resulted congestion probability does not exceed an acceptable value. Other immediate extensions include profit-maximizing objectives where each device is associated with a profit contribution as a function of its throughput rate and with some distinct buffer allocation costs.

### Acknowledgment

The authors acknowledge the referees for their helpful comments. Partial support for this study has been provided by the IBM program of support for Education in the Management of Information Systems.

### References

- Daskalaki, S., and Smith J.M. (1988), "Real-time routing in finite queueing networks", Working Paper, AT&T Bell Labs, Middletown, NJ.
- Fox, B. (1966), "Discrete optimization via marginal analysis", *Management Science* 13, 210–216.
- Gelenbe, E., and Pujolle, G. (1987), *Introduction to Queueing Networks*, Wiley, New York.
- Genedenko, B.V., and Kovalenko, I.N. (1968), *Queueing Theory*, Weiner Binders, Jerusalem.
- Hahne, E.L. (1981), "Dynamic routing in unreliable manufacturing networks with limited storage, Technical Report LIDS-TH-1603, MIT, Cambridge, MA.
- Jafari, M.A., and Shanthikumar, J.G. (1989), "Determination of optimal buffer storage capacities and optimal allocation in multistage automatic transfer lines", *IIE Transactions* 21, 130–135.
- Kingman, J.F.C. (1969), "Markov population processes", *Journal of Applied Probability* 6, 1–18.
- Kleinrock, L. (1976), *Queueing Systems*, Wiley, New York.
- Knuth, D.E. (1973), *Sorting and Searching*, Addison-Wesley, Reading, MA.
- Larsen, R.L. (1981), "Control of multiple exponential servers with application to computer systems", Ph.D. Dissertation, Department of Computer Science, University of Maryland, College Park, MD.
- Lin, W., and Kumar, P.R. (1984), "Optimal control of a queueing system with two heterogeneous servers", *IEEE Transactions on Automatic Controls* 29, 696–703.
- Onvural, R.O. (1988), "On the exact decomposition of closed queueing networks with finite buffers", *Proceedings of the*

- ple component cost functions", Working Paper QM 87-15, W.E. Simon Graduate School of Business Administration, University of Rochester, Rochester, NY.
- Shalev-Oren, S., Seidmann, A., and Schweitzer, P.J. (1985), "Analysis of Flexible Manufacturing System with priority scheduling: PMVA", *Annals of Operations Research* 3, 115-140.
- Tenenbaum, A.S. (1981), *Computer Networks*, Prentice-Hall, Englewood Cliffs, NJ.
- Tijms, H.C., and Eikeboom, A.M. (1986), "A simple technique in Markovian control with applications to resource allocation in communication networks", *Operations Research Letters* 5, 25-32.
- Yamashita, H., and Suzuki, S. (1987), "An approximate solution method for optimal buffer allocation in serial  $n$ -stage production lines", *Transactions of the Japanese Society of Mechanical Engineers* 53C, 807-814.
- First International Workshop on Queuing Networks with Blocking, Raleigh, May, 84-107.
- Petros, H.G. (1984), "Queuing networks with blocking: a review", *ACM Sigmetrics* 12, 8-14.
- Pourbaba, B. (1989a), "Optimal selection of the local storage in an assembly system", *International Journal of Production Research* 28, 337-352.
- Pourbaba, B. (1989b), "Optimal selection of buffers in a tandem finite capacity G/M/1 queuing system", *Automatica* 25/6, 879-906.
- Pourbaba, B. (1990), "Optimal utilization of a finite capacity assembly system", *International Journal of Production Research* 28, 337-352.
- Seidmann, A., and Schweitzer, P.J. (1984), "Part selection policy for a flexible manufacturing cell feeding several production lines", *IEEE Transactions* 16, 355-362.
- Seidmann, A., and Tenenbaum, A. (1987), "A computational approach to optimal queuing systems controls with multi-